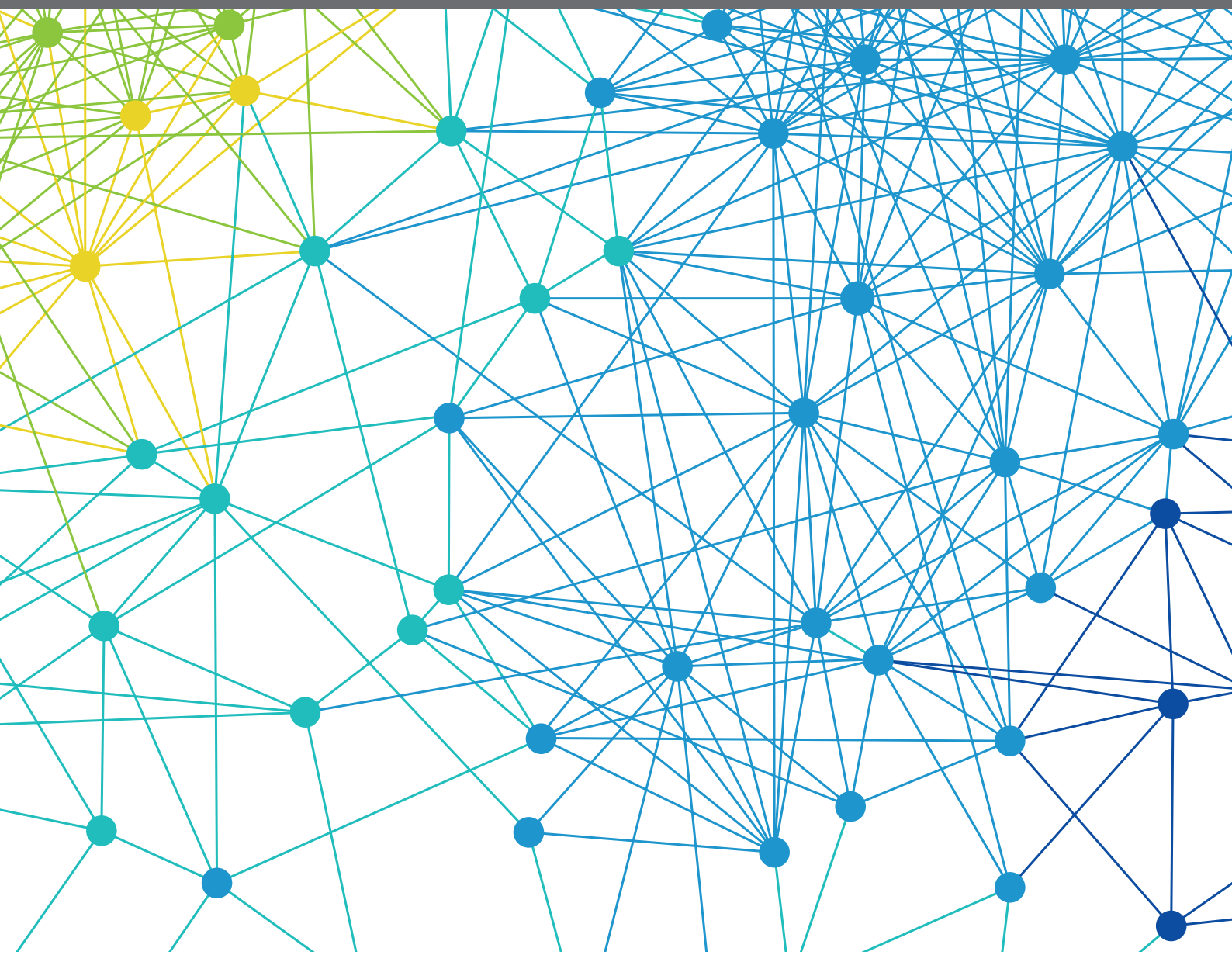
COMPUTATIONAL APPROACHES FOR AGEING AND AGE-RELATED DISEASES

EDITED BY: Stanley Durrleman, Daniel C. Alexander, Ninon Burgos,
Holger Fröhlich, Neil P. Oxtoby and Viktor Wottschel

PUBLISHED IN: Frontiers in Big Data and Frontiers in Artificial Intelligence



frontiers Research Topics

ISSN 1664-8714
ISBN 978-2-88976-766-3
DOI 10.3389/978-2-88976-766-3

About Frontiers

Frontiers is more than just an open-access publisher of scholarly articles: it is a pioneering approach to the world of academia, radically improving the way scholarly research is managed. The grand vision of Frontiers is a world where all people have an equal opportunity to seek, share and generate knowledge. Frontiers provides immediate and permanent online open access to all its publications, but this alone is not enough to realize our grand goals.

Frontiers Journal Series

The Frontiers Journal Series is a multi-tier and interdisciplinary set of open-access, online journals, promising a paradigm shift from the current review, selection and dissemination processes in academic publishing. All Frontiers journals are driven by researchers for researchers; therefore, they constitute a service to the scholarly community. At the same time, the Frontiers Journal Series operates on a revolutionary invention, the tiered publishing system, initially addressing specific communities of scholars, and gradually climbing up to broader public understanding, thus serving the interests of the lay society, too.

Dedication to Quality

Each Frontiers article is a landmark of the highest quality, thanks to genuinely collaborative interactions between authors and review editors, who include some of the world's best academicians. Research must be certified by peers before entering a stream of knowledge that may eventually reach the public - and shape society; therefore, Frontiers only applies the most rigorous and unbiased reviews.
Frontiers revolutionizes research publishing by freely delivering the most outstanding research, evaluated with no bias from both the academic and social point of view. By applying the most advanced information technologies, Frontiers is catapulting scholarly publishing into a new generation.

What are Frontiers Research Topics?

Frontiers Research Topics are very popular trademarks of the Frontiers Journals Series: they are collections of at least ten articles, all centered on a particular subject. With their unique mix of varied contributions from Original Research to Review Articles, Frontiers Research Topics unify the most influential researchers, the latest key findings and historical advances in a hot research area! Find out more on how to host your own Frontiers Research Topic or contribute to one as an author by contacting the Frontiers Editorial Office: frontiersin.org/about/contact

COMPUTATIONAL APPROACHES FOR AGEING AND AGE-RELATED DISEASES

Topic Editors:
Stanley Durrleman, Institut National de Recherche en Informatique et en Automatique (INRIA), France
Daniel C. Alexander, University College London, United Kingdom
Ninon Burgos, Centre National de la Recherche Scientifique (CNRS), France
Holger Fröhlich, Fraunhofer Institute for Algorithms and Scientific Computing (FHG), Germany
Neil P. Oxtoby, University College London, United Kingdom
Viktor Wottschel, Amsterdam University Medical Center, Netherlands

Citation: Durrleman, S., Alexander, D. C., Burgos, N., Fröhlich, H., Oxtoby, N. P., Wottschel, V., eds. (2022). Computational Approaches for Ageing and Age-Related Diseases. Lausanne: Frontiers Media SA. doi: 10.3389/978-2-88976-766-3

Table of Contents

frontiers in Big Data

ORIGINAL RESEARCH
published: 12 August 2020
doi: 10.3389/fdata.2020.00024

Statistical Disease Progression Modeling in Alzheimer Disease

Lars Lau Raket[1,2]* *for the Alzheimer's Disease Neuroimaging Initiative*[†]

[1] H. Lundbeck A/S, Copenhagen, Denmark, [2] Clinical Memory Research Unit, Department of Clinical Sciences, Lund University, Lund, Sweden

OPEN ACCESS

Edited by:
Neil P. Oxtoby,
University College London,
United Kingdom

Reviewed by:
Dan Li,
Ionis Pharmaceuticals, Inc.,
United States
Igor Koval,
Institut National de la Santé et de la Recherche Médicale (INSERM), France

***Correspondence:**
Lars Lau Raket
llra@lundbeck.com

[†] Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: 20 April 2020
Accepted: 24 June 2020
Published: 12 August 2020

Citation:
Raket LL (2020) Statistical Disease Progression Modeling in Alzheimer Disease. Front. Big Data 3:24. doi: 10.3389/fdata.2020.00024

Background: The characterizing symptom of Alzheimer disease (AD) is cognitive deterioration. While much recent work has focused on defining AD as a biological construct, most patients are still diagnosed, staged, and treated based on their cognitive symptoms. But the cognitive capability of a patient at any time throughout this deterioration reflects not only the disease state, but also the effect of the cognitive decline on the patient's pre-disease cognitive capability. Patients with high pre-disease cognitive capabilities tend to score better on cognitive tests that are sensitive early in disease relative to patients with low pre-disease cognitive capabilities at a similar disease stage. Thus, a single assessment with a cognitive test is often not adequate for determining the stage of an AD patient. Repeated evaluation of patients' cognition over time may improve the ability to stage AD patients, and such longitudinal assessments in combinations with biomarker assessments can help elucidate the time dynamics of biomarkers. In turn, this can potentially lead to identification of markers that are predictive of disease stage and future cognitive decline, possibly before any cognitive deficit is measurable.

Methods and Findings: This article presents a class of statistical disease progression models and applies them to longitudinal cognitive scores. These non-linear mixed-effects disease progression models explicitly model disease stage, baseline cognition, and the patients' individual changes in cognitive ability as latent variables. Maximum-likelihood estimation in these models induces a data-driven criterion for separating disease progression and baseline cognition. Applied to data from the Alzheimer's Disease Neuroimaging Initiative, the model estimated a timeline of cognitive decline that spans ∼15 years from the earliest subjective cognitive deficits to severe AD dementia. Subsequent analyses demonstrated how direct modeling of latent factors that modify the observed data patterns provides a scaffold for understanding disease progression, biomarkers, and treatment effects along the continuous time progression of disease.

Conclusions: The presented framework enables direct interpretations of factors that modify cognitive decline. The results give new insights to the value of biomarkers for staging patients and suggest alternative explanations for previous findings related to accelerated cognitive decline among highly educated patients and patients on symptomatic treatments.

Keywords: cognitive decline, dementia, Alzheimer disease, disease staging, biomarkers, disease progression modeling, progression curves, cognitive reserve

BACKGROUND

Alzheimer disease (AD) is slowly progressing with preclinical and prodromal phases lasting many years before the onset of dementia. The stage of the underlying disease process of an AD patient entering a clinical trial is largely unknown, but may be estimated by a combination of, for example, cognitive testing, clinical evaluation, and biomarker results. While these procedures for evaluating disease severity are useful for creating coarse groupings of patients, the factors used to create groupings may be systematically affected by a wealth of factors not directly tied to the disease process, for example, comorbidities, intelligence, level of education, and genetics.

So far, efforts to develop therapies that delay or halt the progression of AD have generally been unsuccessful, and the vast majority of trials testing symptomatic agents in AD have failed. These failures may be due to wrong therapeutic targets or non-efficacious therapies, but it is conceivable that a proportion of trial failures could be attributed to other factors, such as study design, endpoints, and non-optimal patient population selection. For disease-modifying drugs, for example, the current standard durations for interventional studies may not be adequate. Simulations based on cohort studies suggest that prevention of disease in cognitively normal individuals may require study lengths far beyond the current standard to achieve high statistical power for detecting an effect of even very efficacious drugs (Anderson et al., 2017; Insel et al., 2019). Better patient selection and an improved understanding of patient-level cognitive decline could potentially address this problem.

Cognitive Decline and Symptom Onset

The characterizing symptom of AD is cognitive deterioration. The cognitive capability of a patient at any time throughout this deterioration will not directly reflect the disease state, but the cumulative effect of the cognitive decline on the patient's pre-disease cognitive capability.

Many factors influence instantaneous cognitive ability, and low cognitive ability at a single time point is not necessarily an indication of cognitive decline. Cognitive decline can only be established by repeated evaluations of patients' cognition over time. Longitudinal assessments of patient cognition also offer the benefit of hindsight—once cognitive decline or dementia is established, one can traverse back in time along the cognitive trajectory and predict when the decline started and search for patterns that are indicative of future cognitive decline in its earliest stages. If done properly, one can synchronize individual observed trajectories to one long-term timeline representative of the full span and variation of cognitive decline over the course of disease.

Disease Progression Modeling

Alzheimer disease typically presents in a sporadic late-onset form. The autosomal dominant forms of AD (ADAD) caused by rare genetic mutations have earlier onset than sporadic AD, but otherwise, the pathogenesis is largely similar (Bateman et al., 2011). In ADAD, age at symptom onset is strongly affected by mutation type, parental age at symptom onset, APOE genotype, and sex (Ryman et al., 2014). These factors can be used to calculate expected patient age at symptom onset for ADAD patients, which can be used to construct a more synchronized time scale for studying biomarkers and the pathological cascade of the disease (Wang et al., 2019). Furthermore, this makes it possible to do primary prevention studies in a highly efficient manner (Bateman et al., 2017).

In sporadic AD, age at onset cannot be predicted accurately from demographic or genetic factors. Assessment of biomarkers, such as amyloid and tau load in cerebrospinal fluid (CSF) or by positron emission tomography (PET) may be used to diagnose the disease even in the earliest stages (Jack et al., 2018), but such assessments can be both invasive and expensive, and data are sparse. There are, however, rich datasets with longitudinal cognitive measurements that span different parts of the disease. An appealing use of this data is to assemble the individual observed short-term trajectories to one long-term timeline representative of the full span of cognitive decline over the disease.

Different approaches to construct disease progression models for AD have been taken. A classic approach is to formulate the changes in cognitive scores using differential equations (Ito et al., 2011; Gomeni et al., 2012; Samtani et al., 2012; Delor et al., 2013). One major drawback of this type of modeling is that covariate effects and different sources of random variation should be formulated in the differential equation framework and may be very difficult to handle and interpret. A more direct approach to disease progression modeling in AD is event-based models (Young et al., 2014; Oxtoby et al., 2018) where cutoff points of abnormality are inferred from observed biomarkers or clinical scales, and disease stage is mapped to a discrete set of biomarker-abnormality events. Event-based models can improve robustness, but the dichotomization of variables also reduces the granularity of the results, especially for variables that do not show a bimodal distribution and/or continuously evolve with disease progression.

An alternative class of disease progression models relies on direct modeling of the observed longitudinal trajectories and explicit modeling of the patient-level disease stage (Jedynak et al., 2012). An important example of this type of approach is the model by Donohue et al. (2014), which simultaneously models multiple observations of cognitive measures and biomarkers. This modeling approach has been powerful in illuminating the multivariate nature of AD progression. The approach was recently generalized to a wider class of Bayesian latent-time joint mixed-effects models (Li et al., 2018). This generalized class of models allows dependencies between different outcomes and inclusion of covariates, but covariates can only model variation in outcomes and not disease stage or progression rate.

For modeling disease progression of very high-dimensional data with rich structure, such as brain imaging, disease progression models are often considered in the context of Riemannian geometry (Louis et al., 2019). While there have been recent advances in the range implementable models (Schiratti et al., 2017; Koval et al., 2018), the complexity and computational demand are still restricting the types data and effects that can be modeled by these approaches.

For practical applications of disease progression modeling, there are several important considerations to make. Simultaneous modeling of multiple outcomes is desirable as one can detect signals in multiple outcomes that may reduce noise in the staging of patients. However, it typically comes with an assumption of all outcomes being synchronized in the same disease time model (Donohue et al., 2014; Li et al., 2018). Therefore, care should be taken when deciding which outcomes to include. For example, if a group of individuals with different age-related neurodegenerative diseases is modeled, these individuals may all experience progressive dementia that can be mapped to a common trajectory of cognitive decline, but their individual biomarker measurements may be very different along this trajectory, and including these as outcomes may deteriorate the quality of the staging. Conversely, if most individuals have the same cause of their cognitive decline (e.g., AD), including biomarkers may help staging patients in the early stages of disease where no or little cognitive deficit is detectable. Another important consideration is the time scale at which to model disease progression. Age is typically considered the major risk factor for developing AD, but age at first diagnosis of AD can vary by decades between patients, and because this span is much greater than the entire course of cognitive decline associated with AD, patient age is not an appropriate scale for understanding the pattern of cognitive decline in AD because it may amplify the *a priori* dis-synchronization between patients by orders of magnitude. For example, two individuals diagnosed with AD dementia at 60 and 90 years of age, respectively, may have similar courses of cognitive decline, but an age-indexed model would have to compensate for the additional 30 years' difference when compared to a diagnosis-indexed model. The negative consequence of this can, for example, be seen in Figure 1 in Li et al. (2018), where patient-level trajectories go from minimal to maximal severity over 10–15 years, while variation of when maximal severity is reached between patients is spread out over 30-years periods. Therefore, a more natural scale for studying the patterns of cognitive decline is time since symptom onset. However, self- or caregiver-reported age at symptom onset is not perfect either. It may be imprecise because of the patient's memory problems; recall bias, where early sporadic cognitive issues are believed to be symptoms of the disease; or personal differences in sensitivity and interpretation of the earliest cognitive problems.

In this article, we propose a new approach to disease progression modeling that separates disease stage and deviations from the mean pattern in a fully data-driven manner. The model enables more detailed modeling and analysis of some of the aspects of cognitive decline compared to previous models. For example, it allows investigation of whether observed variables are related to cognitive ability, disease stage, or rate of decline. In the presented form, the model is estimating a disease timeline from repeated assessments of a univariate measure, such as a cognitive scale. The model is inspired by the statistical framework presented by Raket et al. (2014), where systematic patterns of variation in both vertical (observed cognitive score) and horizontal (disease timing) directions are modeled simultaneously on both the population and individual levels. The model allows covariate effects on both outcomes and disease progression, and all model parameters are estimated simultaneously using maximum likelihood estimation.

The goal of this work was to explore whether the proposed disease progression model could align observed cognitive trajectories to a precise timeline of cognitive decline associated with AD and to evaluate if this modeling would shed new light on aspects related to disease progression and biomarkers. When the model was fitted to cognitive scores from Alzheimer's Disease Neuroimaging Initiative (ADNI), the presented model aligned the cognitive trajectories of patients to a consistent shape of cognitive decline with a span of ∼15 years from the earliest subjective cognitive deficits to severe AD dementia. It was shown that the model's predictions of patients' disease stages based on their longitudinal cognitive scores could predict time since symptom onset and diagnosis. It was further demonstrated that the predicted disease stages provided a more suitable time scale for modeling the evolution of biomarkers over the course of disease than group-wise modeling based on patient symptoms at baseline. The model was used to estimate the effects of sex, age, and education on cognitive decline and to evaluate the effects of cholinesterase inhibitor (ChEI) treatment on cognitive decline. Finally, the model was fitted to the cognitive trajectories of a subset of patients with a rich set of biomarkers available at baseline to estimate if baseline biomarker profile could predict disease stage. The results of the model in an independent held-out validation dataset confirmed that baseline biomarker profiles could predict the disease stage of unseen individuals—even in the preclinical phases of disease where no clinically detectable cognitive impairment was present.

METHODS

Data

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public–private partnership, led by the principal investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early AD. For up-to-date information, see www.adni-info.org.

Patients included in the current study were required to have a valid classification at baseline [cognitively normal, significant memory concern, MCI (early), MCI (late), or dementia].

Outcomes

The main outcome measure considered was the total score of the 13-item Alzheimer's Disease Assessment Scale–Cognitive Subscale (ADAS-Cog; range = 0–85; lower score indicates less impairment) (Mohs et al., 1997). Included patients were required to have at least one valid ADAS-Cog total score to be included in the present study.

Other outcomes reported were onset of various symptoms related to cognitive impairment and AD; Clinical Dementia Rating scale—sum of boxes (Hughes et al., 1982); Functional

Activities Questionnaire (Pfeffer et al., 1982); fluorodeoxyglucose (FDG) PET meta region of interest (meta-ROI) (Landau et al., 2011); cross-sectional hippocampal volume extracted from MRI using FreeSurfer (Fischl, 2012); florbetapir PET SUVr (Landau et al., 2015); $A\beta_{1-42}$, total tau, and $p\text{-tau}_{181}$ concentrations in CSF as measured using the Roche Elecsys® immunoassay (Bittner et al., 2016); the ratio of $A\beta_{1-42}/A\beta_{1-40}$ concentrations in CSF as measured by two-dimensional ultraperformance liquid chromatography tandem mass spectrometry; and neurofilament light chain (NfL) levels in plasma measured using a single-molecule array platform (Mattsson et al., 2019).

Disease Progression Model

Let y_{ij} represent the observed cognitive score of patient i at time t_{ij} ($i = 1, \ldots, n, j = 1, \ldots, m_i$). We assume that y_{ij} is generated by a model of the form

$$y_{ij} = \theta\left(w_i\left(t_{ij}\right)\right) + x_i\left(t_{ij}\right) + \varepsilon_{ij}$$

where θ is a function that represents the shape of cognitive decline; w_i is a *warping* function that transforms observation time t_{ij} to a disease time scale $w_i\left(t_{ij}\right)$ that is aligned across patients; x_i is the idiosyncratic patient-level deviation from the mean shape that represents consistent deviations over time; and ε_{ij} is independent measurement noise.

Cognitive scores can be extremely noisy because of many different sources of variation, and one will have to make suitable model choices to accurately infer the shape of the disease timeline of cognitive decline θ, to predict patient-level disease stage w_i, and to predict the entire patient-level course of decline $\hat{y}_i$. In the following, we describe the basic model choices taken here and their motivations.

Because we are modeling cognitive decline in pathological aging, it is natural to assume that the representative shape of decline θ is a function that has a stable left asymptote (pre-disease cognitive normality) and a monotone decline. In this article, we focus on ADAS-Cog scores that show a distinct exponential decline in dementia (Yang et al., 2011), and thus we will work with a parametrized family of exponential functions to model the mean progression pattern

$$\theta(t) = l \cdot \exp\left(\frac{t+s}{\exp(g)}\right) + v,$$

It is worth noting that this choice of θ is overparametrized unless restrictions are put on some of the parameters. The constraint used here (discussed further below) is that $s = 0$ for the cognitively normal individuals, in which case v is the left asymptote representing the average stable pre-disease cognitive score and where the remaining parameters determine the shape of the decline.

The mean progression pattern θ can be modeled differently to achieve other properties, for example, as a generalized logistic function or as a monotone spline (Ramsay, 1988). Other modeling options are available in the progmod R package (Raket, 2020) accompanying this article.

The mapping of observed time to disease time w_i should allow the model to assemble short-term longitudinal observations to a long-term timeline of cognitive decline. Because the major source of horizontal variation can likely be ascribed to differences in how long the patient has had the disease before we begin observing them, we model w_i as a shift of study time.

$$w_i(t) = t + s_i.$$

Random Effects

When modeling longitudinal data for groups of individuals, it is often natural to describe systematic differences between individuals using random effect. The proposed disease progression model has three types of random effects.

- s_i: Random patient-level shift that models the disease stage of patient i. Assumed to follow a zero-mean normal distribution with unknown variance τ^2.
- x_i: Random patient-level systematic deviation from the mean curve. Assumed to be a sum of discrete-time observation of a Brownian motion $x_{i,\,BM}$ and an independent zero-mean normally distributed starting level $x_{i,\,0}$ with unknown variance. The covariance function of x_i is thus $C(t, t') = \sigma^2_{BM} \cdot \min(t, t') + \sigma^2_0$ where σ^2_{BM} is an unknown parameter controlling variance scale of the Brownian motion, and σ^2_0 is an unknown parameter controlling the variance of the random starting level.
- ε_{ij}: Random observation noise. Assumed to be independent zero-mean normally distributed with unknown variance σ^2.

A free correlation between s_i and the starting level $x_{i,\,0}$ is included in the model; the remaining effects are assumed independent.

Fixed Effects

The basic model parameters l, g, s, and v that describe the shape of θ are modeled as fixed effects.

- l is a scaling parameter of the exponential function. Because a goal of disease progression modeling is to find a common pattern of decline, l will be modeled as a single free parameter.
- g is a scaling parameter of time. Patient-level differences in rate of decline that can be ascribed to a covariate or factor can be modeled as a regression-type model on g. Initially, this parameter will be modeled as a single free parameter.
- s is a shift of observed time. Patient-level differences in disease stage that can be ascribed to a covariate or factor can be modeled as fixed effects. Because the present study includes several cohorts at different disease stages (e.g., cognitively normal individuals, patients with dementia), the initial modeling will have different s parameters for non-cognitively normal cohorts and $s = 0$ for the cognitively normal individuals to ensure structural identifiability of the model (Lavielle and Aarons, 2016). Thus, s is modeling disease time since the average baseline stage of the cognitively normal individuals.
- v is an intercept parameter describing the left asymptote. Patient-level differences in pre-disease cognition that can be ascribed to a covariate or factor can be modeled as a

regression-type model on v. Initially, this parameter will be modeled as a single free parameter.

Final Model Formulation

The basic model used has the form

$$y_{ij} = l \cdot \exp\left(\frac{t_{ij} + s_i + X_{s,ij}\beta_s}{\exp(g)}\right) + v + x_{i,\,BM}\left(t_{ij}\right) + x_{i,\,0} + \varepsilon_{ij}$$

where l, g, $v \in \mathrm{R}$ and $\beta_s \in R^4$ are free fixed effects, and $X_{s,ij}$ is the four-dimensional dummy row-vector indicating which, if any, of the symptomatic baseline groups individual i belongs to. The random effects follow a joint zero-mean normal distribution

$$\begin{pmatrix} s_i \\ x_{i,\,0} \\ x_{i,\,BM}\left(t_{ij}\right) + x_{i,\,0} + \varepsilon_{ij} \\ x_{i,\,BM}\left(t_{ik}\right) + x_{i,\,0} + \varepsilon_{ik} \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \tau^2 & \rho & \rho & \rho \\ \rho & \sigma_0^2 & \sigma_0^2 & \sigma_0^2 \\ \rho & \sigma_0^2 & \sigma_{BM}^2 \cdot t_{ij} + \sigma_0^2 + \sigma^2 & \sigma_{BM}^2 \cdot \min\left(t_{ij}, t_{ik}\right) + \sigma_0^2 \\ \rho & \sigma_0^2 & \sigma_{BM}^2 \cdot \min\left(t_{ij}, t_{ik}\right) + \sigma_0^2 & \sigma_{BM}^2 \cdot t_{ik} + \sigma_0^2 + \sigma^2 \end{bmatrix}\right)$$

where τ^2, σ_{BM}^2, σ_0^2, $\sigma^2 > 0$ are unknown variance parameters, and $\rho \in R$ is a parameter that controls the correlation between the random time shift s_i and the random starting level $x_{i,\,0}$.

Statistical Analysis

Estimation in the disease progression model was done with maximum likelihood using the two-step algorithm of Lindstrom and Bates (1990). Random effects were predicted as the most likely values given the data and maximum likelihood parameter estimates (i.e., they maximized the posterior under the parameter estimates).

To investigate the effect of covariates on the pattern of disease progression, forward selection was used to evaluate models with all combinations of covariate effects on rate of decline g, disease stage s, and pre-disease cognition v. The search was continued as long as the Akaike Information Criterion (Akaike, 1998) improved, but the model selection was based on the more conservative Schwarz's Bayesian Information Criterion (BIC) (Schwarz, 1978).

To investigate if predicted disease time was predictive of time since reported symptom onset, linear regression was done on time since reported symptom onset (at baseline) using predicted disease time as a covariate. P-values were computed using t-tests.

Linear mixed-effects modeling was used to investigate if predicted disease time offered a better time scale for modeling other longitudinal outcomes (e.g., biomarkers) than time since baseline for the five baseline groups. To allow for non-linear trends in the mean pattern, the outcome was modeled using a cubic B-spline function with 3 degrees of freedom plus an intercept across predicted disease time and time since baseline (one pattern per baseline group), respectively. Patient-level random slopes and intercepts were included to model longitudinal deviations within an individual. P-values were computed using likelihood ratio tests with maximum likelihood estimation.

Comparisons of quantitative outcomes between groups with two levels were done using Wilcoxon rank sum tests, and correlations were evaluated with Spearman rank correlation coefficients.

Software

All analyses were done using R version 4.0.0 (R Core Team, 2020). Maximum likelihood estimation in the disease progression models was done using the progmod R package (Raket, 2020), which builds on the estimation procedures available in the nlme and covBM R packages (Pinheiro et al., 2019).

RESULTS

Basic Model

The basic model described above was fitted on longitudinal ADAS-Cog data from ADNI. The data comprised 9,830 ADAS-Cog scores across 2,142 individuals. The ADAS-Cog scores plotted against study time are shown in the top panel of **Figure 1**. The middle panel in **Figure 1** shows the fixed-effects staging of the baseline status groups relative to the cognitively normal group on the predicted time scale ("disease month"). The bottom panel of **Figure 1** shows the predicted individual staging (both fixed and random effects) of trajectories on the predicted time scale. Relative to the average baseline disease stage of the cognitively normal group, the model estimated that the significant memory concern group was 29 months later into the trajectory of cognitive decline, whereas the early and late MCI groups were, respectively, 42 and 88 months later and that the dementia group was 136 months later. The model had 12 degrees of freedom, and twice the negative log likelihood of the fitted model was 59,468.52. AIC and BIC were 59,492.52 and 59,578.84, respectively.

Validation of the Basic Model

The presented disease progression model aggregates the information in baseline status groups and the longitudinal trajectories of participants to a single number, the predicted disease month. For this continuous disease progression scale to be relevant to AD, it should also hold information that describes other aspects of the disease than the cognitive deterioration observed on ADAS-Cog that the model was fitted on.

To evaluate whether the disease progression model captured milestones of cognitive deterioration, we investigated the model's ability to predict self-reported onset of cognitive symptoms, MCI symptoms, AD symptoms, or diagnosis of AD. There were 1,142 participants who had at least one entry of these data during the study follow-up. Age at symptom onset or diagnosis plotted against the age at the model's predicted disease time 0

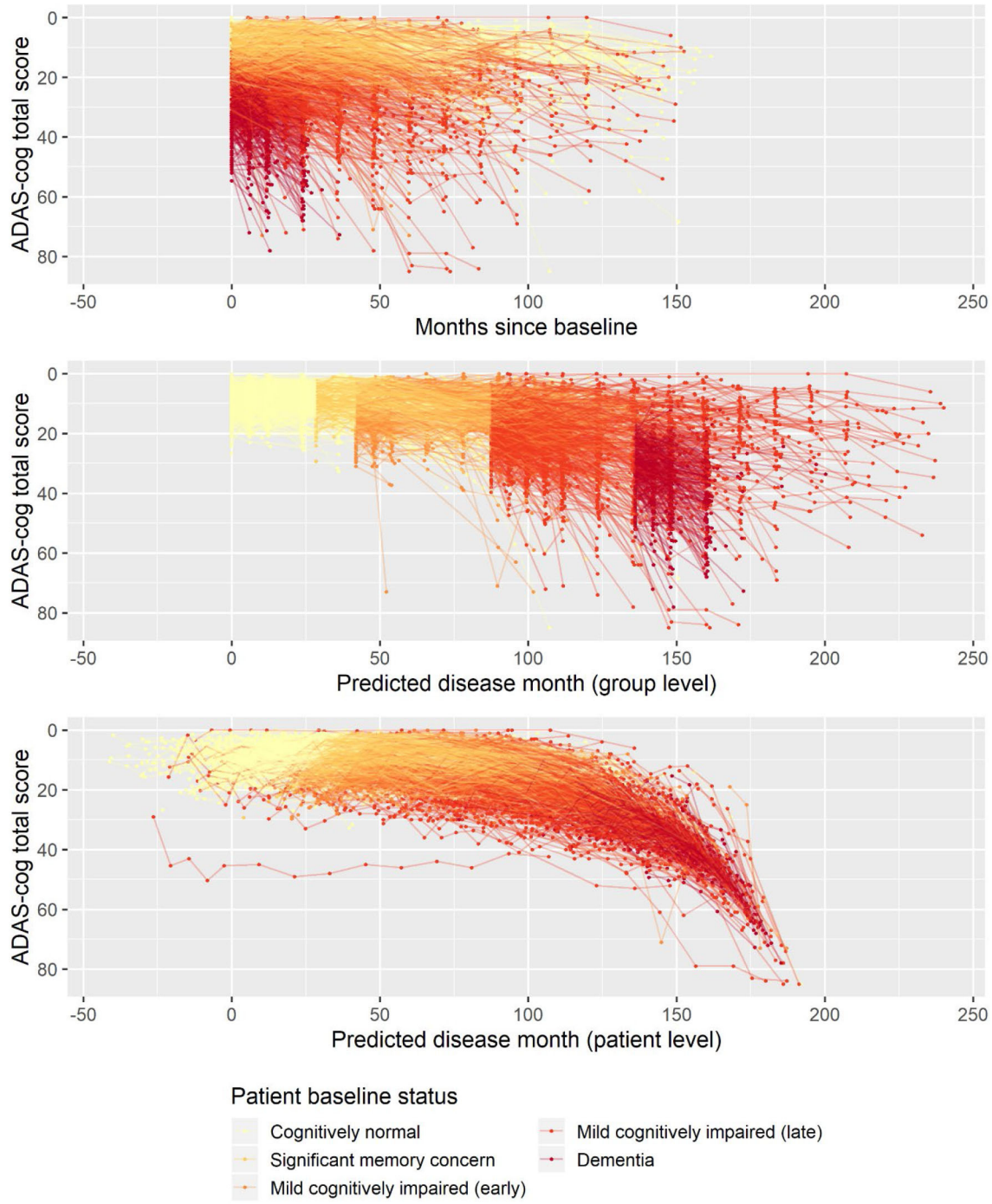


FIGURE 1 | Observed longitudinal ADAS-Cog trajectories for 2,142 ADNI participants plotted against time in study (top), predicted disease time based on the fixed-effects staging of the different patient baseline status groups relative to the cognitively normal group (middle), and predicted individual disease time based on both fixed group and random individual effects.

(computed as age at baseline minus predicted shift in disease time in years) is shown in **Figure 2**. In an ideal setting where trajectories were perfectly aligned and onset/diagnosis would be perfectly consistently reported across individuals, the results of each measure would lie on a line with slope 1, and the intercept would represent the difference in years between age at disease time 0 and the age at onset/diagnosis time. For the age at onset of cognitive symptoms, there seem to be different intercepts for the different baseline groups, where more severe baseline groups tend to report symptom onset later relative to the model prediction of the less severe groups. This may be an effect of different subjective definitions of onset of cognitive symptoms across

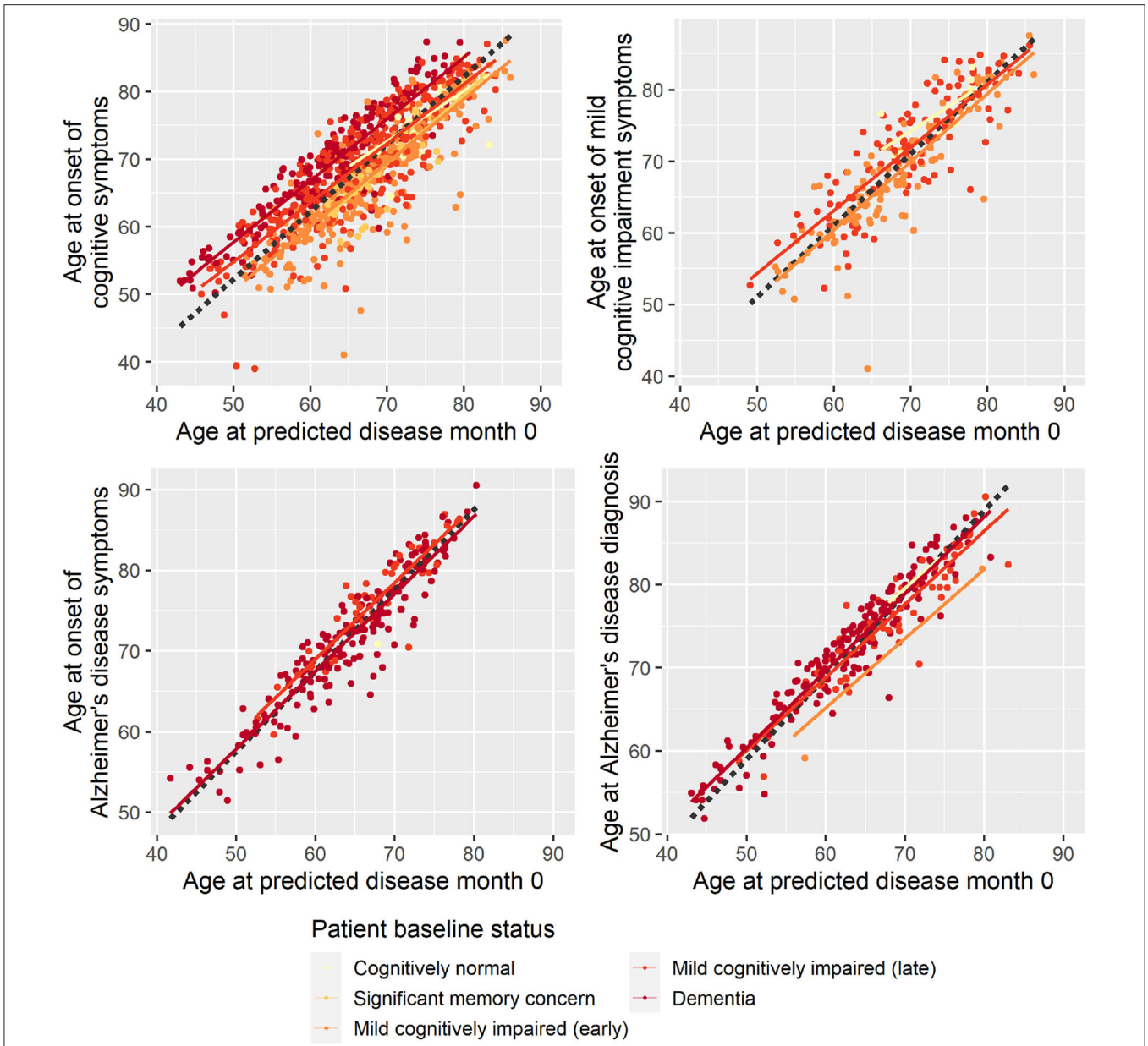


FIGURE 2 | Reported age at onset of cognitive symptoms (top left), cognitive impairment symptoms (top right), Alzheimer disease symptoms (bottom left), and Alzheimer disease diagnosis (bottom right) as a function of age at predicted disease month 0. Dotted lines represent the best-fitting least-squares estimated lines with slope 1.

baseline groups, it may be because of biased model estimates of the staging of the baseline groups, or a combination.

Based on linear regression, predicted disease month was predictive of time since cognitive symptom onset ($p < 0.0001$), time since AD symptoms onset ($p < 0.0001$), and time since Alzheimer diagnosis ($p < 0.0001$)—all times relative to study baseline. Predicted disease month was not significantly predictive for time since MCI symptom onset ($p = 0.558$).

Second, to validate that the predicted disease time also synchronized other independently captured aspects of the disease than cognition as measured by ADAS-Cog, we analyzed if the predicted continuous disease scale better captured patterns of variation in other clinical scales and biomarkers than separate modeling of the different baseline groups. We found that predicted disease time better described the patterns of variation compared to allowing separate patterns per baseline group in 7 of the 10 outcomes when measured by log likelihood (**Table 1**), even though the latter model had 16 degrees of freedom more than the former. When measured using AIC and BIC that both adjust for additional degrees of freedom to compare model quality, the predicted disease time model was better in 8 of the 10 cases for AIC and 10 of the 10 cases for BIC. Interestingly, the three

TABLE 1 | Comparison of longitudinal modeling of clinical scales and biomarkers based on patient baseline group vs. continuous disease time.

Outcome measure	Number of observations (number of patients)	One trajectory per baseline group ($df = 24$)			One trajectory across predicted disease time ($df = 8$)		
		−2·Log likelihood	AIC	BIC	−2·Log likelihood	AIC	BIC
Clinical Dementia Rating Scale—sum of boxes	9,712 (2,142)	29,584.0	29,632.0	29,804.3	**29,062.3**	**29,078.3**	**29,135.8**
Functional Activities Questionnaire	9,715 (2,126)	51,328.2	51,376.2	51,548.5	**50,526.3**	**50,542.3**	**50,599.8**
FDG-PET (meta-ROI)	3,461 (1,454)	−7,185.3	−7,137.3	−6,989.7	**−7,594.8**	**−7,578.8**	**−7,529.6**
Hippocampal volume (MRI)	6,052 (1,675)	88,404.7	88,452.7	88,613.7	**88,372.1**	**88,388.1**	**88,441.7**
Florbetapir PET SUVr	2,568 (1,224)	**−3,466.3**	−3,418.3	−3,277.9	−3,430.2	**−3,414.2**	**−3,367.4**
$A\beta_{1-42}$ (CSF)	2,342 (1,252)	**33,958.5**	**34,006.5**	34,144.7	34,011.1	34,027.1	**34,073.1**
$A\beta_{1-42}/A\beta_{1-40}$ (CSF)	1,425 (867)	**−5,838.2**	**−5,790.2**	−5,663.9	−5,816.5	−5,800.5	**−5,758.4**
Total tau (CSF)	2,334 (1,247)	26,441.1	26,489.1	26,627.2	**26,353.7**	**26,369.7**	**26,415.7**
p-tau_{181} (CSF)	2,330 (1,246)	15,849.8	15,897.8	16,035.9	**15,793.6**	**15,809.6**	**15,855.6**
NfL (plasma)	4,219 (1,576)	37,584.5	37,632.5	37,784.8	**37,517.8**	**37,533.8**	**37,584.6**

Comparison in terms of −2·log likelihood, AIC and BIC (smaller is better for all measures). Bold numbers indicate the best-fitting model for a given measure. df, degrees of freedom.

biomarkers where group-wise modeling was better as measured by log likelihood were all measures related to amyloid burden (CSF $A\beta_{1-42}$ and $A\beta_{1-42}/A\beta_{1-40}$ ratio, florbetapir PET). These biomarkers are known to have a bimodal distribution (Palmqvist et al., 2015) and are thus poorly modeled by a single trajectory. The estimated trajectories of the two types of models for cognitive scales are shown in **Figure 3**, imaging data trajectories are shown in **Figure 4**, and CSF and plasma biomarker trajectories are shown in **Figures 5**, **6**. For some outcomes, the per-baseline group modeling approach had too many degrees of freedom for the significant memory concern and dementia groups. In these cases, the estimated mean trajectories oscillate during time periods where no data were collected. For the non-amyloid biomarkers, reducing the degrees of freedom for these group would have no bearing on the results in **Table 1**.

Age, Sex, Education, and Cognitive Decline

There were systematic differences in follow-up time, age at baseline, and length of education between male and female participants (**Supplementary Table 1**). Compared to female participants, male participants on average had 3.2 months' longer follow-up (Wilcoxon $p = 0.0085$), were on average 2.0 years older at baseline (Wilcoxon $p < 0.0001$), and had 0.89 years more education (Wilcoxon $p < 0.0001$). Age at baseline and years of education were not significantly correlated (Spearman $\rho = -0.04$, $p = 0.0792$).

To explore whether age at baseline, sex, and length of education affected the pattern of cognitive decline, stepwise forward model selection was done to include these factors in the model. The best model included fixed covariate effects of age and sex on g, s, and v, and fixed covariate effects of years of education on g and v. While there were some substantial differences in marginal parameter estimates due to age, sex, and length of education (e.g., men are predicted to be 57 months later in disease compared to women in the same baseline groups; **Supplementary Table 2**), the estimates should not be interpreted in isolation because all parameters simultaneously affect the shape of the disease trajectory and may counteract each other. **Figure 7** shows how age, sex, and education differences systematically affected the mean trajectories. From the figure, we see that male participants consistently scored lower on ADAS-Cog throughout the disease (3.1 points), but that they remained more stable in the initial 100 months where female participants had a more gradual decline. Lower age at baseline and longer education were both associated with higher cognitive scores, but also slightly increased rates of decline as evident in the stages of overt dementia (predicted disease time >120 months).

Cholinesterase Inhibitors and Cognitive Decline

Using the search terms described in the Supplementary Material, we identified 1,347 individuals that were treated with ChEIs, which are approved for symptomatic treatment of AD. There were no restrictions to brand or dose used. Only 64 of the identified patients had records of initiation or discontinuation of treatment during the observation time (total of nine initiations and 60 discontinuations).

To explore if treatment with ChEIs affected the shape of the decline trajectories, stepwise forward model selection from the basic model was done to include ChEI treatment in the model. The best model included fixed effects of treatment on s and v, but not on rate of decline g (14 degrees of freedom, twice the negative log likelihood = −59,277.59, AIC = 59,305.59, BIC = 59,406.29). The model found that patients treated with ChEIs generally had worse level of cognition (effect on v was 5.50 ADAS-Cog points for treated individuals, $p < 0.0001$) and a delayed progression within baseline groups (effect on s was 7.53 months, $p < 0.0001$). The average trajectories and distribution of data across treatment are shown in **Figure 8**.

Biomarkers for Disease Staging

The disease progression model relies on observing patients longitudinally and uses the temporal patterns of cognitive scores to predict the patient's status at baseline. This type of approach

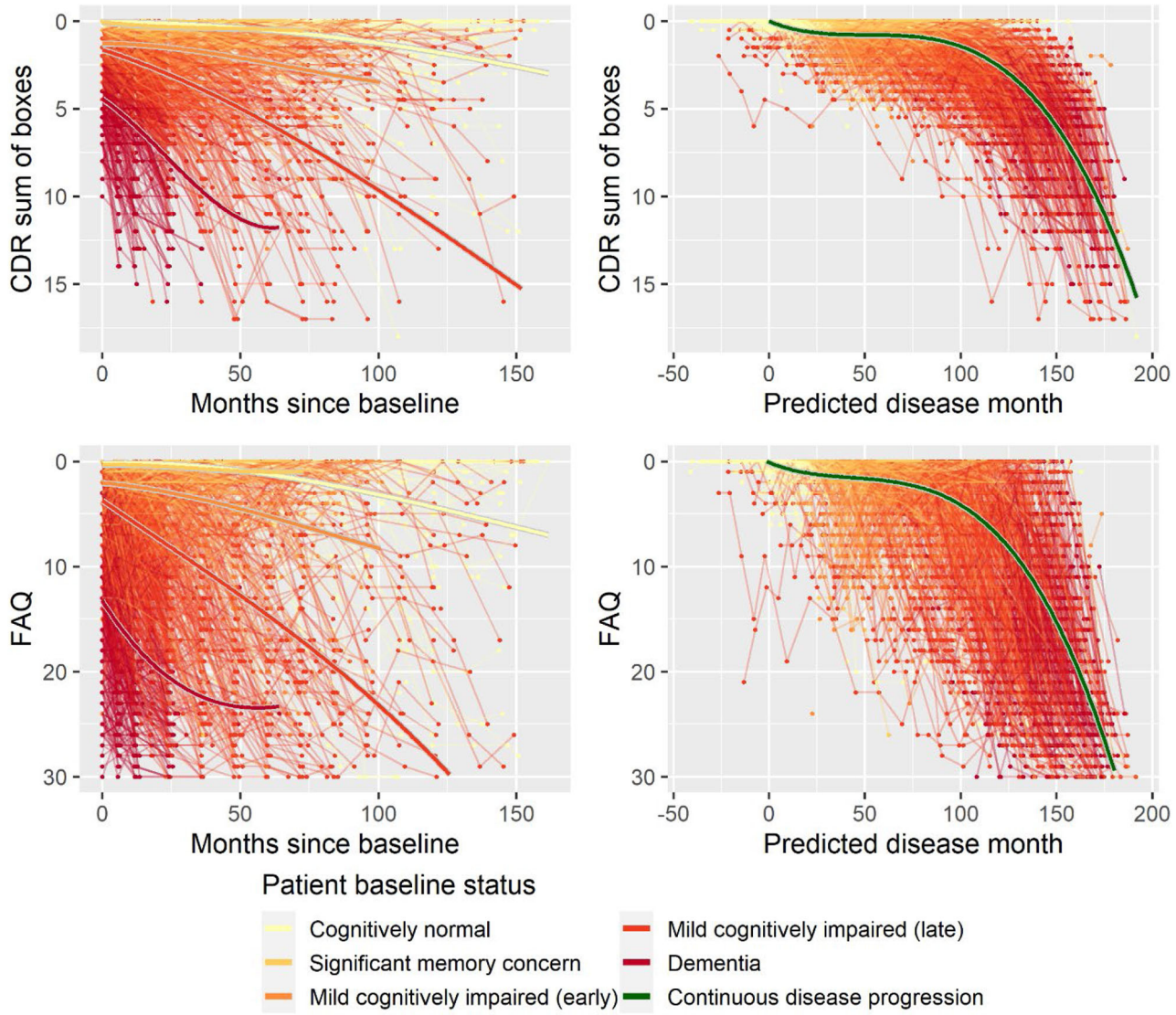


FIGURE 3 | Data and estimated biomarker trajectories for CDR sum of boxes and FAQ. Left column shows results when allowing different trajectories for the five baseline groups; right column shows the results when requiring a single trajectory over predicted disease time.

is needed for understanding the progression of disease and is valuable in retrospective cohort analyses. But the models presented thus far offer only little insight into the disease stage of a patient that has not been followed longitudinally, for example, a patient entering a clinical trial. In this setting, only the baseline classification of the patient, the cognitive score, and possibly other demographic data would be able to inform the stage of the patient. However, as shown in *Validation of the Basic Model*, several biomarkers have distinct temporal patterns over the course of predicted disease time. Biomarker data collected at baseline may thus enable a better assessment of the stage of an individual.

The following analyses were done on the 688 individuals who had complete biomarker data at baseline for the eight biomarkers considered in *Validation of the Basic Model*. These individuals had 3,301 visits with valid ADAS-Cog scores.

Training and Validation Data

Five hundred forty individuals (80%) were randomly selected for the training cohort, and the remaining 148 (20%) comprised the validation cohort.

Model Development

Using the BIC-based model selection procedure described previously, we searched for the best model among models that included adjustment for sex, baseline age, and education (on parameters g, s, v), as well as adjustment for the eight baseline biomarkers on disease stage (parameter s). The model selection was done on the training data. The best model included the biomarkers FDG-PET (meta-ROI), hippocampal volume (MRI), florbetapir PET SUVr, $A\beta_{1-42}/A\beta_{1-40}$ (CSF), and NfL (plasma) (22 degrees of freedom, $-2\cdot$log likelihood

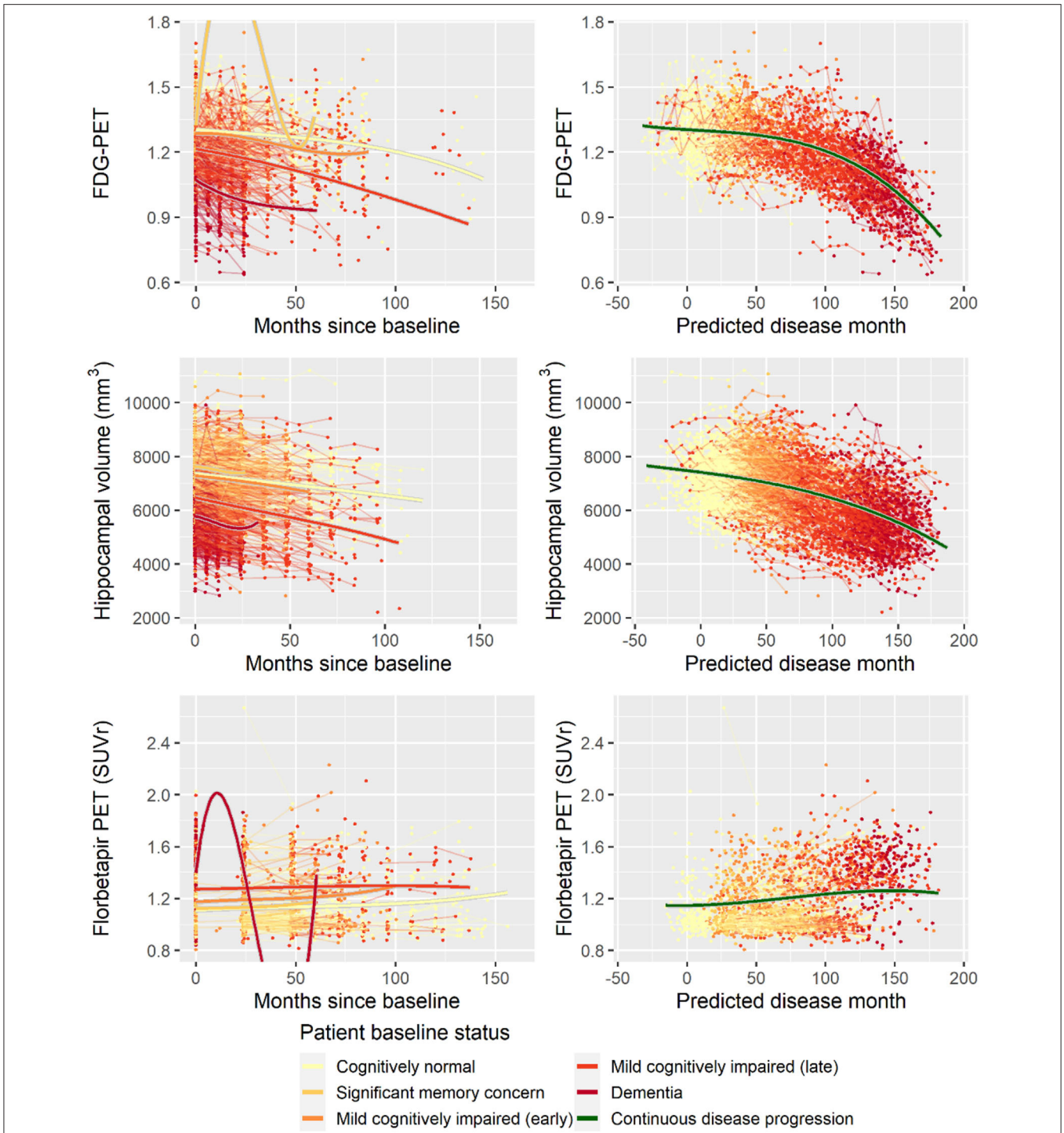


FIGURE 4 | Data and estimated biomarker trajectories for FDG-PET, hippocampal volume (MRI), and florbetapir PET. Left column shows results when allowing different trajectories for the five baseline groups; right column shows the results when requiring a single trajectory over predicted disease time. Note that the oscillations for the significant memory concern and dementia groups on FDG-PET and florbetapir PET, respectively, occur in time periods where no data were collected for these groups.

= 15,182.34, AIC = 15,226.34, BIC = 15,355.45). The parameter estimates for the model are given in **Supplementary Table 3**.

Model Validation

To validate the biomarker model, the model fitted on training data was used to predict disease stage in two different

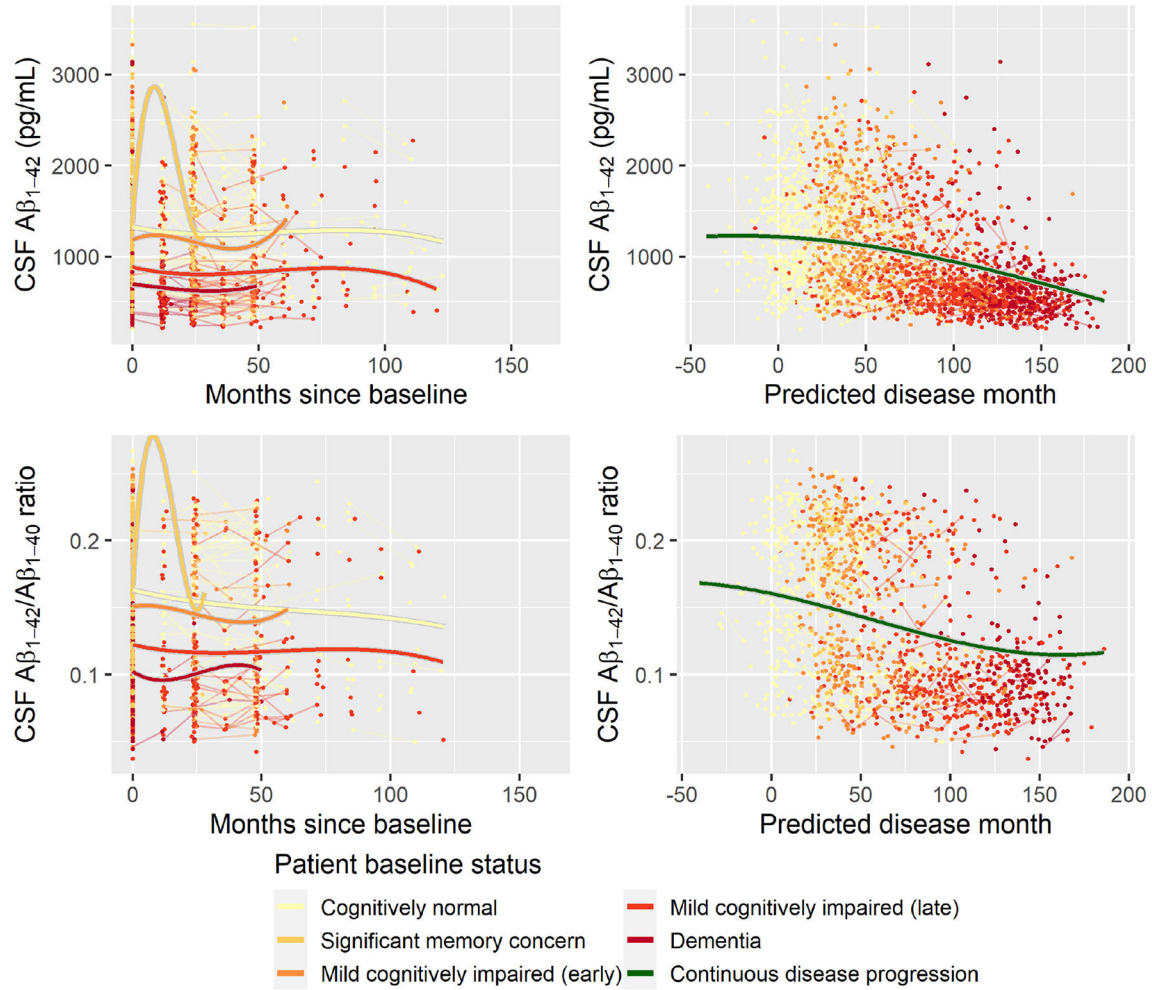


FIGURE 5 | Data and estimated biomarker trajectories for $A\beta_{1-42}$ (CSF) and $A\beta_{1-42}/A\beta_{1-40}$ ratio (CSF). Left column shows results when allowing different trajectories for the five baseline groups; right column shows the results when requiring a single trajectory over predicted disease time. Note that the oscillations for the significant memory concern group occur in time periods where no data were collected for this group.

scenarios. In addition to patient status, the first used only baseline biomarker data, whereas the second also used baseline ADAS-Cog total score. Visual inspection of the longitudinal ADAS-Cog trajectories suggests that the baseline data do hold information that improves prediction of disease stage in the test data (**Figure 9**). To quantify this, the predictive accuracy of the biomarker model was compared to the basic model that did not include biomarker on the longitudinal ADAS-Cog total score trajectories (**Table 2**). Inclusion of biomarker data clearly reduced the mean squared error (MSE) and median absolute error (MAE) of predictions on both test and training data (MSE/MAE 65.1/4.21 vs. 100.0/4.98 on test data). Including the baseline ADAS-Cog total score improved the post-baseline predictive accuracy of the biomarker model further (MSE/MAE 55.1/3.48 for baseline biomarkers + ADAS-Cog model vs. 69.8/4.31 for biomarker model on test data).

DISCUSSION

Disease Progression Modeling

In this article, we presented a model for progression of dementia based on longitudinal cognitive assessments. Disease stages of individual patients were modeled using a latent variable approach. As opposed to conventional latent variable models, for example, those used in item response theory for modeling cognitive tests (Balsis et al., 2012; Embretson and Reise, 2013), the proposed model imposes explicit structures to ensure that the longitudinal modeling respects the known course of disease (e.g., that disease progression is an increasing function of elapsed time and that cognition on average declines with disease progression). By imposing these structures, the model provides a scaffold for understanding disease progression in pathological aging in terms of three continuous measures, *disease stage*, *rate of decline*, and *cognitive deviation from the mean*.

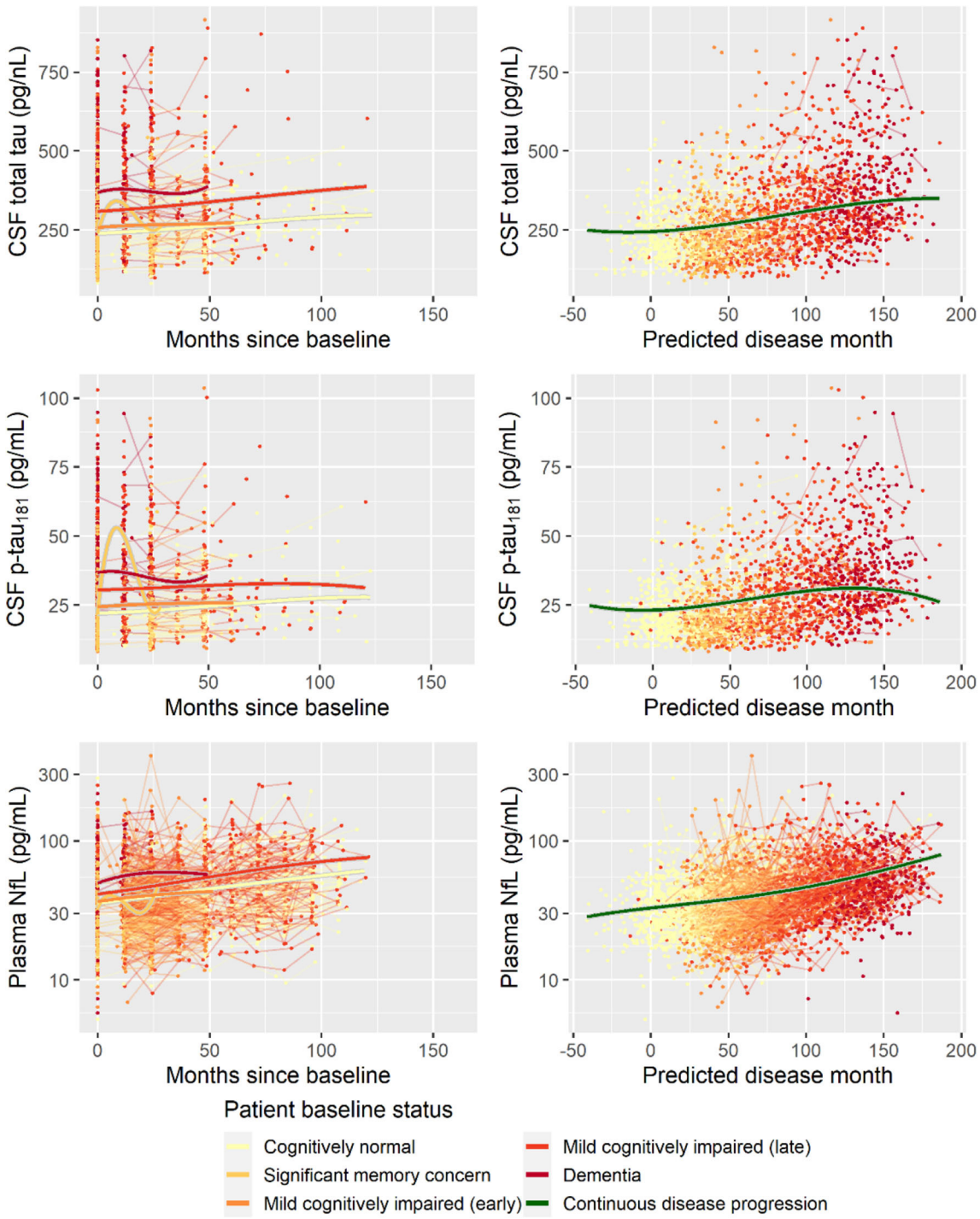


FIGURE 6 | Data and estimated biomarker trajectories for total tau (CSF), p-tau$_{181}$ (CSF), and NfL (plasma). Left column shows results when allowing different trajectories for the five baseline groups; right column shows the results when requiring a single trajectory over predicted disease time. Note that the oscillations for the significant memory concern group occur during time periods where no data were collected on the respective biomarkers.

The proposed model aligned trajectories of cognitive decline. To demonstrate that this approach provided valid insights about other aspects of the disease, it was shown that predicted disease time was predictive of various measures of disease onset. Furthermore, the use of ADAS-Cog trajectories to map patients to a one-dimensional disease timeline was shown to consistently provide a better explanation of other clinical scales and biomarker trajectories than

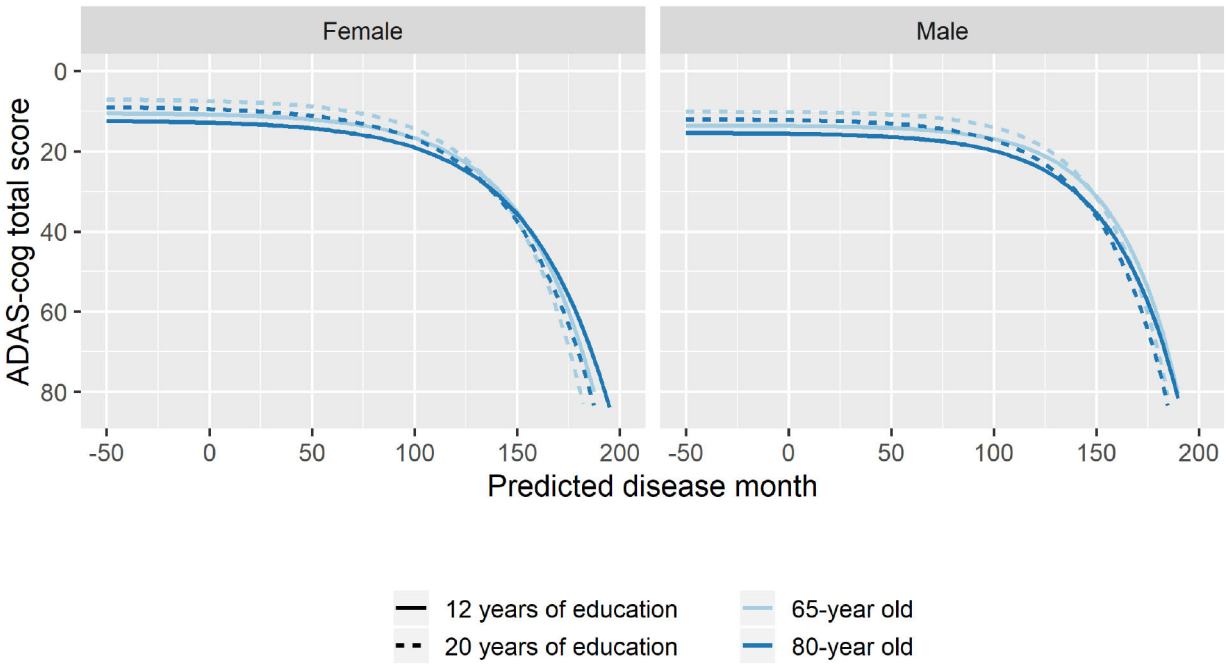


FIGURE 7 | Estimated trajectories for different combination of patient age, sex, and length of education. The trajectories are aligned at predicted month 0 that corresponds to the average cognitive stage of cognitively normal individuals at baseline.

a conventional approach that grouped patients based on baseline symptoms.

The presented model was formulated for one-dimensional outcomes. This choice allowed formulation and implementation of the model in a non-linear mixed-effects modeling framework using maximum likelihood estimation. This in turn enabled modeling of covariate effects on different aspects of disease progression, sophisticated models for random variation in data, and the possibility of taking advantage of the large body of developed statistical methodology for mixed-effects modeling (Pinheiro and Bates, 2006). Because AD is multifaceted, and different measures are sensitive of disease stage and progression at different times during AD, progression models for multivariate outcomes can be more sensitive than univariate models. However, realistic specification of covariate effects, cross-covariance structures, and dependence between random effects for the different outcomes may be very difficult and require a large number of free parameters. While model classes and estimation procedures for similar longitudinal multivariate outcomes have been proposed in other fields (Olsen et al., 2018), existing multivariate models for AD progression generally rely on simple modeling of different sources of variation. Future work should address this gap in the current available methodology: while existing multivariate models can achieve high-quality staging of patients and outcomes with simple noise modeling by taking advantage of the aggregated information across many outcomes (Donohue et al., 2014; Jedynak et al., 2015), they can likely not take advantage of this aggregated information to address specific questions about whether a covariate affects a specific aspect of disease progression.

Age, Sex, Education, and Cognitive Decline

The effect of demographic and socioeconomic factors on disease risk and manifestation in AD has been the subject of much study. In this work, we focused on the combined effects of age, sex, and length of education.

When considered individually, these factors have been observed to result in differences in disease progression. While age is typically considered the major risk factor for developing AD, higher age at AD onset has been observed to be associated with a slower rate of cognitive decline (Gardner et al., 2013; Stanley et al., 2019). Similarly, female sex has been identified as a major risk factor, with almost two-thirds of AD cases being women (Alzheimer's Association., 2018). While this difference has been known for a long time, it has only become apparent more recently that there are sex differences in symptomatology, rate of decline, and possibly in neural anatomy (Ferretti et al., 2018; Oveisgharan et al., 2018). The effects of cognitive reserve on age-related cognitive decline have been the subject of much study (Tucker and Stern, 2011). Cognitive reserve is often studied using educational attainment as an operational proxy for cognitive reserve. It has consistently been found that higher education is associated with increased rate of cognitive decline in incident AD (Teri et al., 1995; Rasmusson et al., 1996; Wilson et al., 2004; Andel et al., 2006; Scarmeas et al., 2006; Musicco et al., 2009; Thomas et al., 2016), with several of these studies also reporting that education is associated with higher baseline cognition.

Differences in cognitive decline are often studied by comparing slopes in statistical models that assume that cognitive decline follows a linear pattern. The argumentation and interpretation around the cognitive reserve model are somewhat

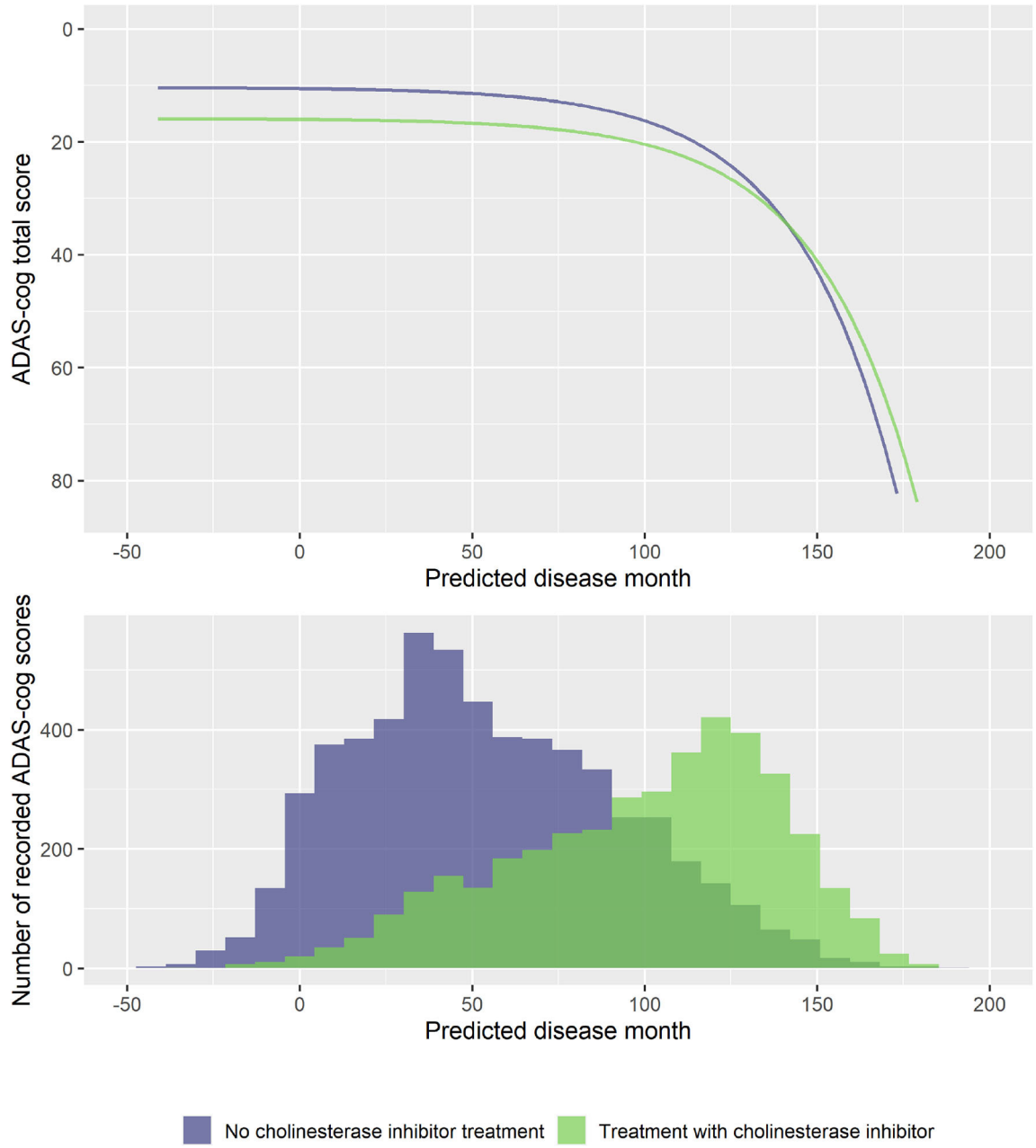


FIGURE 8 | Estimated trajectories for patients with and without cholinesterase inhibitor treatment (top) and corresponding distribution of number of observed ADAS-Cog scores at corresponding predicted disease times. The trajectories are aligned at predicted month 0 that corresponds to the average cognitive stage of cognitively normal individuals that are not treated with cholinesterase inhibitors at baseline.

more sophisticated, but still largely centered on an assumption of a linear rate of decline (e.g., illustrated in Figure 1 in Stern, 2012). The prevailing hypothesis within the field of cognitive reserve research is that, compared to individuals with low cognitive reserve, individuals with high cognitive reserve have higher pre-disease cognitive scores and that their brains tolerate a higher load of neuropathology before cognitive decline is seen. At a sufficiently high level of neuropathology, cognitive ability reaches its floor for all participants. If the timescale of neuropathological buildup is similar across individuals, this suggests that individuals with high cognitive reserve will have to decline a wider range of cognitive scores in a shorter time, thus leading to an accelerated rate of decline (Stern, 2012).

The analyses in the present article clearly illustrate that rate of cognitive decline as measured on ADAS-Cog is not constant but increases over the course of AD. Thus, findings of an increased

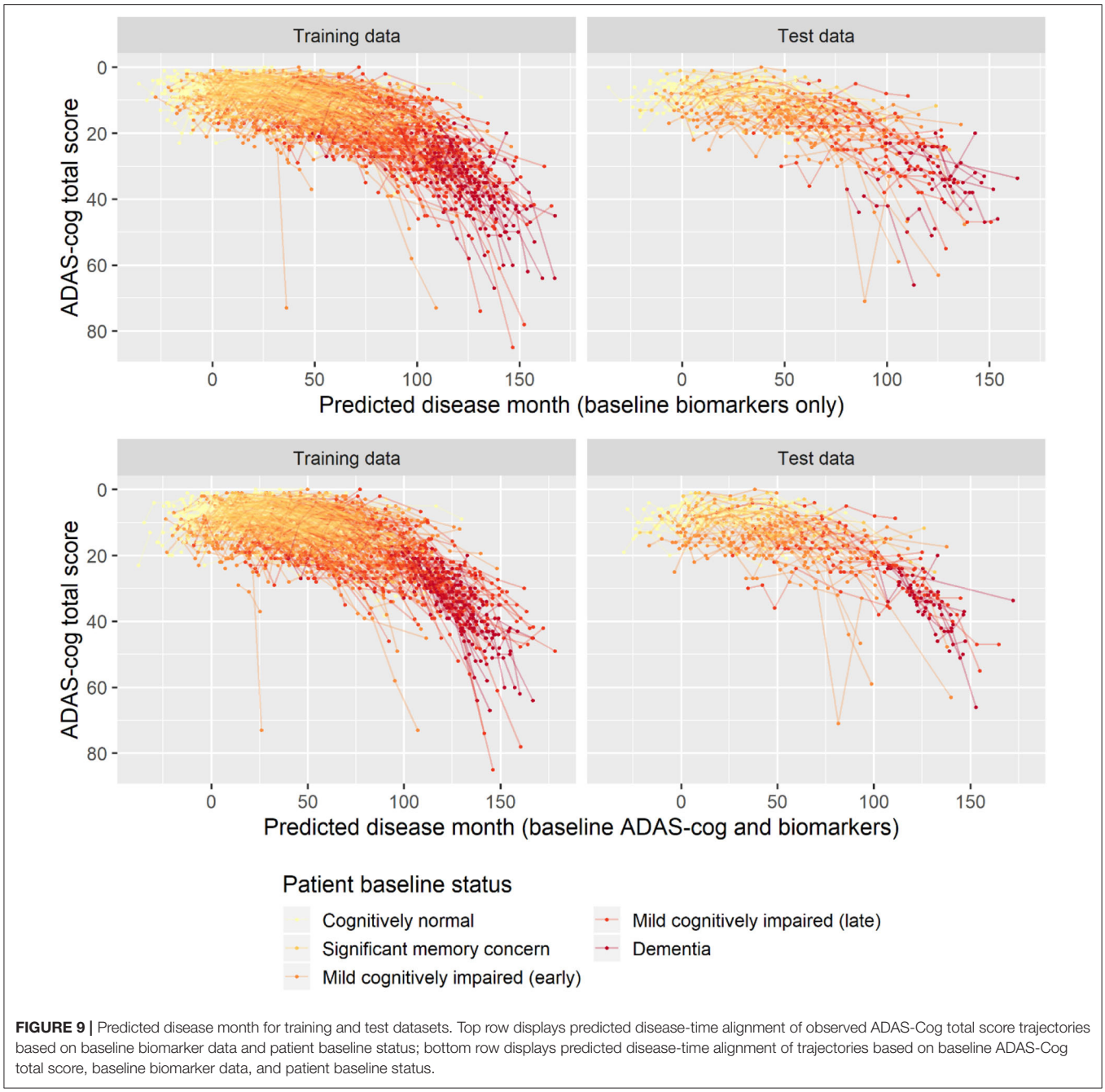


FIGURE 9 | Predicted disease month for training and test datasets. Top row displays predicted disease-time alignment of observed ADAS-Cog total score trajectories based on baseline biomarker data and patient baseline status; bottom row displays predicted disease-time alignment of trajectories based on baseline ADAS-Cog total score, baseline biomarker data, and patient baseline status.

rate of decline in a certain group of patients using slope models could either be because the group of patients has accelerated decline, because they are at a later disease stage, or a combination. The proposed disease progression model seeks to align cognitive trajectories on a disease timeline, and thus it allows one to separate the hypothesized mechanisms of cognitive decline. The best model that adjusted for effects of age at baseline, sex, and length of education on, respectively, disease stage, rate of decline, and cognitive deviation found that all three factors affected all three disease measures except for disease stage, which was not affected by length of education.

When considering the combination of effects (**Figure 7**), the results suggested that higher age at baseline was associated with lower cognition throughout disease time and a slightly reduced rate of decline. Women tended to have not only better pre-disease cognition but also an accelerated decline. Finally, longer education was associated with slightly faster rate of decline and a systematically better cognition throughout the disease.

While these findings are largely consistent with previous findings, they also illustrate that previous results that do not take the long-term disease trajectories into account may be systematically biased. In particular, the fact that highly educated

TABLE 2 | Predictive accuracies of predicted ADAS-Cog total score trajectories for the basic model and the biomarker model both with and without the baseline ADAS-Cog total score.

Model and data	Mean squared error		Median absolute error	
	Training	Test	Training	Test
Basic model (baseline status)	90.0	100.0 (110.6[a])	5.48	4.98 (5.01[a])
BM model (baseline status + BMs)	58.3	65.1 (69.8[a])	4.19	4.21 (4.31[a])
Basic model (baseline status + ADAS-Cog)	78.2[a]	102.7[a]	3.49[a]	3.70[a]
BM model (baseline status + ADAS-Cog + BMs)	53.5[a]	55.1[a]	3.29[a]	3.48[a]

Predictions were censored to the interval [0, 85] to respect the range of the ADAS-Cog scores.
BM, biomarker.
[a]*Baseline ADAS-Cog measurements excluded in computation of prediction errors.*

patients tend to have above-mean cognition throughout the early stages of disease means that they will meet cognitive cutoffs used for inclusion criteria in clinical studies longer into their disease than patients with less education. Because of the accelerated cognitive decline in the later stages of disease, these patients will have a much faster rate of decline when using conventional slope models, but this difference will primarily be due to their later disease stage.

Symptomatic Medications for Alzheimer Disease and Cognitive Decline

Cholinesterase inhibitors have consistently shown a symptomatic benefit in mild to severe dementia due to AD in randomized, double-blind, placebo-controlled trials (Birks, 2006). It has, however, been questioned whether long-term treatment with ChEIs could be harmful (Schneider, 2012). A recent meta-analysis found that AD patients treated with symptomatic treatments had a faster rate of cognitive decline (Kennedy et al., 2018). This could be interpreted as a harmful side effect, but because the included studies were not randomized with respect to symptomatic treatments, such causal link cannot be made. An alternative explanation is simply that ChEIs work—that patients who are being treated at study inclusion have a cognitive benefit that, similarly to higher levels of education, means that they meet inclusion criteria for clinical studies further into their disease. The optimal disease progression model identified in the model search did not include effects of ChEI treatment on rate of decline. Instead, the results of this model showed that patients treated generally had lower cognition compared to untreated patients (which points to confounding by indication; patients are prescribed ChEIs because of their cognitive impairment) and that their progression was slightly delayed.

Biomarker-Based Disease Staging

The final application of the model examined how a patient's biomarker profile at study entry could be used to predict his/her disease stage. Based on training data used for model development, a set of five biomarkers were included in the model. Biomarker profiles considerably improved prediction of future ADAS-Cog trajectories in the unseen validation dataset, and inclusion of baseline ADAS-Cog score further improved the prediction. Among the biomarkers, FDG-PET explained most variation followed by CSF $A\beta_{1-42}/A\beta_{1-40}$ and florbetapir SUVr. Hippocampal volume and plasma NfL explained the least.

This modeling of baseline biomarkers for patients in the earliest stages of disease takes advantage of the long-term follow-up that is unique to ADNI. The modeling essentially relies on hindsight because the patients' disease stage can only be predicted with high reliability once a systematic pattern of cognitive decline has been observed. By using these patterns, the model identified how combinations of biomarkers could be used to predict disease stage. The results of the model suggest that biomarker profiles at a single time point may be used to predict the disease stage of an individual even in the preclinical phases of disease where no clinically detectable cognitive impairment is present.

With further validation, these results can be used to define a space of permissible biomarker profiles to use as inclusion criteria in clinical trials. Such biomarker-based synchronization of patient's disease stage would enable testing a drug in a more homogeneous population. This would in turn greatly increase the power of clinical trials in AD where it is common to see extreme levels of variability in patient trajectories (Cummings et al., 2018; Ballard et al., 2019).

DATA AVAILABILITY STATEMENT

ADNI data can be accessed after accepting its data use agreement and submitting an online application for access. For more information, please see the ADNI website http://adni.loni.usc.edu/.

AUTHOR CONTRIBUTIONS

LR developed the disease progression model, analyzed the data, and wrote the paper.

ACKNOWLEDGMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering,

and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research was providing funds to support ADNI clinical sites in Canada. Private sector contributions were facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization was the Northern California Institute for Research and Education, and the study was coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data were disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2020.00024/full#supplementary-material

REFERENCES

Akaike, H. (1998). "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Aakaike*, eds E. Parzen, K. Tanabe, and G. Kitagawa (New York, NY: Springer), 199–213. doi: 10.1007/978-1-4612-1694-0_15

Alzheimer's Association. (2018). 2018 Alzheimer's disease facts and figures. *Alzheimers Dement.* 14, 367–429. doi: 10.1016/j.jalz.2018.02.001

Andel, R., Vigen, C., Mack, W. J., Clark, L. J., and Gatz, M. (2006). The effect of education and occupational complexity on rate of cognitive decline in Alzheimer's patients. *J. Int. Neuropsychol. Soc.* 12, 147–152. doi: 10.1017/S1355617706060206

Anderson, R. M., Hadjichrysanthou, C., Evans, S., and Wong, M. M. (2017). Why do so many clinical trials of therapies for Alzheimer's disease fail? *Lancet* 390, 2327–2329. doi: 10.1016/S0140-6736(17)32399-1

Ballard, C., Atri, A., Boneva, N., Cummings, J. L., Frölich, L., Molinuevo, J. L., et al. (2019). Enrichment factors for clinical trials in mild-to-moderate Alzheimer's disease. *Alzheimers Dement. Transl. Res. Clin. Intervent.* 5, 164–174. doi: 10.1016/j.trci.2019.04.001

Balsis, S., Unger, A. A., Benge, J. F., Geraci, L., and Doody, R. S. (2012). Gaining precision on the Alzheimer's disease assessment scale-cognitive: a comparison of item response theory-based scores and total scores. *Alzheimers Dement.* 8, 288–294. doi: 10.1016/j.jalz.2011.05.2409

Bateman, R. J., Aisen, P. S., de Strooper, B., Fox, N. C., Lemere, C. A., Ringman, J. M., et al. (2011). Autosomal-dominant Alzheimer's disease: a review and proposal for the prevention of Alzheimer's disease. *Alzheimers Res. Ther.* 3:1. doi: 10.1186/alzrt59

Bateman, R. J., Benzinger, T. L., Berry, S., Clifford, D. B., Duggan, C., Fagan, A. M., et al. (2017). The DIAN-TU next generation Alzheimer's prevention trial: adaptive design and disease progression model. *Alzheimers Dement* 13, 8–19. doi: 10.1016/j.jalz.2016.07.005

Birks, J. S. (2006). Cholinesterase inhibitors for Alzheimer's disease. *Cochrane Database Syst. Rev.* 1:CD005593. doi: 10.1002/14651858.CD005593

Bittner, T., Zetterberg, H., Teunissen, C. E., Ostlund, R. E., Militello, M., Andreasson, U., et al. (2016). Technical performance of a novel, fully automated electrochemiluminescence immunoassay for the quantitation of β-amyloid (1–42) in human cerebrospinal fluid. *Alzheimers Dement.* 12, 517–526. doi: 10.1016/j.jalz.2015.09.009

Cummings, J. L., Atri, A., Ballard, C., Boneva, N., Frölich, L., Molinuevo, J. L., et al. (2018). Insights into globalization: comparison of patient characteristics and disease progression among geographic regions in a multinational Alzheimer's disease clinical program. *Alzheimers Res. Ther.* 10:116. doi: 10.1186/s13195-018-0443-2

Delor, I., Charoin, J. E., Gieschke, R., Retout, S., and Jacqmin, P. (2013). Modeling Alzheimer's disease progression using disease onset time and disease trajectory concepts applied to CDR-SoB scores from ADNI. *CPT Pharmacometr. Syst. Pharmacol.* 2:e78. doi: 10.1038/psp.2013.54

Donohue, M. C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R. G., Raman, R., Gamst, A. C., et al. (2014). Estimating long-term multivariate progression from short-term data. *Alzheimers Dement.*10, 200–410. doi: 10.1016/j.jalz.2013.10.003

Embretson, S. E., and Reise, S. P. (2013). *Item Response Theory*. New York, NY: Psychology Press. doi: 10.4324/9781410605269

Ferretti, M. T., Iulita, M. F., Cavedo, E., Chiesa, P. A., Schumacher Dimech, A., Santuccione Chadha, A., et al. (2018). Sex differences in Alzheimer disease—the gateway to precision medicine. *Nat Rev Neurol.* 14, 457–469. doi: 10.1038/s41582-018-0032-9

Fischl, B. (2012). FreeSurfer. *Neuroimage* 62, 774–781. doi: 10.1016/j.neuroimage.2012.01.021

Gardner, R. C., Valcour, V., and Yaffe, K. (2013). Dementia in the oldest old: a multi-factorial and growing public health issue. *Alzheimers Res. Ther.* 5:27. doi: 10.1186/alzrt181

Gomeni, R, Simeoni, M., Zvartau-Hind, M., Irizarry, M. C., Austin, D., and Gold, M. (2012). Modeling Alzheimer's disease progression using the disease system analysis approach. *Alzheimers Dement.* 8, 39–50. doi: 10.1016/j.jalz.2010.12.012

Hughes, C. P., Berg, L., Danziger, W., Coben, L. A., and Martin, R. L. (1982). A new clinical scale for the staging of dementia. *Br. J. Psychiatr.* 140, 566–572. doi: 10.1192/bjp.140.6.566

Insel, P. S., Weiner, M., Mackin, R. S., Mormino, E., Lim, Y. Y., Stomrud, E., et al. (2019). Determining clinically meaningful decline in preclinical Alzheimer disease. *Neurology* 93, e322–e333. doi: 10.1212/WNL.0000000000007831

Ito, K., Corrigan, B., Zhao, Q., French, J., Miller, R., Soares, H., et al. (2011). Disease progression model for cognitive deterioration from Alzheimer's disease neuroimaging initiative database. *Alzheimers Dement.* 7, 151–160. doi: 10.1016/j.jalz.2010.03.018

Jack, C. R., Bennett, D. A., Blennow, K., Carrillo, M. C., Dunn, B., Haeberlein, S. B., et al. (2018). NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement.* 14, 535–562. doi: 10.1016/j.jalz.2018.02.018

Jedynak, B.M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B.T., et al. (2012). A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease neuroimaging initiative cohort. *Neuroimage* 63, 1478–1486. doi: 10.1016/j.neuroimage.2012.07.059

Jedynak, B. M., Liu, B., Lang, A., Gel, Y., Prince, J. L., and Alzheimer's Disease Neuroimaging Initiative. (2015). A computational method for computing an Alzheimer's disease progression score; experiments and validation with the ADNI data set. *Neurobiol. Aging* 36, S178–S184. doi: 10.1016/j.neurobiolaging.2014.03.043

Kennedy, R. E., Cutter, G. R., Fowler, M. E., and Schneider, L. S. (2018). Association of concomitant use of cholinesterase inhibitors or memantine with cognitive decline in alzheimer clinical trials: a meta-analysis. *JAMA Netw Open* 1:e184080. doi: 10.1001/jamanetworkopen.2018.4080

Koval, I., Schiratti, J. B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., et al. (2018). Spatiotemporal propagation of the cortical atrophy: population

and individual patterns. *Front. Neurol.* 9:235. doi: 10.3389/fneur.2018.00235

Landau, S. M., Fero, A., Baker, S. L., Koeppe, R., Mintun, M., Chen, K., et al. (2015). Measurement of longitudinal β-amyloid change with 18F-florbetapir PET and standardized uptake value ratios. *J. Nucl. Med.* 56, 567–574. doi: 10.2967/jnumed.114.148981

Landau, S. M., Harvey, D., Madison, C. M., Koeppe, R. A., Reiman, E. M., Foster, N. L., et al. (2011). Alzheimer's disease neuroimaging initiative. Associations between cognitive, functional, and FDG-PET measures of decline in AD and MCI. *Neurobiol. Aging* 32, 1207–1218. doi: 10.1016/j.neurobiolaging.2009.07.002

Lavielle, M., and Aarons, L. (2016). What do we mean by identifiability in mixed effects models? *J. Pharmacokinet. Pharmacodyn.* 43, 111–122. doi: 10.1007/s10928-015-9459-4

Li, D, Iddi, S., Thompson, W. K., Rafii, M. S., Aisen, P. S., and Donohue, M. C. (2018). Alzheimer's disease neuroimaging initiative. Bayesian latent time joint mixed-effects model of progression in the Alzheimer's disease neuroimaging initiative. *Alzheimers Dement.* 10, 657–668. doi: 10.1016/j.dadm.2018.07.008

Lindstrom, M. L., and Bates, D. M. (1990). Nonlinear mixed effects models for repeated measures data. *Biometrics* 673–687. doi: 10.2307/2532087

Louis, M., Couronne, R., Koval, I., Charlier, B., and Durrleman, S. (2019). "Riemannian geometry learning for disease progression modelling," in *International Conference on Information Processing in Medical Imaging* (Cham: Springer), 542–553. doi: 10.1007/978-3-030-20351-1_42

Mattsson, N., Cullen, N. C., Andreasson, U., Zetterberg, H., and Blennow, K. (2019). Association between longitudinal plasma neurofilament light and neurodegeneration in patients with Alzheimer disease. *JAMA Neurol.* 76, 791–799. doi: 10.1001/jamaneurol.2019.0765

Mohs, R. C., Knopman, D., Petersen, R. C., Ferris, S. H., Ernesto, C., Grundman, M., et al. (1997). Development of cognitive instruments for use in clinical trials of antidementia drugs: additions to the Alzheimer's disease assessment scale that broaden its scope. *Alzheimer Dis. Assoc. Disord.* 11(Suppl. 2), S13–S21. doi: 10.1097/00002093-199700112-00003

Musicco, M., Palmer, K., Salamone, G., Lupo, F., Perri, R., Mosti, S., et al. (2009). Predictors of progression of cognitive decline in Alzheimer's disease: the role of vascular and sociodemographic factors. *J. Neurol.* 256, 1288–1295. doi: 10.1007/s00415-009-5116-4

Olsen, N. L., Markussen, B., and Raket, L. L. (2018). Simultaneous inference for misaligned multivariate functional data. *J. R. Stat. Soc. C* 67, 1147–1176. doi: 10.1111/rssc.12276

Oveisgharan, S, Arvanitakis, Z., Yu, L., Farfel, J., Schneider, J. A., and Bennett, D. A. (2018). Sex differences in Alzheimer's disease and common neuropathologies of aging. *Acta Neuropathol.* 136, 887–900. doi: 10.1007/s00401-018-1920-1

Oxtoby, N. P., Young, A. L., Cash, D. M., Benzinger, T. L. S., Fagan, A. M., Morris, J. C., et al. (2018). Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 141, 1529–1544. doi: 10.1093/brain/awy050

Palmqvist, S., Zetterberg, H., Mattsson, N., Johansson, P., Minthon, L., et al. (2015). Detailed comparison of amyloid PET and CSF biomarkers for identifying early Alzheimer disease. *Neurology* 85, 1240–1249. doi: 10.1212/WNL.0000000000001991

Pfeffer, R. I., Kurosaki, T. T., Harrah C. H. Jr., Chance, J. M., and Filos, S. (1982). Measurement of functional activities in older adults in the community. *J. Gerontol.* 37, 323–329. doi: 10.1093/geronj/37.3.323

Pinheiro, J., Bates, D., DebRoy, D., Sarkar, D., and R Core Team. (2019). *nlme: Linear and Nonlinear Mixed Effects Models*. R package version 3.1-141.

Pinheiro, J., and Bates, D. M. (2006). *Mixed-Effects Models in S and S-PLUS*. Springer Science and Business Media.

R Core Team (2020). *R: A Language and Environment for Statistical Computing*. Vienna. Available online at: https://www.R-project.org/

Raket, L. L. (2020). *progmod*. Available online at: https://github.com/larslau/progmod

Raket, L. L., Sommer, S., and Markussen, B. (2014). A nonlinear mixed-effects model for simultaneous smoothing and registration of functional data. *Pattern Recogn. Lett.* 38, 1–7. doi: 10.1016/j.patrec.2013.10.018

Ramsay, J. O. (1988). Monotone regression splines in action. *Stat. Sci.* 3, 425–441. doi: 10.1214/ss/1177012761

Rasmusson, D. X., Carson, K. A., Brookmeyer, R., Kawas, C., and Brandt, J. (1996). Predicting rate of cognitive decline in probable Alzheimer's disease. *Brain Cogn.* 31, 133–147. doi: 10.1006/brcg.1996.0038

Ryman, D. C., Acosta-Baena, N., Aisen, P. S., Bird, T., Danek, A., Fox, N. C., et al. (2014). Symptom onset in autosomal dominant Alzheimer disease: a systematic review and meta-analysis. *Neurology* 83, 253–260. doi: 10.1212/WNL.0000000000000596

Samtani, M. N., Farnum, M., Lobanov, V., Yang, E., Raghavan, N., DiBernardo, A., et al. (2012). An improved model for disease progression in patients from the Alzheimer's disease neuroimaging initiative. *J. Clin. Pharmacol.* 52, 629–644. doi: 10.1177/0091270011405497

Scarmeas, N., Albert, S. M., Manly, J. J., and Stern, Y. (2006). Education and rates of cognitive decline in incident Alzheimer's disease. *J. Neurol. Neurosurg. Psychiatr.* 77, 308–316. doi: 10.1136/jnnp.2005.072306

Schiratti, J.-B., Allassonniere, S., Colliot, O., and Durrleman, S. (2017). A Bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *J. Mach. Learn. Res.* 18, 4840–4872. Available online at: http://www.jmlr.org/papers/volume18/17-197/17-197.pdf

Schneider, L. S. (2012). Could cholinesterase inhibitors be harmful over the long term? *Int Psychogeriatr.* 24, 171–174. doi: 10.1017/S1041610211002389

Schwarz, G. (1978). Estimating the dimension of a model. *Ann. Stat.* 6, 461–464. doi: 10.1214/aos/1176344136

Stanley, K., Whitfield, T., Kuchenbaecker, K., Sanders, O., Stevens, T., and Walker, Z. (2019). Rate of cognitive decline in Alzheimer's disease stratified by age. *J. Alzheimers Dis.* 69, 1153–1160. doi: 10.3233/JAD-181047

Stern, Y. (2012). Cognitive reserve in ageing and Alzheimer's disease. *Lancet Neurol.* 11, 1006–1012. doi: 10.1016/S1474-4422(12)70191-6

Teri, L., McCurry, S. M., Edland, S. D., Kukull, W. A., and Larson, E. B. (1995). Cognitive decline in Alzheimer's disease: a longitudinal investigation of risk factors for accelerated decline. *J. Gerontol. A Biol. Sci. Med. Sci.* 50, M49–M55. doi: 10.1093/gerona/50A.1.M49

Thomas, R. G., Albert, M., Petersen, R. C., and Aisen, P. S. (2016). Longitudinal decline in mild-to-moderate Alzheimer's disease: analyses of placebo data from clinical trials. *Alzheimers Dement.* 12, 598–603. doi: 10.1016/j.jalz.2016.01.002

Tucker, A. M., and Stern, Y. (2011). Cognitive reserve in aging. *Curr. Alzheimer Res.* 8, 354–360. doi: 10.2174/156720511795745320

Wang, G., Coble, D., McDade, E. M., Hassenstab, J., Fagan, A. M., Benzinger, T. L., et al. (2019). Dominantly inherited Alzheimer network. Staging biomarkers in preclinical autosomal dominant Alzheimer's disease by estimated years to symptom onset. *Alzheimers Dement.* 15, 506–514. doi: 10.1016/j.jalz.2018.12.008

Wilson, R. S., Li, Y., Aggarwal, N. T., Barnes, L. L., McCann, J. J., Gilley, D. W., et al. (2004). Education and the course of cognitive decline in Alzheimer disease. *Neurology* 63, 1198–1202. doi: 10.1212/01.WNL.0000140488.65299.53

Yang, E., Farnum, M., Lobanov, V., Schultz, T., Verbeeck, R., Raghavan, N., et al. (2011). Quantifying the pathophysiological timeline of Alzheimer's disease. *J. Alzheimers Dis.* 26, 745–753. doi: 10.3233/JAD-2011-110551

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137, 2564–2577. doi: 10.1093/brain/awu176

Conflict of Interest: LR was employed by company H. Lundbeck A/S.

frontiers in Big Data

ORIGINAL RESEARCH
published: 04 May 2021
doi: 10.3389/fdata.2021.637724

Association vs. Prediction: The Impact of Cortical Surface Smoothing and Parcellation on Brain Age

Yashar Zeighami [1,2]* *and Alan C. Evans* [1,2]

[1] *Montreal Neurological Institute, McGill University, Montreal, QC, Canada,* [2] *Ludmer Centre for Neuroinformatics and Mental Health, McGill University, Montreal, QC, Canada*

Association and prediction studies of the brain target the biological consequences of aging and their impact on brain function. Such studies are conducted using different smoothing levels and parcellations at the preprocessing stage, on which their results are dependent. However, the impact of these parameters on the relationship between association values and prediction accuracy is not established. In this study, we used cortical thickness and its relationship with age to investigate how different smoothing and parcellation levels affect the detection of age-related brain correlates as well as brain age prediction accuracy. Our main measures were resel numbers—resolution elements—and age-related variance explained. Using these common measures enabled us to directly compare parcellation and smoothing effects in both association and prediction studies. In our sample of $N = 608$ participants with age range 18–88, we evaluated age-related cortical thickness changes as well as brain age prediction. We found a negative relationship between prediction performance and correlation values for both parameters. Our results also quantify the relationship between delta age estimates obtained based on different processing parameters. Furthermore, with the direct comparison of the two approaches, we highlight the importance of correct choice of smoothing and parcellation parameters in each task, and how they can affect the results of the analysis in opposite directions.

Keywords: brain aging, cortical thickness, prediction, delta age, smoothing, parcellation, association

OPEN ACCESS

Edited by:
Holger Fröhlich,
University of Bonn, Germany

Reviewed by:
Lingzhong Fan,
Institute of Automation, Chinese Academy of Sciences (CAS), China
Ye Wu,
University of North Carolina at Chapel Hill, United States
Yu Zhang,
Zhejiang Lab, China

***Correspondence:**
Yashar Zeighami
yashar.zeighami@mcgill.ca

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *04 December 2020*
Accepted: *06 April 2021*
Published: *04 May 2021*

Citation:
Zeighami Y and Evans AC (2021) Association vs. Prediction: The Impact of Cortical Surface Smoothing and Parcellation on Brain Age. Front. Big Data 4:637724. doi: 10.3389/fdata.2021.637724

INTRODUCTION

From a biological standpoint, aging is defined by the structural and functional alterations in living organisms (López-Otín et al., 2013). Traditionally, brain imaging studies have used neuroimaging data to find associations between age and tissue alterations across brain areas, using chronological age as the ground truth (Lemaître et al., 2005; Curiati et al., 2009; Takahashi et al., 2011; Ziegler et al., 2012; Booth et al., 2013; Hu et al., 2014). However, biological age might vary between individuals with identical chronological age as well as across different tissues within the same person (Horvath, 2013). To non-invasively measure the biological age of the brain, neuroimaging data is used to predict age. The difference between predicted age and chronological age is then defined as "delta" or brain age gap estimate i.e., "BrainAGE" to compare the subjects' chronological age with the

predicted brain age in a given reference population (Franke et al., 2012; Cole and Franke, 2017; Franke and Gaser, 2019; Smith et al., 2019).

Both age related brain alterations and delta age have been studied and used extensively in the neuroimaging literature. Age association studies translate and generalize easily across different datasets. These association studies are applied across brain regions and can distinguish the differential effect of age on different brain areas (Storsve et al., 2014). Furthermore, they directly relate to biological measures and mechanistic changes in the brain (Khundrakpam et al., 2015). More recently, it has been recognized that association studies are prone to overfitting and more studies focus on prediction as the main goal of the study (Yarkoni and Westfall, 2017; Bzdok et al., 2020). Brain age studies (i.e., age prediction studies based on neuroimaging data) rely on modeling and prediction accuracy. This goal is generally achieved by using a feature set that can capture the variability between and within subjects. On the other hand, prediction tasks face a trade-off between a more accurate whole brain model with no regional specificity vs. a model with lower accuracy and increased spatial resolution (Cole and Franke, 2017; Franke and Gaser, 2019). This limitation also results in a more indirect relationship between delta age and other phenotypes without a direct mechanistic and biological model. Nonetheless, the difference between brain age and chronological age is associated with cognitive decline (Gaser et al., 2013), predisposition to neuropsychiatric and neurodegenerative disorders (Kaufmann et al., 2019), and mortality (Cole et al., 2018). While evidence supports the application of delta age as a valuable measure to study aging in health and disease, it has been criticized due to its reliance on prediction accuracy (i.e., more accurate models result in lower delta values) (Cole and Franke, 2017).

The results of both association studies and delta estimation studies are impacted by processing steps such as data normalization, spatial resolution, and parcellation level (i.e., size of the parcels) of the analysis. Most association studies use smoothing to (i) normalize the distributions of cortical thickness across subjects, (ii) minimize registration and anatomical misalignment across subjects, (iii) reduce measurement noise, and (iv) increase statistical power (Worsley et al., 1999; Lerch and Evans, 2005; Lerch et al., 2006; Zhao et al., 2013). These advantages are gained at the cost of losing individual variability and spatial resolution. In fact, smoothing has been studied and optimized for best performance in association studies, using simulation as well as in real datasets. The smoothing level has been proposed as a dimension within the parameter space in the association analysis that needs to be searched for the given statistical contrast (Lerch and Evans, 2005; Zhao et al., 2013).

Brain age prediction studies have been conducted with various levels of data smoothing. Moreover, these studies rely on various dimension reduction techniques, brain parcellations, or a combination of the two approaches for feature extraction (Franke and Gaser, 2019; Smith et al., 2019). The optimal parcellation for a given task is an open research topic and it can vary between studies (Gorgolewski et al., 2016; Eickhoff et al., 2018; Salehi et al., 2020). While some studies have predicted brain age with multiple parcellation resolutions (Khundrakpam et al., 2015; Lewis J. D. et al., 2019), others have used a predetermined number of parcels. However, the effect of smoothing and parcellation in brain age prediction is not studied systematically. Furthermore, these changes in prediction accuracy also affect the delta estimate (i.e., the variable of interest), and it is not clear whether the delta estimates are robust or sensitive toward these initial choices.

In this study, we used cortical thickness as the brain measure of interest and examined the effect of smoothing and parcellation level on both brain associations with age and brain age prediction. Using different levels of parcellation and smoothing, we projected brain measures onto a lower dimension data representation space and investigated how this mapping affects the derived associations and predictions. We further examined the relationship between the two approaches. Finally, we examined how delta age estimates alter based on different smoothing and parcellation levels.

METHODS

Data

Data used in this study included subjects with T1-weighted MRI data available from the second stage of the Cambridge Centre for Ageing and Neuroscience (CamCAN, https://www.cam-can.org/index.php?content=dataset) dataset, described in more detail in Shafto et al. (2014) and Taylor et al. (2017). Subjects were screened for neurological and psychiatric conditions and those with such underlying disorders were excluded from the study.

MRI Acquisition

T1-weighted MRIs were acquired on a 3T Siemens TIM Trio, with a 32 channel head-coil using a 3D magnetization-prepared rapid gradient echo (MPRAGE) sequence (TR = 2,250 ms, TE = 2.99 ms, TI = 900 ms; FA = 9 deg; FOV = 256 × 240 × 192 mm; 1 mm isotropic; GRAPPA = 2; TA = 4 min 32 s). For detailed acquisition parameters see: https://camcan-archive.mrc-cbu.cam.ac.uk/dataaccess/pdfs/CAMCAN700_MR_params.pdf.

MRI Processing

We used CIVET 2.1.1 (http://www.bic.mni.mcgill.ca/ServicesSoftware/CIVET, release December 2019), a fully automated structural image analysis pipeline developed at the Montreal Neurological Institute, to perform surface extraction and cortical thickness estimation. Briefly, each subject's T1-weighted MRI is corrected for non-uniformity artifacts using the N3 algorithm (N3 distance = 125 mm) (Sled et al., 1998) and linearly registered to stereotaxic MNI152 space (voxel resolution = 0.5 mm) (Collins et al., 1994). The brain is extracted and undergoes tissue classification into three classes: white matter (WM) tissue, gray matter (GM) tissue, and cerebrospinal fluid (CSF) (Zijdenbos et al., 2002; Tohka et al., 2004). White and gray matter surfaces are extracted using the marching cube algorithm and constrained Laplacian-based automated segmentation with proximities (CLASP) algorithms, respectively (MacDonald et al., 2000; Kabani et al., 2001; Kim et al., 2005). Using the extracted surfaces, cortical thickness is measured as the distance between the white and gray cortical surfaces using the Laplace's equation (Jones et al., 2000). For blurring, a surface-based diffusion

smoothing kernel (not to be confused with volumetric kernels) is used, which generalizes Gaussian kernel smoothing and applies it to the curved cortical surfaces (Chung et al., 2002). We applied 6 different smoothing levels with FWHM = 0, 5, 10, 20, 30, and 40 mm. Cortical thickness was measured across the cortical surface for 81,924 vertices (40,962 vertices per hemisphere). The results underwent visual inspection, specifically subjects with major errors in extracted pial and gray–white surfaces were excluded.

Cortical Parcellations

We used the Schaefer functional MRI parcellations (Schaefer et al., 2018), a data-driven atlas based on the widely used seven large-scale functional network parcellations by Thomas Yeo et al. (2011). We used Schaefer parcellation with 100, 200, 400, and 1,000 regions (referred to as parcellation levels). All atlases were registered to the MNI cortical surface template and used in the MNI space (Lewis L. B. et al., 2019). Cortical thickness measurements with different smoothing levels were averaged across these parcellations. These parcellation based measures of cortical thickness were used alongside vertex-wise measurements to examine the interaction between the effect of brain parcellation averaging and smoothing on statistical associations as well as brain age prediction accuracies.

Cortical Resels and Effective Smoothing

In order to compare the findings between smoothing levels and different parcellations, first all obtained cortical thickness were projected to the brain surface. We used the number of resels (i.e., resolution elements) as the measure of interest, since it takes the statistical dependence of the brain map into consideration and is independent of the analysis resolution (at least from a theoretical standpoint) (Worsley et al., 1992, 1999; Worsley, 1996; Lerch et al., 2006). Using the statistical maps between aging and cortical thickness, we estimated the number of resels for each smoothing and parcellation level and used it to quantify the similarity between these conditions. Resels are the number of resolution elements approximated for a given search space [i.e., D(S2), S2 = brain surface] and a given smoothness level FWHM. While the effective FWHM measure varies across brain areas, we defined the overall effective smoothness of the brain map as the square root of the surface search space divided by the number of resels estimated across brain areas (Hayasaka et al., 2004). For the purpose of the current study, the main statistical maps considered are the linear associations between cortical thickness and the chronological age of the participants. All analysis were performed using SurfStat toolbox https://www.math.mcgill.ca/keith/surfstat/ (see **Supplementary Methods** for further details).

Statistical Methods

To examine the effect of the smoothing and parcellations, mean (μ) and standard deviation (σ) of cortical thickness for each vertex/parcel was calculated across the population. The coefficient of variation (CV), $CV = \frac{\sigma}{\mu}$, was used as the main measure of variability. The CV was averaged across the 7 main cytoarchitectural brain regions (von Economo and Koskinas, 1927) in order to examine the effect of parcellation and smoothing across major cytoarchitectural regions and identify any differential impact on a given brain region. Finally, to measure the association between chronological age and cortical thickness across lifespan, correlation coefficient (r) for each vertex/region was calculated. Variance explained (r^2) was used to visualize the results.

Brain Age Prediction

We used principal component analysis (PCA), a singular value decomposition based data factorization method, as the dimensionality reduction approach for our predictive variables (i.e., cortical thickness data) (Smith et al., 2019). This approach allowed us to use the same number of features across parcellation levels and smoothing kernels and therefore made it possible to compare model performance across these conditions. Our analysis for each condition included 1 to 100 first principal components as features to study different levels of dimensionality reduction. Hundred is used as the maximum possible number of independent components for the lowest number of parcels (i.e., Schaefer 100). To predict brain age, we used linear regression as the main prediction model, and to ensure generalizability and avoid overfitting, we used 10-fold cross validation. Finally, to increase robustness, results averaged over 100 repetitions are reported, however as discussed these repetitions are not necessary and had no impact on the conclusions. Root-mean-squared error (RMSE) was used as the natural cost function for linear regression models. Mean absolute error (MAE) and correlation between chronological age and predicted age (two other common error metrics in the age prediction literature; Franke and Gaser, 2019; Franke et al., 2020) are also reported in the **Supplementary Materials**. Finally, we have repeated the same procedure using a support vector machine (SVM) regression method with linear kernel as well as linear regression models with lasso and ridge regularization (results reported in the **Supplementary Materials**).

The Relationship Between Brain Age Prediction and Age Related Brain Association

To compare brain age association and age prediction, we used the variance explained between dependent and independent variables as the main measure of interest for each model. This common measure enabled us to quantify the two analyses in relation to each other. Furthermore, we examined how the number of resels affects whole brain associations with age as well as brain age prediction. To translate the age prediction error into variance explained, we used the predictive features in a linear model, calculating the variance explained for age using adjusted R^2. Finally, the overfitting bias between the variance explained (i.e., adjusted R^2) using this linear model and the cross validated prediction (i.e., r^2 between predicted age and chronological age) is reported.

Delta Age

The main goal of brain age prediction studies is to calculate the deviation from chronological age based on the population

norm, also known as delta age. Here, we examined the effect of smoothing and parcellation on delta age estimation:

$$\mathbf{Y}=\mathbf{X}\beta_1-\delta_1 \quad \delta_1=\mathbf{X}\beta_1-\mathbf{Y}$$

where Y denotes chronological age, X denotes the neuroimaging features, and δ_1 denotes the difference between predicted and chronological age. δ_1 is a measure of brain state/health compared to the population with similar chronological age, and is used to study the predisposition to different brain disorders as well as individual cognitive abilities in neuroimaging literature.

δ_1 being residual of the predictive model is by definition: (1) orthogonal to the predictive measures *X*, and in the case of linear models (2) correlated with the output *Y*(i.e., chronological age) (Le et al., 2018; Liang et al., 2019; Smith et al., 2019). The first feature is unfavorable, since we are interested in brain related discrepancy between chronological and predicted age. The lack of association between δ_1 and brain features predicting age undermines the interpretability of δ_1 in relation to brain measures. The second property is also an adverse feature, since it makes it difficult to distinguish the effect of the chronological age from the additional biological delta age (due to their collinearity). Therefore, in the current study, we followed the recommendation of smith and colleagues (Smith et al., 2019) and used δ_2, the orthogonalized residuals against chronological age:

$$\delta_2=\delta_1-\mathbf{Y}\beta_2$$

δ_2 is then used as the main measure of interest for association across conditions. The results for δ_1 is provided in the **Supplementary Materials**. Note that δ_2 is also consistently calculated using the same 10-fold cross validation with 100 repeats as δ_1, however as discussed these repetitions are not necessary and had no impact on the conclusions. All statistical and prediction analyses were performed using MATLAB 2018a.

RESULTS

Cortical Thickness Aging, Resels, and Practical Smoothness

The parcellations have a considerable impact on the number of resels and function as region-based smoothing kernels applied across the brain (**Figure 1A**). This change in the number of resels affects the statistical power and the association as well as prediction results. Across parcellation levels from 100 to 1,000, the effect of the smaller smoothing kernels with FWHM 0–10 mm is negligible, while applying larger kernels reduces the number of resels dramatically. This equivalency plot also suggests that at the vertex level, the smoothing kernels act as a non-specific parcellation (from an anatomical perspective) across the brain.

Cortical Thickness Variability

While keeping the mean cortical thickness measure intact, smoothing resulted in underestimation of the cortical thickness in the gyri areas and overestimation in the sulci regions. The results are similar for parcellations in the case of uniformly sized parcels and balanced inclusion of gyri and sulci in each parcel (both criteria are met in Schaefer parcellations). Cortical thickness variability (i.e., CV) reduces significantly both as a result of using greater smoothing and larger parcels (**Figure 2A**).

The association cortices have the lowest CV across resolutions and parcellations. Both smoothing and parcellation result in the highest decrease in CV in limbic and insular cortices, while primary sensory and motor areas show the lowest change (**Figure 2B**). The results are shown for 0 mm smoothing across parcellations. The greatest change occurs with increasing the FWHM value from 10 to 20 mm, as well as decreasing the number of parcels from 400 to 200. The results for different smoothing kernels at vertex level were also similar (**Supplementary Figure 1**).

Statistical Association Between Cortical Thickness and Aging

Figure 3A shows the association between age and cortical thickness (using variance explained r^2), calculated for each voxel/parcel for all conditions, after Bonferroni correction to account for the multiple comparisons at each level. The correlation increases with greater smoothing and larger parcels. Changing smoothing kernel size results in the highest variability in the correlation distribution across the brain at vertex level resolution (**Figure 3B**, top panel), whereas smoothing doesn't change the results within Schaefer 100 parcellations (**Figure 3B**, bottom panel). The same pattern is evident between parcellation levels with 0 mm smoothing showing the highest variability, and 40 mm smoothing with lowest variability across parcellations. These findings are further explained with reference to the number of resels and effective smoothing in section The Relationship Between Prediction and Association. Finally, while present across all brain areas, the variability between correlation maps is the highest within association cortices, primary motor, and insular cortex.

Brain Age Prediction Based on Cortical Thickness

For age prediction, vertex-level data outperformed all parcellation-based data using the same (or a smaller) number of principal components as predictive features. The accuracy was also higher for lower smoothing kernel size. However, this effect was more pronounced for FWHMs >10 mm, and the results for FWHM values of 0, 5, and 10 mm showed a very similar performance in the vertex-level analysis. A similar pattern was present within each parcellation level. The best performing models (i.e., 0 and 5 mm smoothed vertex-wise), reach their minimum error using the first 20–30 principal components as features in the prediction model (i.e., a sample to feature ratio of 28–18). The pattern was similar for MAE and correlation between predicted age and chronological age (**Supplementary Figures 2, 3**).

The Relationship Between Prediction and Association

As expected, there was a negative relationship between the overall correlation between age and cortical thickness across

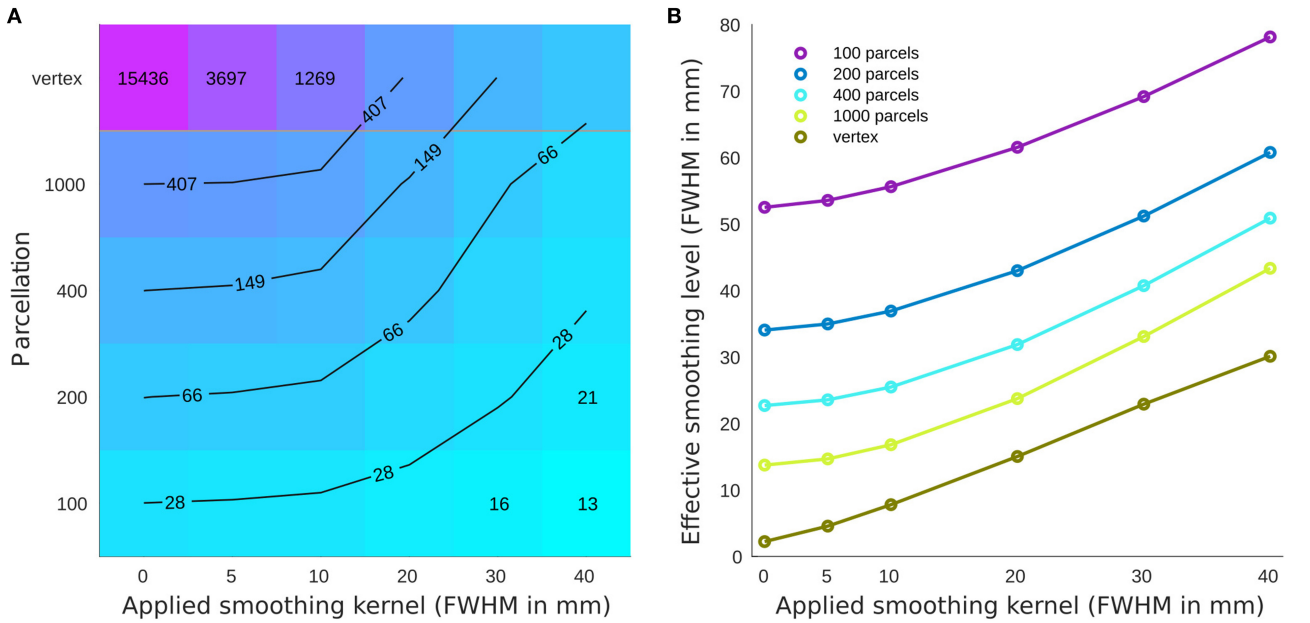


FIGURE 1 | Number of resels and effective smoothing for cortical thickness association with age. **(A)** Number of resels estimated for different parcellations/smoothing pairs. The lines show the interpolated iso-response values. **(B)** Effective smoothing based on the number of resels for each condition. The results show the initial effective smoothing as a result of parcellation with additional smoothing with applied smoothing kernels.

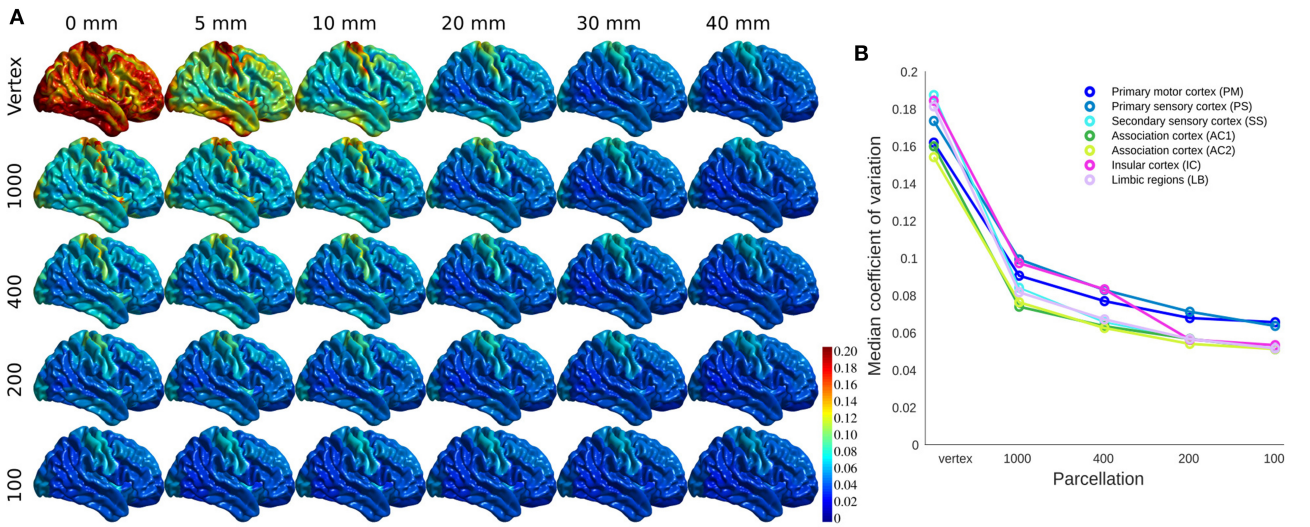


FIGURE 2 | The coefficient of variation (CV) of cortical thickness across population. **(A)** CV projected across brain vertices for each parcellations/smoothing pair. **(B)** CV shown at 0 mm smoothing level for each cytoarchitectural region across parcellation resolutions.

brain regions (measured by median r^2) and the number of resels within each condition (**Figure 5A**). Interestingly, we found a positive association between the number of resels and the overall ability of cortical thickness features to explain the variance of chronological age (as measured by adjusted R^2 of the linear model) shown in **Figure 5B**. These results suggest that the higher number of resels results in lower correlation values, but since resels are independent based on their relationship with age, they can explain different modes of chronological age within the population (hence the higher adjusted R^2), whereas, in conditions with lower resel numbers (i.e., higher smoothing and larger parcels) the correlation values are higher but homogenous across the brain and therefore explain a lower proportion of the age variance.

Finally, there was a strong linear relationship between (i) the overall variance explained (adjusted R^2) using a linear model with

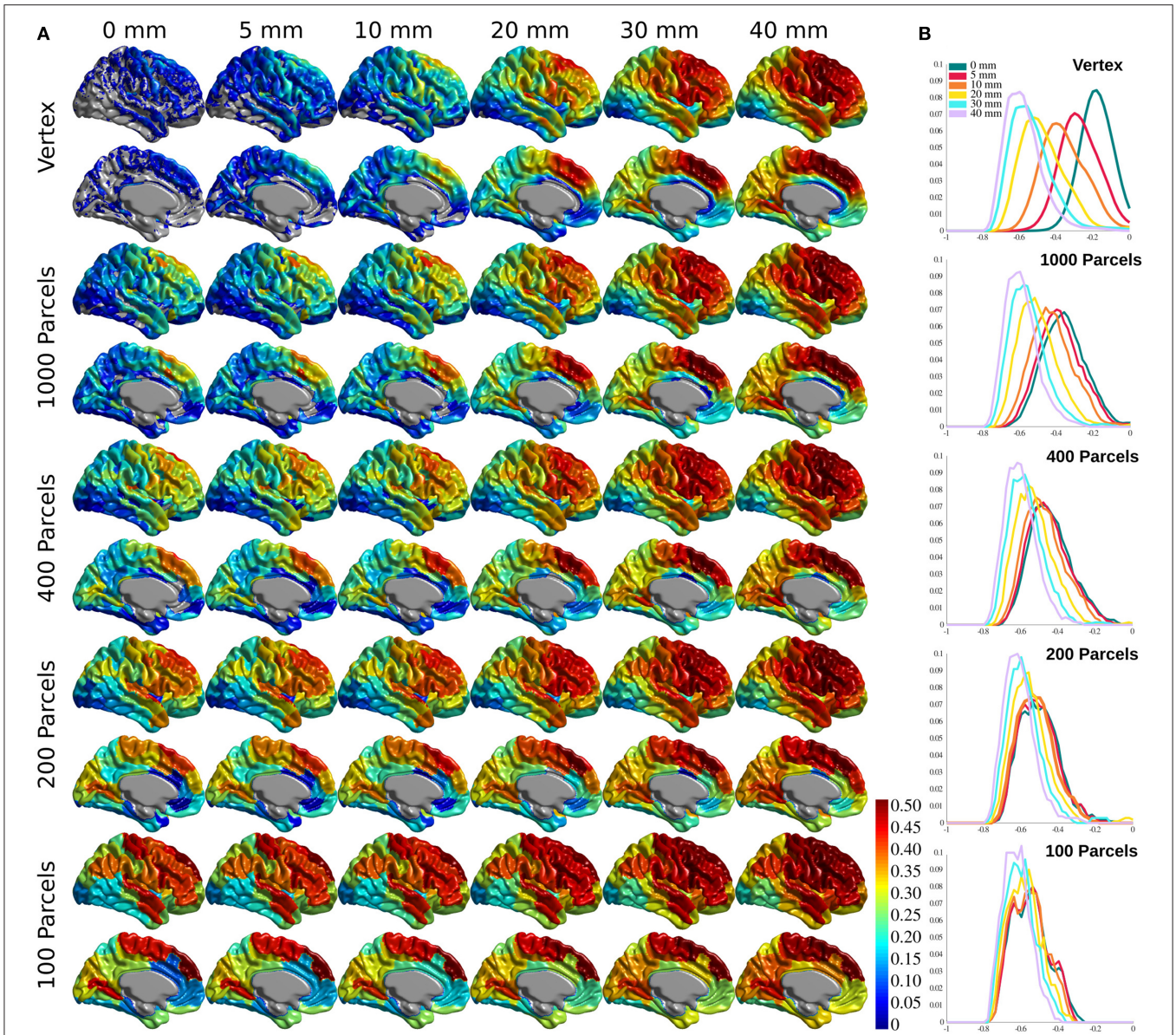


FIGURE 3 | Cortical thickness variance explained by age (r^2). **(A)** Cortical thickness variance explained by age (r^2) for each vertex/parcel across smoothing/parcellation conditions. **(B)** Histograms for correlation values for each parcellation conditions, grouped by smoothing level.

age as dependent variable and PCs as independent variable and (ii) the predictive performance of the linear regression model, with a bias due to overfitting in the linear model (**Figure 5C**). **Figure 5D** shows the overfitting bias of the adjusted R^2 compared to the cross-validated prediction, as a function of the number of features in the model. Taken together, these results explain the opposing directions between correlation results and prediction accuracy across parcellation and smoothing conditions.

The Effect of Smoothing and Parcellation on the Estimation of Brain Age Delta

In this section, we present δ_2 age prediction accuracy results with 10-fold cross validation. The prediction accuracy based on the modified δ_2 is presented in **Figure 6**. One of the main assumptions in age prediction studies is that delta age measured in different studies using different processing parameters are similar and can be interpreted as the same measure. We have examined the relationship between the optimal δ_2 across different parcellations and smoothing kernels (**Figure 7**). These results demonstrate the degree of sensitivity of δ_2 as a function of our choice for parcellation and smoothing kernel. While there is high correlation for large smoothing kernels (20–40 mm) as well as lower number of parcels, these conditions have the lowest prediction accuracies. The correlations between these conditions and higher accuracy conditions (i.e., vertex-wise and 1,000 parcels with 0–10 mm smoothing) are lower ($r \sim 0.55$). See the results for δ_1 in the **Supplementary Figure 4**.

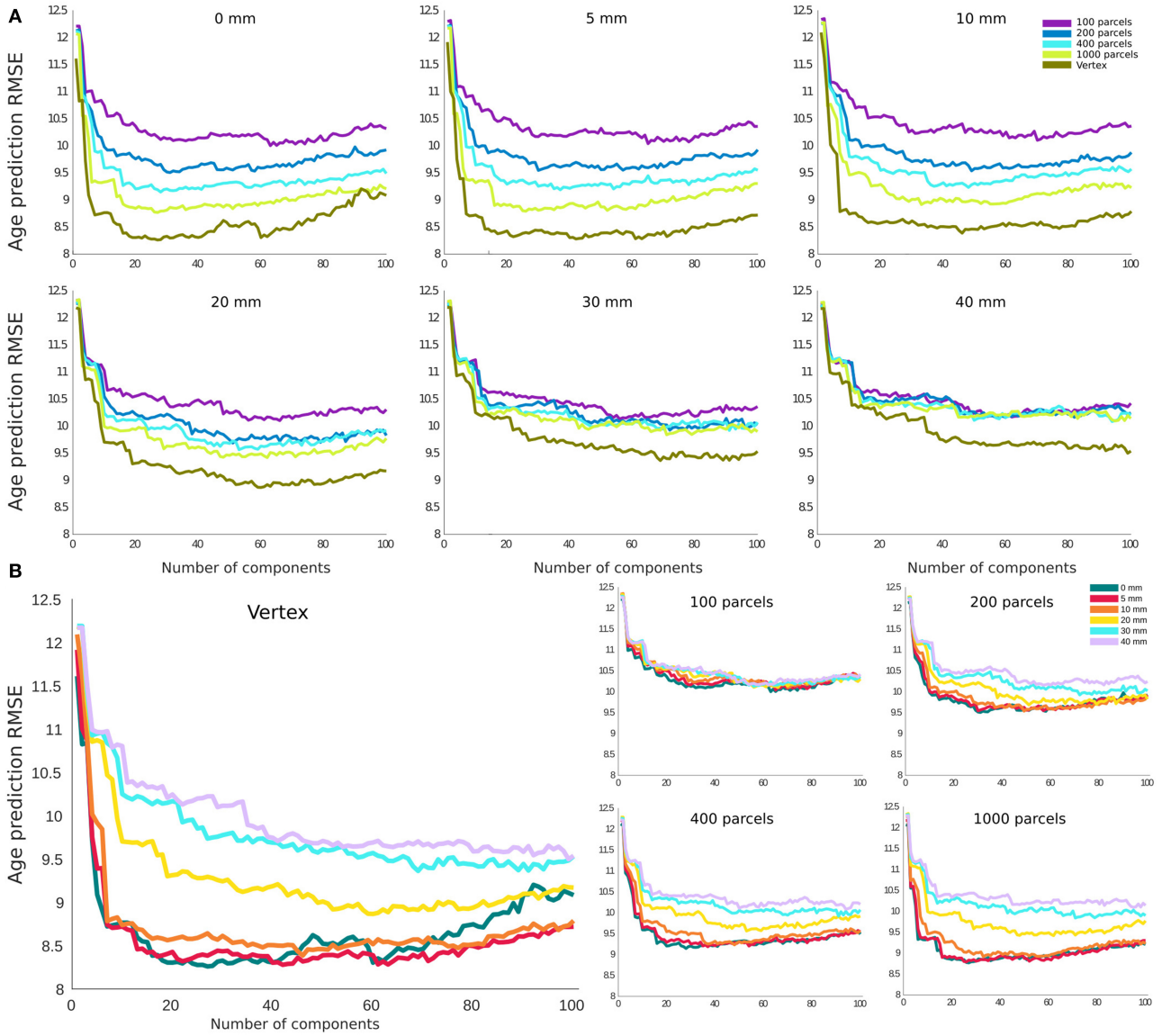


FIGURE 4 | Root mean square error (RMSE) for age prediction as a function of number of principal components included as features in the predictive model. **(A)** Results grouped together based on the smoothing level. **(B)** Results grouped together based on the parcellation resolution.

DISCUSSION

In this article, we compared the effect of different smoothing and parcellation on associations between cortical thickness and chronological age as well as brain age prediction accuracy. We showed that the optimal choice for association analysis might indeed undermine age prediction accuracy, and vice versa. We further investigated this relationship and demonstrated the underlying differences that lead to this trade-off between the two analyses. Finally, we examined the effect of smoothing and parcellation on delta age estimation and showed that the initial smoothing and parcellation choices can change the delta estimation which in turn will affect any downstream analysis.

We used brain association with age and brain age prediction as our target analyses, since age is used as the main variable of interest or at least a confounding variable in most neuroimaging studies. We used cortical thickness as the main measure of interest. Due to the wide availability of T1-weighted MRI in research and clinical settings, cortical thickness is a suitable measure which has been widely used to study brain anatomy in general (Toga, 2015), and more specifically, brain aging and predicting brain age (Wang and Pham, 2011; Groves et al., 2012; Kandel et al., 2013; Liem et al., 2017). Finally, our results are presented based on a sample size of N∼600 which is a common sample size for publicly available datasets in the field of neuroimaging.

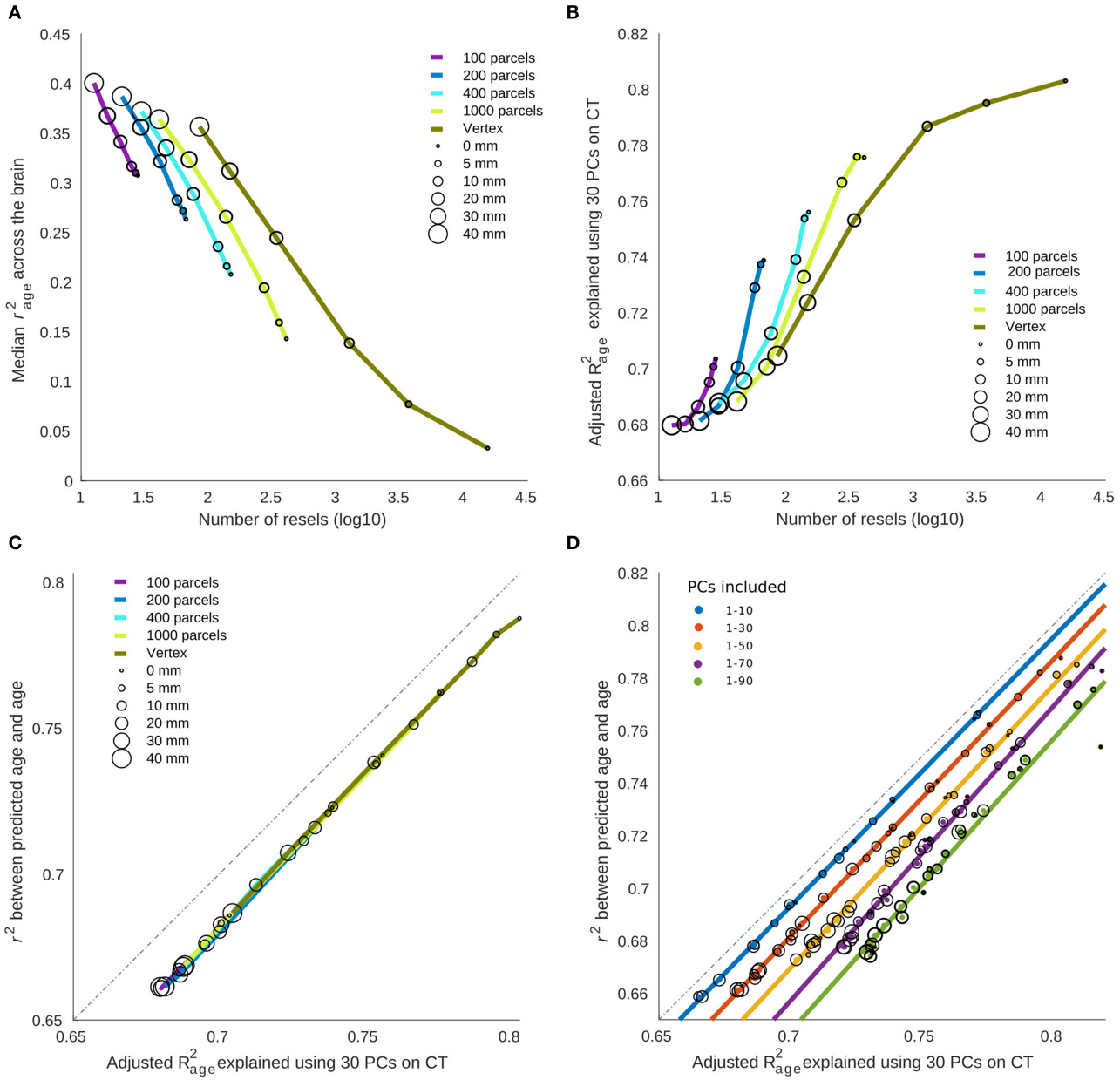


FIGURE 5 | The relationship between cortical thickness association with age vs. brain age prediction. **(A)** Median variance explained of cortical thickness across the brain. The results are grouped based on the parcellation. Circles represent the something level within each parcellation. **(B)** Total variance explained of age by the first 30 principal components (PCs) of cortical thickness as independent variables. **(C)** The relationship between age prediction accuracy and total variance explained of age. In the case of prediction, the first PCs are used as predictive features alongside cross validation to prevent overfitting. The total variance explained of age is the same as depicted in **(B)**. **(D)** The overfitting bias of linear model compared to the same model used with cross validation. As expected, a higher number of predictive features results in higher level of overfitting bias.

Given the limited number of subjects in neuroimaging studies compared to potential features (number of vertices/voxels), most prediction studies apply dimension reduction as an initial step. We used PCA for dimension reduction of the cortical thickness data. Due to its simplicity and interpretability, PCA has been widely used in the brain age prediction literature. Furthermore, we employed linear regression with cross-validation as our prediction model (Smith et al., 2019). As expected, we observed an initial drop in the prediction error, followed by a plateau/increase in the error as the sample to feature ratio increases (Hastie et al., 2009). At each parcellation level, the accuracy drops with increased smoothing, and for each smoothing level, the accuracy decreases with larger parcels/regions.

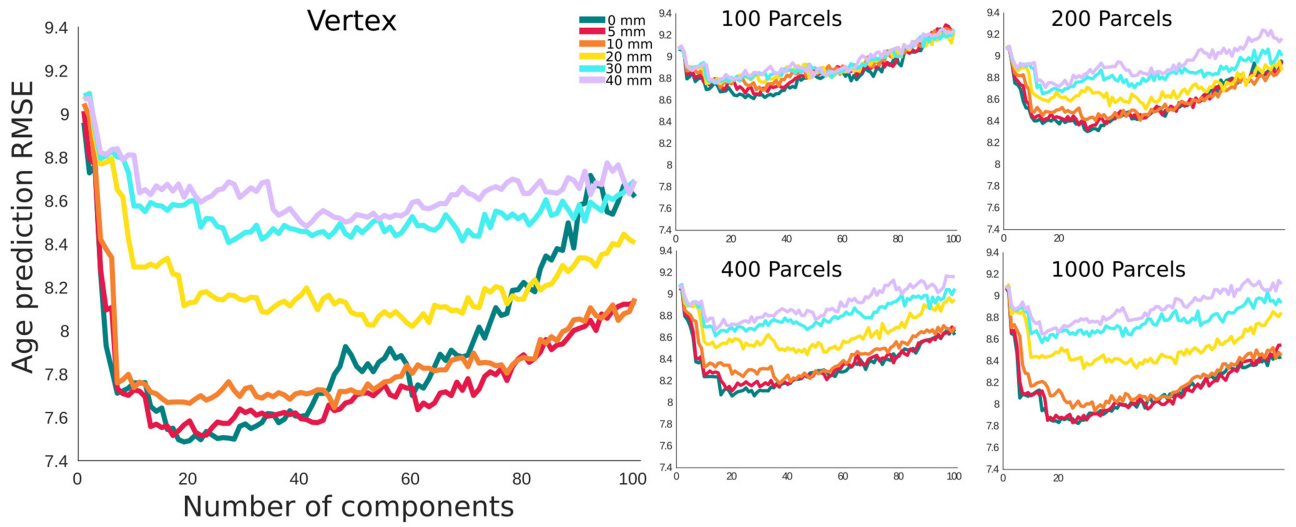


FIGURE 6 | Root mean square error (RMSE) for Age prediction with δ_2 as the error term. The x axis shows the number of principal components included as features. The results are grouped based on the parcellation resolution.

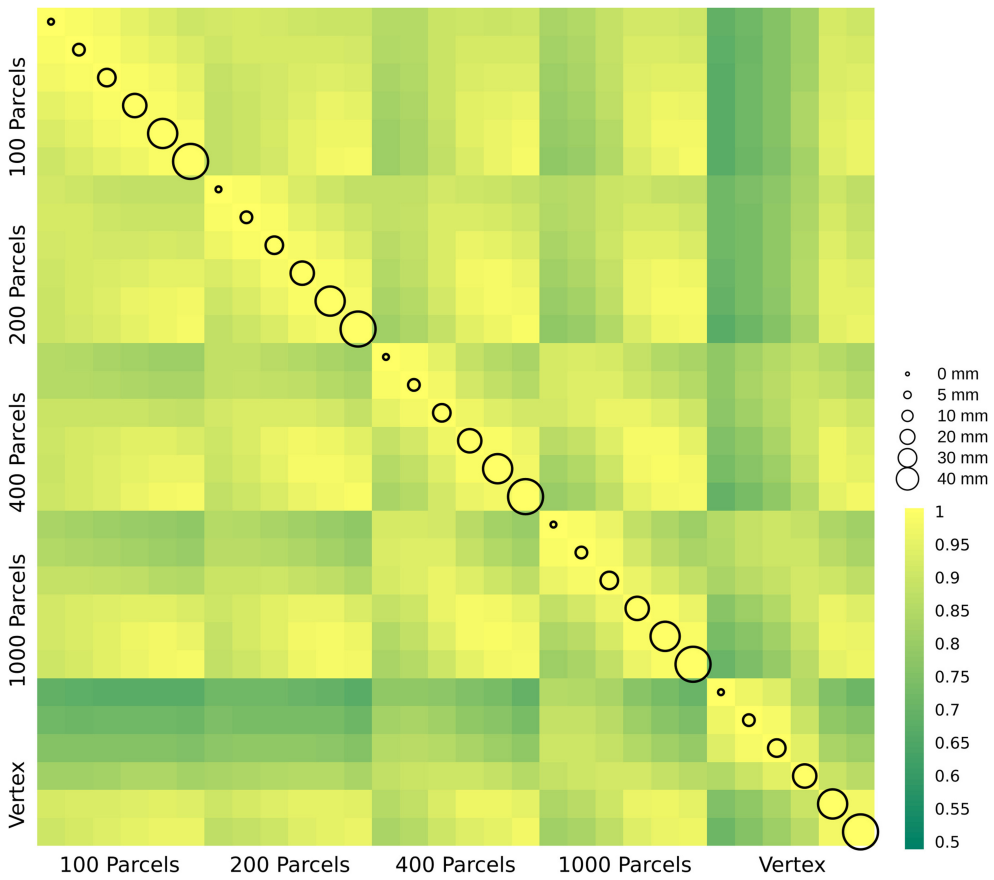


FIGURE 7 | δ_2 age prediction error. The correlation between delta age (as measured by δ_2) across parcellation resolutions (x and y axis labels) and smoothing kernels (represented by circle size).

The age prediction results presented in the main manuscript are based on linear regression analysis and PCA based features. We have also examined the performance of support vector machines with different kernel types as well as linear regression using lasso and ridge regularization methods. All these methods were applied on both raw cortical thickness values as well as PCA based features as predictors. In all cases, the PCA based features outperformed the same method using the raw cortical thickness values. These results can be due to the relatively small sample size and/or sample to feature ratio in the current study. Furthermore, with the exception of linear regression models with lasso regularization, the presented linear regression method outperformed all other methods. In the case of lasso, we optimized the method for the regularization weight (i.e., lambda parameter). With this optimization, we gained a 2% increase in accuracy. All the results based on the explained models are reported as **Supplementary Tables 1–4** and **Supplementary Figures 6–8**. Although the accuracy might vary slightly between methods, the higher accuracy in smaller smoothing kernels (i.e., 0 and 5 mm smoothing) and smaller parcels (i.e., vertex based level) is consistent across all methods and the explained relationship between age- association and prediction holds true across these prediction models. Furthermore, the anatomical correlates of aging or the main anatomical features contributing to the brain age prediction were not the main target or in the scope of the current study. However, the mapping of the first 100 PCs (used in the prediction analysis) to the Schaeffer parcellation as well as whole brain vertices of CIVET is provided in the **Supplementary Tables 5, 6**. They can be used alongside the other **Supplementary Tables** to infer variables of interest and their anatomical distribution.

While 10-fold cross validation is enough in our case, for relatively small samples, the random assignment of data in the cross-validation partitions can lead to differences in the distribution of the training and test data in some of the folds, leading to highly variable performances across some folds. To exercise the best practice and ensure that the reported results are robust, the 10-fold cross validation procedure was repeated and the results were averaged so that such inhomogeneous assignments (however unlikely) do not impact the reported results. The randomized performances were very similar (mean correlation between repetitions was between 0.97 and 0.99) and the standard deviation of the repetitions is $<1\%$ of the reported value across repetitions, suggesting that our results are indeed robust and the repetitions were not necessary for the conclusions in the manuscript.

While not exceptionally high, the brain age prediction accuracy in this study is comparable to similar studies in the field (see Franke et al., 2020; Table 3). Furthermore, the accuracy of brain age prediction is dependent on two factors which can significantly impact prediction performance (1) Age range and variance: With the current population's ages ranging between 18 and 88 years (mean age = 53.52, standard deviation = 18.07), CamCAN dataset is one of the more challenging datasets for prediction. (2) Distribution of age: Prediction models tend to favor values close to the mean of the population. Therefore, data with a Gaussian distribution (which is generally used in other similar brain age prediction studies) will result in a much better prediction performance compared to a rather uniform distribution of age which is the case for the CamCAN dataset.

In terms of variability within the cytoarchitectural regions, there is a distinction between the change in the insular cortex compared to the rest of the regions. The main shift occurs between 400 and 200 parcellation levels (where Insular cortex parcels are combined from 23 to 15 parcels). As a result, several regions (in both right and left hemispheres) with distinct cortical thickness values are combined and averaged together, resulting in a drop in variance and consequently coefficient of variation across regions. This might be due to the unique morphometric properties of insular regions as well as the limited number of parcels in the insular cortex compared to other cytoarchitectural regions. This misalignment might also be due to the functional nature of the Schaefer cortical parcellation, which doesn't necessarily have a one-to-one to correspondence with the structural variability in the same areas.

It is commonplace for neuroimaging studies to use smoothing and parcellation as the first step of their analysis to achieve higher statistical power with reducing the individual variability within the data. Furthermore, with increased availability of public neuroimaging datasets, it is commonplace to release a preprocessed version of the data with a fixed smoothing level and averaged based on a given parcellation. Many research groups in the field use preprocessed and parcellation-based data releases as the starting point for their analyses. In fact, in many cases, the raw data is not publicly distributed, and the preprocessed parcellated data is the only version of data available. For example, some of the most influential public datasets in the field of neuroimaging such as Adolescent Brain Cognitive Development (ABCD, for details see https://nda.nih.gov/abcd) Study and UKBiobank (for details see https://www.ukbiobank.ac.uk) provide cortical thickness data using Desikan-Killiany-Tourville parcellations (Klein and Tourville, 2012) with 62 regions (smoothing varies across studies) as one of their pre-calculated measures. Our findings can help provide a guide to interpret these available measures and shed light on the effect of these preselected parameters/parcellation when applied in aging studies.

Higher correlation values across brain regions (as a result of smoothing) can be explained by increased signal to noise ratio and reduced individual variability (**Figure 2**). The effect of smoothing on brain related associations has previously been studied (Lerch and Evans, 2005). Indeed, Zhao and colleagues propose smoothing as a scaling dimension which needs optimization for any given target analysis (Zhao et al., 2013). The effect of parcellation on brain association has been addressed in several studies. However, the optimal parcellation level is still an open question dependent on the specific case of interest (Eickhoff et al., 2018). Here, we showed that parcellation level has a similar impact, by reducing variability, using both CV (**Figure 2**) and number of resels (**Figure 1**).

Association/correlation analyses reflect the general patterns across the population (suitable for studies that investigate population specific trends), whereas prediction analyses aim to determine the likely value of a certain measure of interest at

the individual level (suitable for diagnosis/prognosis purposes). Association analyses in general benefit from averaging, since it lowers the levels of noise and improves the obtained correlations, allowing the analysis to draw out the overall trends of the population. In contrast, prediction is inherently a much more challenging task, since it aims to provide accurate estimates at the individual level. Averaging methods decrease individual variability and differences, preventing the prediction models from accurately capturing the individual variabilities.

Resel numbers are statistical constructs based on the association analysis, while at the same time informing our interpretation of prediction analysis. As such resels don't have any inherent biological interpretation. Even within the same dataset and using the same metrics, the number of resels will differ between different statistical analyses (e.g., the number of resels will change if we use fluid intelligence or working memory measures instead of age) since it is a construct that evaluates the number of resolution elements by considering the dependency across regions/vertices with regards to a certain variable within an association or statistical contrast analysis over the region of interest (in our case entire brain surface). A higher number of resels reflects a higher number of independent features (with regards to age), which in turn captures the individual variability across the population and increases prediction accuracy (as confirmed by our findings presented in **Figure 5**). However, this higher level of individual variability represented in the features will result in lower correlation values across the brain (also shown in our results in **Figure 5**).

In neuroimaging, smoothing and parcellations are generally studied separately. In this study, we used a unified metric to directly compare the effect of smoothing and parcellation. Using resel numbers and variance explained in the model, we have calculated common measures for both association and prediction results. Our results show that with increased smoothing and larger parcels (i.e., lower number of resels), cortical thickness variability reduces. This will remove inter-individual differences across brain regions and result in higher associations between cortical thickness and aging (**Figure 5A**). However, while this improves the regional correlation with age, most of this general trend can be captured in a few PCs (mainly the first component) and the rest of the PCs do not explain the remaining variance of age. On the other hand, this relationship is reversed in the conditions with higher resel numbers (i.e., lower smoothing and higher spatial resolutions). While in these cases higher regional variability results in lower correlation with age, the age related associations capture different portions of age variance in different PCs and overall they have a higher adjusted R^2 (**Figure 5B**). There was a consistent bias in the adjusted R^2 across conditions (**Figures 5C,D**), however, the effects remained similar after removing the overfitting with cross-validation. Altogether, these analyses explain the seeming opposite direction of correlation values and prediction accuracies for different smoothing/parcellation levels in section Statistical Association Between Cortical Thickness and Aging and Brain Age Prediction Based on Cortical Thickness.

One should also consider that while the objective function in linear regression and its variants is based on RMSE (shown in **Figure 4**), considering the interdependence between the features, there is a close linear relationship between adjusted R^2 and the prediction accuracy based on the RMSE. Furthermore, our conclusions were independent of the use of RMSE and R^2 as shown in **Supplementary Figure 3**. With these considerations, without loss of generality, we have used r^2 from correlation analysis and adjusted R^2 from the linear regression model alongside the resel numbers (as shown in **Figure 5**) to study the relationship between the association and prediction analysis.

While delta age in itself is not the target of the current study, it is important in so far as it is the main measure derived from age prediction studies. The discrepancy between predicted age and chronological age (i.e., delta age) is used to study other phenotypes (either demographic, biological, or clinical) (Cole and Franke, 2017). Based on this definition, subjects with higher delta age are assumed to have accelerated aging (i.e., their brain is similar to brains of older individuals). Several studies have found relationships between delta age and brain disorders including but not limited to traumatic brain injury, schizophrenia, epilepsy, mild cognitive impairment, and Alzheimer's disease. (See Cole and Franke, 2017; Franke and Gaser, 2019; Franke et al., 2020 for a complete review of the topic). A recent study using 45,615 subjects simultaneously investigated the relationship between delta age and 10 different brain disorders and found that subjects with Schizophrenia, Multiple Sclerosis, Mild cognitive impairment, and dementia show higher delta age compared to the controls (Kaufmann et al., 2019). The increase in the studies of brain age emphasizes that not only a better understanding of the biological nature of delta age is needed, but also a systematic study of the effect of analytical and computational methods used to obtain delta age is necessary. However, the effect of the preprocessing condition on delta age estimation is not studied. Here we have examined the effect of parcellation and smoothing levels as an important factor that can change delta age estimation and consequently the aforementioned relationships with other measures. In the current manuscript, we found a range of associations (0.5–1) between δ_2s obtained in different conditions. These results suggest not only that each study needs to optimize their choice of the smoothing and parcellation level, but also when interpreting results from different studies in the field, these parameters should be considered.

One of the main limitations of the current study is the number of subjects (N∼600), particularly given that their age spans across 70 years. This leads to overfitting as the number of features increase. In fact, for vertex-wise prediction (with 0 mm smoothing), the first 30 PCs only explain 20% of the variability in the data. This number is around 40% for 10 mm smoothing. In comparison, the first 30 PCs for 100 parcels explain 80 and 90% of the variance of the cortical thickness data for 0 mm and 40 mm smoothing levels, respectively (**Supplementary Figure 5**). Given the higher performance of the vertex-wise PCs at 0–10 mm smoothing, it is likely that with a larger sample size and increased sample to feature ratio, the accuracy can be further improved. It should be noted that in each case the variance explained corresponds to the total variability for the corresponding smoothing and parcellation condition. Another limitation in the current study is the use of functionally driven Schaefer

parcellations. While this does not automatically suggest a disadvantage, multi-resolution anatomically driven parcellations have the theoretical advantage of a more relevant initial feature space for cortical thickness studies. Finally, CamCAN data used in our study is cross-sectional. This potentially decreases the detection power of our study, since we can only estimate the effect of time between subjects with individual variability as part of the measurement, whereas a longitudinal dataset can decrease variability by estimating the effect of aging within subjects.

Traditionally, neuroimaging studies have targeted brain related associations with a given phenotype/symptom or the statistical differences between different groups for a given brain region, followed up with the association of these differences with a given biological or behavioral variable of interest. More recently, there has been an ongoing conversation in the field toward prediction as an alternative approach. Along the same line, the field of brain aging, has pursued age related associations as well as age prediction. The relationship between the two approaches is often taken for granted (since in ideal settings, i.e., large sample size and low inter-individual variability or noise levels, the results would be equivalent) and ignored in practice. In this study, we have directly addressed both age association and prediction as a function of smoothing and parcellation levels. Within our sample size, we found an inverse relationship between regional age related associations and brain age prediction accuracy as a function of smoothing and parcellation level, highlighting the importance of the parameter selection based on the goal of the study.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: https://camcan-archive.mrc-cbu.cam.ac.uk/dataaccess/datarequest.php.

ETHICS STATEMENT

This study is conducted in compliance with the Helsinki Declaration, and has been approved by the local ethics committee, Cambridgeshire 2 Research Ethics Committee (now East of England-Cambridge Central). The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

YZ contributed to the study plan, analyzed the data, and wrote the manuscript. AE contributed to the study plan and revision of the manuscript. All authors contributed to the article and approved the submitted version.

FUNDING

This work was supported by funding from the Canada First Research Excellence Fund, awarded to McGill University for the Healthy Brains, Healthy Lives (HBHL) initiative, CBRAIN for Multidisciplinary Reproducible Science Grant (CANARIE RS3-031), and Canadian Institutes of Health Research Operating Grant (CIHR PJT-173236).

ACKNOWLEDGMENTS

Data collection and sharing for this project was provided by the Cambridge Centre for Ageing and Neuroscience (CamCAN). CamCAN funding was provided by the UK Biotechnology and Biological Sciences Research Council (grant number BB/H008217/1), together with support from the UK Medical Research Council and University of Cambridge, UK. Authors thank Compute Canada (https://www.computecanada.ca/home) for the usage of the computing resources in the current work. Authors thank Dr. Lindsay B Lewis for providing the surface parcellations registered to the MNI surface space as well as Drs. Mahsa Dadar and Filip Morys for their inputs and comments regarding data analysis and the final manuscript.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.637724/full#supplementary-material

REFERENCES

Booth, T., Starr, J. M., and Deary, I. (2013). Modeling multisystem biological risk in later life: allostatic load in the lothian birth cohort study 1936. *Am. J. Hum. Biol.* 25, 538–543. doi: 10.1002/ajhb.22406

Bzdok, D., Varoquaux, G., and Steyerberg, E. W. (2020). Prediction, not association, paves the road to precision medicine. *JAMA Psychiatry* 78, 127–128. doi: 10.1001/jamapsychiatry.2020.2549

Chung, M., Worsley, K., Paus, T., Robbins, S., Evans, A. C., Taylor, J., et al. (2002). *Tensor-Based Surface Morphometry*. University of Wisconsin. Available online at: http://www.stat.wisc.edu/$\sim$mchung

Cole, J. H., and Franke, K. (2017). Predicting age using neuroimaging: innovative brain ageing biomarkers. *Trends Neurosci.* 40, 681–690. doi: 10.1016/j.tins.2017.10.001

Cole, J. H., Ritchie, S. J., Bastin, M. E., Valdés Hernández, M. C., Muñoz Maniega, S., Royle, N., et al. (2018). Brain age predicts mortality. *Mol. Psychiatry* 23, 1385–1392. doi: 10.1038/mp.2017.62

Collins, D. L., Neelin, P., Peters, T. M., and Evans, A. C. (1994). Automatic 3d intersubject registration of mr volumetric data in standardized talairach space. *J. Comput. Assist. Tomogr.* 18, 192–205. doi: 10.1097/00004728-199403000-00005

Curiati, P. K., Tamashiro, J. H., Squarzoni, P., Duran, F. L. S., Santos, L. C., Wajngarten, M., et al. (2009). Brain structural variability due to aging and gender in cognitively healthy elders: results from the São Paulo ageing and health study. *Am. J. Neuroradiol.* 30, 1850–1856. doi: 10.3174/ajnr.A1727

Eickhoff, S. B., Yeo, B. T. T., and Genon, S. (2018). Imaging-based parcellations of the human brain. *Nat. Rev. Neurosci.* 19, 672–686. doi: 10.1038/s41583-018-0071-7

Franke, K., Bublak, P., Hoyer, D., Billiet, T., Gaser, C., Witte, O. W., et al. (2020). *In vivo* biomarkers of structural and functional brain development and aging in humans. *Neurosci. Biobehav. Rev.* 117, 142–164. doi: 10.1016/j.neubiorev.2017.11.002

Franke, K., and Gaser, C. (2019). Ten years of brainage as a neuroimaging biomarker of brain aging: what insights have we gained? *Front. Neurol.* 10:789. doi: 10.3389/fneur.2019.00789

Franke, K., Luders, E., May, A., Wilke, M., and Gaser, C. (2012). Brain maturation: predicting individual BrainAGE in children and adolescents using structural MRI. *Neuroimage* 63, 1305–1312. doi: 10.1016/j.neuroimage.2012.08.001

Gaser, C., Franke, K., Klöppel, S., Koutsouleris, N., and Sauer, H. (2013). BrainAGE in mild cognitive impaired patients: predicting the conversion to Alzheimer's disease. *PLoS ONE* 8:e67346. doi: 10.1371/journal.pone.0067346

Gorgolewski, K., Tambini, A., Durnez, J., Sochat, V., Wexler, J., and Poldrack, R. (2016). "Evaluation of full brain parcellation schemes using the NeuroVault database of statistical maps," in *Organisation for Human Brain Mapping* 2016 *Annual Meeting*. Available online at: https://54.246.141.91/posters/6-1986

Groves, A. R., Smith, S. M., Fjell, A. M., Tamnes, C. K., Walhovd, K. B., Douaud, G., et al. (2012). Benefits of multi-modal fusion analysis on a large-scale dataset: life-span patterns of inter-subject variability in cortical morphometry and white matter microstructure. *Neuroimage* 63, 365–380. doi: 10.1016/j.neuroimage.2012.06.038

Hastie, T., Tibshirani, R., and Friedman, J. (2009). *Springer Series in Statistics The Elements of Statistical Learning Data Mining, Inference, and Prediction*. Available online at: https://books.google.ca/books?hl=enandlr=andid=tVIjmNS3Ob8Candoi=fndandpg=PR13anddq=trevor$+$hastie$++$bookandots=EOBcP9J5X5andsig=w9Dod2i1zZD9tkzSKn0TPwDs1UE

Hayasaka, S., Phan, K. L., Liberzon, I., Worsley, K. J., and Nichols, T. E. (2004). Nonstationary cluster-size inference with random field and permutation methods. *Neuroimage* 22, 676–687. doi: 10.1016/j.neuroimage.2004.01.041

Horvath, S. (2013). DNA methylation age of human tissues and cell types. *Genome Biol.* 14:3156. doi: 10.1186/gb-2013-14-10-r115

Hu, S., Chao, H. H. A., Zhang, S., Ide, J. S., and Li, C. S. R. (2014). Changes in cerebral morphometry and amplitude of low-frequency fluctuations of BOLD signals during healthy aging: correlation with inhibitory control. *Brain Struct. Funct.* 219, 983–994. doi: 10.1007/s00429-013-0548-0

Jones, S. E., Buchbinder, B. R., and Aharon, I. (2000). Three-dimensional mapping of cortical thickness using Laplace's equation. *Hum. Brain Mapp.* 11, 12–32. doi: 10.1002/1097-0193(200009)11:1andlt;12::AID-HBM20andgt;3.0.CO;2-K

Kabani, N., Le Goualher, G., Macdonald, D., and Evans, A. C. (2001). Measurement of cortical thickness using an automated 3-D algorithm: a validation study. *Neuroimage* 13, 375–380. doi: 10.1006/nimg.2000.0652

Kandel, B. M., Wolk, D. A., Gee, J. C., and Avants, B. (2013). Predicting cognitive data from medical images using sparse linear regression. *Inf. Process Med Imaging* 23, 86–97. doi: 10.1007/978-3-642-38868-2_8

Kaufmann, T., van der Meer, D., Doan, N. T., Schwarz, E., Lund, M. J., Agartz, I., et al. (2019). Common brain disorders are associated with heritable patterns of apparent aging of the brain. *Nat. Neurosci.* 22, 1617–1623. doi: 10.1038/s41593-019-0471-7

Khundrakpam, B. S., Tohka, J., Evans, A. C., Ball, W. S., Byars, A. W., Schapiro, M., et al. (2015). Prediction of brain maturity based on cortical thickness at different spatial resolutions. *Neuroimage* 111, 350–359. doi: 10.1016/j.neuroimage.2015.02.046

Kim, J., Singh, V., Jun, K. L., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., et al. (2005). Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *Neuroimag.* 27, 210–221. doi: 10.1016/j.neuroimage.2005.03.036

Klein, A., and Tourville, J. (2012). 101 labeled brain images and a consistent human cortical labeling protocol. *Front. Neurosci.* 6:171. doi: 10.3389/fnins.2012.00171

Le, T. T., Kuplicki, R. T., McKinney, B. A., Yeh, H.-W., Thompson, W. K., and Paulus, M. P. (2018). A nonlinear simulation framework supports adjusting for age when analyzing BrainAGE. *Front. Aging Neurosci.* 10:317. doi: 10.3389/fnagi.2018.00317

Lemaître, H., Crivello, F., Grassiot, B., Alpérovitch, A., Tzourio, C., and Mazoyer, B. (2005). Age- and sex-related effects on the neuroanatomy of healthy elderly. *Neuroimage* 26, 900–911. doi: 10.1016/j.neuroimage.2005.02.042

Lerch, J. P., and Evans, A. C. (2005). Cortical thickness analysis examined through power analysis and a population simulation. *Neuroimage* 24, 163–173. doi: 10.1016/j.neuroimage.2004.07.045

Lerch, J. P., Worsley, K., Shaw, W. P., Greenstein, D. K., Lenroot, R. K., Giedd, J., et al. (2006). Mapping anatomical correlations across cerebral cortex (MACACC) using cortical thickness from MRI. *Neuroimage* 31, 993–1003. doi: 10.1016/j.neuroimage.2006.01.042

Lewis, J. D., Fonov, V. S., Collins, D. L., Evans, A. C., and Tohka, J. (2019). Cortical and subcortical T1 white/gray contrast, chronological age, and cognitive performance. *Neuroimage* 196, 276–288. doi: 10.1016/j.neuroimage.2019.04.022

Lewis, L. B., Lepage, C. Y., and Evans, A. C. (2019). "An extended MSM surface registration pipeline to bridge atlases across the MNI and the FS/HCP worlds," in *Annual Meeting of the Organization for Human Brain Mapping*. Available online at: https://ww5.aievolution.com/hbm1901/index.cfm?do=abs.viewAbsandabs=1243

Liang, H., Zhang, F., and Niu, X. (2019). Investigating systematic bias in brain age estimation with application to post-traumatic stress disorders. *Hum. Brain Mapp.* 40, 3143–3152. doi: 10.1002/hbm.24588

Liem, F., Varoquaux, G., Kynast, J., Beyer, F., Kharabian Masouleh, S., Huntenburg, J. M., et al. (2017). Predicting brain-age from multimodal imaging data captures cognitive impairment. *Neuroimage* 148, 179–188. doi: 10.1016/j.neuroimage.2016.11.005

López-Otín, C., Blasco, M. A., Partridge, L., Serrano, M., and Kroemer, G. (2013). The hallmarks of aging. *Cell* 153:1194. doi: 10.1016/j.cell.2013.05.039

MacDonald, D., Kabani, N., Avis, D., and Evans, A. C. (2000). Automated 3-D extraction of inner and outer surfaces of cerebral cortex from MRI. *Neuroimage* 12, 340–356. doi: 10.1006/nimg.1999.0534

Salehi, M., Greene, A. S., Karbasi, A., Shen, X., Scheinost, D., and Constable, R. T. (2020). There is no single functional atlas even for a single individual: functional parcel definitions change with task. *Neuroimage* 208:116366. doi: 10.1016/j.neuroimage.2019.116366

Schaefer, A., Kong, R., Gordon, E. M., Laumann, T. O., Zuo, X.-N., Holmes, A. J., et al. (2018). Local-global parcellation of the human cerebral cortex from intrinsic functional connectivity MRI. *Cerebral Cortex* 28, 3095–3114. doi: 10.1093/cercor/bhx179

Shafto, M. A., Tyler, L. K., Dixon, M., Taylor, J. R., Rowe, J. B., Cusack, R., et al. (2014). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) study protocol: a cross-sectional, lifespan, multidisciplinary examination of healthy cognitive ageing. *BMC Neurol.* 14:204. doi: 10.1186/s12883-014-0204-1

Sled, J. G., Zijdenbos, A. P., and Evans, A. C. (1998). A nonparametric method for automatic correction of intensity nonuniformity in mri data. *IEEE Trans. Med. Imaging* 17, 87–97. doi: 10.1109/42.668698

Smith, S. M., Vidaurre, D., Alfaro-Almagro, F., Nichols, T. E., and Miller, K. L. (2019). Estimation of brain age delta from brain imaging. *Neuroimage* 200, 528–539. doi: 10.1016/j.neuroimage.2019.06.017

Storsve, A. B., Fjell, A. M., Tamnes, C. K., Westlye, L. T., Overbye, K., Aasland, H. W., et al. (2014). Differential longitudinal changes in cortical thickness, surface area and volume across the adult life span: regions of accelerating and decelerating change. *J. Neurosci.* 34, 8488–8498. doi: 10.1523/JNEUROSCI.0391-14.2014

Takahashi, R., Ishii, K., Kakigi, T., and Yokoyama, K. (2011). Gender and age differences in normal adult human brain: voxel-based morphometric study. *Hum. Brain Mapp.* 32, 1050–1058. doi: 10.1002/hbm.21088

Taylor, J. R., Williams, N., Cusack, R., Auer, T., Shafto, M. A., Dixon, M., et al. (2017). The Cambridge Centre for Ageing and Neuroscience (Cam-CAN) data repository: structural and functional MRI, MEG, and cognitive data from a cross-sectional adult lifespan sample. *Neuroimage* 144, 262–269. doi: 10.1016/j.neuroimage.2015.09.018

Thomas Yeo, B. T., Krienen, F. M., Sepulcre, J., Sabuncu, M. R., Lashkari, D., Hollinshead, M., et al. (2011). The organization of the human cerebral cortex estimated by intrinsic functional connectivity. *J. Neurophysiol.* 106, 1125–1165. doi: 10.1152/jn.00338.2011

Toga, A. W. (2015). *Brain Mapping: An Encyclopedic Reference.* Academic Press.

Tohka, J., Zijdenbos, A., and Evans, A. (2004). Fast and robust parameter estimation for statistical partial volume models in brain MRI. *Neuroimage* 23, 84–97. doi: 10.1016/j.neuroimage.2004.05.007

von Economo, C. F., and Koskinas, G. (1927). Die cytoarchitektonik der hirnrinde des erwachsenen menschen. *J. Anat*. 61, 264–266.

Wang, B., and Pham, T. D. (2011). MRI-based age prediction using hidden Markov models. *J. Neurosci. Methods* 199, 140–145. doi: 10.1016/j.jneumeth.2011.04.022

Worsley, K. J. (1996). An unbiased estimator for the roughness of a multivariate Gaussian random eld 1 Model, 1–5. Available online at: https://pdfs.semanticscholar.org/4159/a64da50863a945c8af38c42f8b09487a985b.pdf

Worsley, K. J., Andermann, M., Koulis, T., MacDonald, D., and Evans, A. C. (1999). Detecting changes in nonisotropic images. *Hum. Brain Mapp*. 8, 98–101.

Worsley, K. J., Evans, A. C., Marrett, S., and Neelin, P. (1992). A three-dimensional statistical analysis for CBF activation studies in human brain. *J. Cerebral Blood Flow Metab*. 12, 900–918. doi: 10.1038/jcbfm.1992.127

Yarkoni, T., and Westfall, J. (2017). Choosing prediction over explanation in psychology: lessons from machine learning. *Perspect. Psychol. Sci*. 12, 1100–1122. doi: 10.1177/1745691617693393

Zhao, L., Boucher, M., Rosa-Neto, P., and Evans, A. C. (2013). Impact of scale space search on age- and gender-related changes in MRI-based cortical morphometry. *Hum. Brain Mapp*. 34, 2113–2128. doi: 10.1002/hbm.22050

Ziegler, G., Dahnke, R., Jäncke, L., Yotter, R. A., May, A., and Gaser, C. (2012). Brain structural trajectories over the adult lifespan. *Hum. Brain Mapp*. 33, 2377–2389. doi: 10.1002/hbm.21374

Zijdenbos, A. P., Forghani, R., and Evans, A. C. (2002). Automatic "pipeline" analysis of 3-D MRI data for clinical trials: application to multiple sclerosis. *IEEE Trans. Med. Imaging* 21, 1280–1291. doi: 10.1109/TMI.2002.806283

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 20 May 2021
doi: 10.3389/fdata.2021.661110

Inter-Cohort Validation of SuStaIn Model for Alzheimer's Disease

Damiano Archetti[1], Alexandra L. Young[2,3], Neil P. Oxtoby[3], Daniel Ferreira[4,5], Gustav Mårtensson[4], Eric Westman[4], Daniel C. Alexander[3], Giovanni B. Frisoni[6,7] and Alberto Redolfi[1] for Alzheimer's Disease Neuroimaging Initiative and EuroPOND Consortium*

[1]Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy, [2]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [3]Department of Computer Science, UCL Centre for Medical Image Computing, London, United Kingdom, [4]Division of Clinical Geriatrics, Center for Alzheimer Research, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden, [5]Department of Radiology, Mayo Clinic, Rochester, MN, United States, [6]Memory Clinic and LANVIE - Laboratory of Neuroimaging of Aging, University Hospitals and University of Geneva, Geneva, Switzerland, [7]Laboratory of Alzheimer's Neuroimaging and Epidemiology - LANE, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy

Alzheimer's disease (AD) is a neurodegenerative disorder which spans several years from preclinical manifestations to dementia. In recent years, interest in the application of machine learning (ML) algorithms to personalized medicine has grown considerably, and a major challenge that such models face is the transferability from the research settings to clinical practice. The objective of this work was to demonstrate the transferability of the Subtype and Stage Inference (SuStaIn) model from well-characterized research data set, employed as training set, to independent less-structured and heterogeneous test sets representative of the clinical setting. The training set was composed of MRI data of 1043 subjects from the Alzheimer's disease Neuroimaging Initiative (ADNI), and the test set was composed of data from 767 subjects from OASIS, Pharma-Cog, and ViTA clinical datasets. Both sets included subjects covering the entire spectrum of AD, and for both sets volumes of relevant brain regions were derived from T1-3D MRI scans processed with Freesurfer v5.3 cross-sectional stream. In order to assess the predictive value of the model, subpopulations of subjects with stable mild cognitive impairment (MCI) and MCIs that progressed to AD dementia (pMCI) were identified in both sets. SuStaIn identified three disease subtypes, of which the most prevalent corresponded to the typical atrophy pattern of AD. The other SuStaIn subtypes exhibited similarities with the previously defined hippocampal sparing and limbic predominant atrophy patterns of AD. Subject subtyping proved to be consistent in time for all cohorts and the staging provided by the model was correlated with cognitive performance. Classification of subjects on the basis of a combination of SuStaIn subtype

OPEN ACCESS

Edited by:
Tuan D. Pham, Prince Mohammad bin Fahd University, Saudi Arabia

Reviewed by:
Gang Wang, Shanghai Jiao Tong University, China
Dinh Tuan Phan Le, New York City Health and Hospitals Corporation, United States

***Correspondence:**
Damiano Archetti
darchetti@fatebenefratelli.eu

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *30 January 2021*
Accepted: *04 May 2021*
Published: *20 May 2021*

Citation:
Archetti D, Young AL, Oxtoby NP, Ferreira D, Mårtensson G, Westman E, Alexander DC, Frisoni GB and Redolfi A (2021) Inter-Cohort Validation of SuStaIn Model for Alzheimer's Disease. Front. Big Data 4:661110. doi: 10.3389/fdata.2021.661110

Abbreviations: AD, Alzheimer's disease; ADNI, Alzheimer's disease Neuroimaging Initiative; ANOVA, Analysis of Variance; AUC, Area Under Curve; AVRA, Automatic Visual Ratings of Atrophy; $A\beta_{1-42}$, Amyloid-β_{1-42}; CDR, Clinical dementia rating; CN, Cognitively Normal; CSF, Cerebrospinal Fluid; CVIC, Cross Validation Information Criterion; EBM, Event Based Model; ELISA, Enzyme Linked Immunosorbent Assay; GCA-F, Global Cortical Atrophy; GENFI, GENetic Frontotemporal dementia Initiative; MCI, Mild Cognitive Impairment; MMSE, Mini Mental State Examination; MTA, Medial Temporal Atrophy; MRI, Magnetic Resonance Imaging; NFT, Neurofibrillary Tangles; OASIS, Open Access Series of Imaging Studies; PA, Posterior Atrophy; pMCI, progressive Mild Cognitive Impairment; ROC, Receiver-Operator Characteristic; SMC, Subjective memory Complaints; sMCI, stable Mild Cognitive Impairment; SuStaIn, Subtype and Stage inference; TIV, Total Intracranial Volume; ViTA, Vienna Transdanube Aging.

and stage, mini mental state examination and amyloid-β_{1-42} cerebrospinal fluid concentration was proven to predict conversion from MCI to AD dementia on par with other novel statistical algorithms, with ROC curves that were not statistically different for the training and test sets and with area under curve respectively equal to 0.77 and 0.76. This study proves the transferability of a SuStaIn model for AD from research data to less-structured clinical cohorts, and indicates transferability to the clinical setting.

Keywords: alzheiemer's disease, patient subtyping, patient staging, SuStain model, inter-cohort validation

INTRODUCTION

Interest in the application of advanced statistics and machine learning (ML) in medicine has been constantly rising during the last years and their predictive capability allowed advancements in many fields. Particularly, data-driven approaches may contribute greatly to the advancement of neurosciences (Oxtoby et al., 2017; Ten Kate et al., 2018; Redolfi et al., 2020), where diseases are regularly modeled heuristically and patient care is influenced by clinicians' expertise (Braak and Braak, 1991; Jack et al., 2010; Jack et al., 2013).

Alzheimer's disease (AD) is one of the most impactful neurodegenerative diseases, affecting more than 50 million patients worldwide and costing healthcare systems $800 billion per year (Chan et al., 2019). The common underlying pathology of this disease is the combination of deposition of amyloid plaques with tau neurofibrillary tangles (NFT) (Braak and Braak, 1991), which is the driving cause of neurodegeneration and brain atrophy that leads to a progressive cognitive deterioration that affects multiple domains and eventually to a complete loss of function (Jack et al., 2010). Some basic questions still remain unresolved, such as: how homogeneous is AD? Is the course of progression more or less the same for most patients or are there significant variations?

Heuristic models of the temporal evolution of AD have been largely hypothesized (Braak and Braak, 1991; Jack et al., 2010; Jack et al., 2013), but most of these had the limitation of defining a mean average for the disease evolution that fits the majority of the AD patients. Instead, the phenomenology of AD is heterogeneous in terms of spatial distribution of tau NFT (Murray et al., 2011) and detecting rarer disease patterns may help in patient stratification, potentially allowing for specific drug targeting (ten Kate et al., 2018). Another major limitation of most heuristic and data driven models is the lack of validation in independent data, which is fundamental in order to translate models from the research setting to the clinical practice. For all these reasons well-validated ML tools are needed in order to promote advancements in clinical practice.

In recent years, the collection of numerous data sets containing demographic, clinical and biologic data of subjects from all stages of AD made possible the employment of statistical models and ML approaches (Oxtoby and Alexander, 2017). This context helped deploying disease models that allowed the definition of new strategies for biomarker-informed patient staging (Sperling et al., 2011). Among these algorithms, the family of event-based models (EBM) has been proven successful in defining discrete models for a wide battery of brain diseases (Young et al., 2015; Eshaghi et al., 2018; Wijeratne et al., 2018; Venkatraghavan et al., 2019; Firth et al., 2020; Oxtoby et al., 2021), showing utility in fine-grained staging of patients (Young et al., 2014). Generally, the assumption of these EBMs is that the sequence of events describing the disease progression is common for all subjects, which ignores the observed variation between individuals that may indicate the presence of subtypes of AD (Poulakis et al., 2020).

One key limitation of early subtyping approaches in literature (Whitwell et al., 2012; Nettiksimmons et al., 2014; Noh et al., 2014; Hwang et al., 2015), is that they do not account for temporal variation of the disease, implicitly assuming that all subjects were at the same disease stage.

SuStaIn (Young et al., 2018) (Subtype and Stage Inference) generalizes the EBM approach to include both subtyping and staging of subjects simultaneously, by using a full trajectory of change to define each subtype rather than a static pathology pattern. SuStaIn drops the basic EBM hypothesis of a single event sequence that fits all subjects, while also modeling the transition of biomarkers between different intermediate levels of severity rather than just changing from normal to abnormal. SuStaIn enables the discovery of different progression patterns that represent different manifestations of the same disease while avoiding the confounds of temporal change (Young et al., 2018).

However, SuStaIn has been tested so far only on well-defined research datasets or on synthetic data. Well-defined research datasets are not entirely representative of the general population (Ferreira et al., 2017) and transferability of a model to a less-structured clinical data is not granted a priori. In this paper we trained SuStaIn model on the well-defined research dataset of Alzheimer's disease Neuroimaging Initiative (ADNI) (Aisen et al., 2010), and we tested the subtyping and staging utility provided by the resulting disease model on a wider and heterogeneous data cohort composed of independent and less-well-phenotyped datasets representative of clinical settings and routine biomarker collection procedures. Our goal was to assess the transferability of a SuStaIn progression model from research data to an independent clinical data cohort coming from three different multi-centric data sets encompassing the entire AD spectrum that spans from early pre-clinical stages of cognitively normal (CN) elderly individuals to full blown dementia. This is a mandatory step in order to adopt SuStaIn and, more generally, advanced statistical models and ML tools in the clinical environment.

TABLE 1 | Characteristics of the data sets selected.

	Data Set	Full name	Description	Categories
Training Set	ADNI-1	Alzheimer's Disease Neuroimaging Initiative – 1	The Alzheimer's Disease Neuroimaging Initiative Aisen et al. (2010) is a longitudinal multicentre study designed to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of Alzheimer's disease (AD). ADNI was originally launched in 2003 as a public-private partnership; its primary goal has been to test whether magnetic resonance imaging (MRI), biological markers, clinical and neuropsychological assessments can be combined to measure the progression of MCI and Alzheimer's disease. The initial five-year study (ADNI-1) was extended by 2 years in 2009 by a Grand Opportunities grant (ADNI-GO), and in 2011 by further competitive renewal of the ADNI-1 grant (ADNI-2). Through its three phases, it has targeted participants with AD, different stages of MCI, and CN.	CN MCI AD SMC
	ADNI-GO	Alzheimer's Disease Neuroimaging Initiative – Grand Opportunities		MCI SMC
	ADNI-2	Alzheimer's Disease Neuroimaging Initiative – 2		CN MCI AD SMC
Test Set	OASIS	Open Access Series of Imaging Studies	OASIS Marcus et al. (2007) consists of I) a cross-sectional collection of 416 subjects. 100 of the included subjects, over the age of 60, have been clinically diagnosed with very mild to moderate Alzheimer's disease (AD). II) A longitudinal collection of 150 subjects aged from 60 to 96 years. Each subject was scanned on two or more visits, separated by at least 1 year for a total of 373 imaging sessions. In addition, the data set contains socio-demographic, clinical, and genotype information.	CN MCI AD
	PharmaCog (E-ADNI)	Prediction of cognitive properties of new drug candidates for neurodegenerative diseases in early clinical development	PharmaCog is an industry-academic (Innovative Medicines Initiative – IMI) European project aimed at identifying biomarkers sensitive to symptomatic and disease modifying effects of drugs for Alzheimer's disease Galluzzi et al. (2016). Several clinical sites participated in this study across Italy (Brescia, Verona, Milan, Perugia, and Genoa), Spain (Barcelona), France (Marseille, Lille, and Toulouse), Germany (Leipzig and Essen), Greece (Thessaloniki) and Netherland (Amsterdam). 151 MCI patients have been studied longitudinally for 3 years collecting multimodal image scans, clinical variables, and bio-specimens.	MCI AD
	ViTA	Vienna Transdanube Aging	ViTA is a population-based cohort-study of all 75-years old inhabitants of a geographically defined area of Vienna Fischer et al. (2002). VITA is composed of 606 subjects followed longitudinally for 4 years. Recruitment took place between May 2000 and October 2002. The primary focus of the VITA work-group was to establish a prospective age cohort for evaluation of prognostic criteria for the development of AD.	CN MCI AD

AD, Alzheimer's disease; CN, cognitively normal; MCI, mild cognitive impairment; SMC; subjective memory complaints.

MATERIALS AND METHODS

Participants

Data from a total of 1810 subjects gathered from various cohorts (**Table 1**) were used for this study. Subjects were divided into a training set, used to create the disease model, and a test set, used for model validation. The training set was composed of baseline data of 1043 subjects from the ADNI cohort that were either CN, affected by mild cognitive impairment (MCI) or AD dementia (**Table 2**), and were not affected by other major neurological diseases. Subjects diagnosed with subjective memory complaints (SMC) were included in the CN group since Mini-Mental State Examination (MMSE) score of these individuals was 28.1 ± 1.6. Diagnostic criteria used to identify MCI subjects were a clinical dementia rating (CDR) = 0.5 and a mini mental state examination (MMSE) (Tombaugh and McIntyre, 1992) score ≥24, while AD subjects were identified as all subjects with CDR ≥ 1 or subjects with CDR = 0.5 and MMSE<24.

Additionally, two subpopulations of subjects with longitudinal information, namely stable MCI subjects (sMCI) and progressive MCI (pMCI) were identified. Specifically, sMCIs were subjects for which only MCI diagnosis was reported for all available time-points and pMCIs were subjects that had at least one diagnosis of MCI and subsequently one diagnosis of AD and never reverted to MCI in the time-span of 10 years we considered.

The test set was composed of subjects coming from three independent data cohorts characterized by heterogenous and less-structured data collection. Specifically, subjects were selected from the Open Access Series of Imaging Studies (OASIS) (Marcus et al., 2007), PharmaCog (Galluzzi et al.,

TABLE 2 | Demographic, clinical, genetic and biological characteristics of the training and test sets.

		N	Age (years)	Sex (M/F)	Education (years)	MMSE (raw score)	$A\beta_{1-42}$ (positive/negative)	*APOE*-ε4 (carriers/non carriers)
Training set	CN	335	73.5 ± 5.9	46%/54%	16.3 ± 2.6	29.1 ± 1.2	40%/60%	27%/73%
	MCI	537	72.0 ± 7.2	59%/41%	16.0 ± 2.8	27.7 ± 1.8	66%/34%	51%/49%
	AD	171	73.4 ± 8.2	54%/46%	15.5 ± 2.7	23.4 ± 2.0	95%/5%	73%/27%
	Total	1043	72.7 ± 7.0	54%/46%	16.04 ± 2.7	27.4 ± 2.5	62%/38%	46%/54%
	sMCI	271	72.3 ± 7.1	58%/42%	16.1 ± 2.8	28.0 ± 1.7	56%/44%	42%/58%
	pMCI	205	73.1 ± 6.8	59%/41%	15.8 ± 2.8	27.2 ± 1.8	87%/13%	64%/36%
Test Set	CN	440	54 ± 25*	37%/63%*	8.4 ± 5.7*	29.0 ± 1.2	NA	2%/7%
	MCI	283	72.3 ± 7.6	46%/54*	9.2 ± 5.1*	26.3 ± 2.6*	34%/17%	19%/32%*
	AD	44	77.3 ± 7.4*	34%/66%*	5.8 ± 5.3*	21.7 ± 3.8*	NA	0%/5%
	Total	767	62 ± 21*	40%/60%*	8.6 ± 5.4*	27.2 ± 3.0	12%/6%	8%/16%*
	sMCI	152	71.2 ± 7.5	47%/53%	11.5 ± 4.2*	26.7 ± 2.2*	46%/25%	25%/43%
	pMCI	39	69.8 ± 6.4*	49%/51%	11.7 ± 3.9*	25.7 ± 2.4*	44%/5%	33%/26%

*Values from CN, MCI and AD contribute to the totals, MCI subpopulations of pMCIs and sMCIs are reported as well. Values marked with * on the test set are significantly different (p-value of ANOVA for continuous variables and chi-square for discrete variables <0.05) from the corresponding values from training set. Abbreviations: M, male; F, female; N, number.*

2016), and Vienna Transdanube Aging (ViTA) (Fischer et al., 2002) cohorts, totaling 767 subjects with the same clinical labels and diagnostic criteria as the training set. Populations of sMCIs and pMCIs were identified in the test sets with the same criteria as in the training set, but in this case the maximum time-span available was 7.5 years.

The training and test set populations were heterogeneous in terms of demographic, genetic and biological features (**Table 2**). The CN subjects in the test set were younger and less educated compared to the training set. The MCI subjects in the test set were less educated, and had higher prevalence of *APOE*-ε4 non-carriers compared to the training set's. Moreover, the pMCIs in the test set were younger than those in the training set. Finally, the AD dementia subjects in the test set were older and less educated compared to the corresponding subjects of the training set. Importantly, no statistical differences were reported in the frequency of abnormal cerebrospinal fluid (CSF) concentrations of amyloid-β_{1-42} ($A\beta_{1-42}$) protein between the test and the training sets for each diagnostic group. In all test set subgroups, with the exception of pMCIs, the gender prevalence was statistically different compared to the training set.

Clinical, Cognitive, Biological and Imaging Data

Clinical, cognitive, biological and imaging information were collected for each subject from the training and test set. Imaging information was derived from 1.5T or 3T T1-3D magnetic resonance imaging (MRI) scans, and was analyzed with Freesurfer 5.3 cross sectional stream (http://surfer.nmr.mgh.harvard.edu) with Desikan-Killiany atlas to obtain volumes of relevant brain regions of each subject, which were used to build the SuStaIn disease progression model. Freesurfer outputs were visually checked and validated by expert neuroscientists. The volumes of specific regions were used, specifically, we selected volumes of hippocampus, fusiform gyrus, entorhinal cortex, middle temporal cortex, precuneus, amygdala, insula, thalamus putamen, caudate, nucleus accumbens, pallidum and ventricles, which are among the most used regions employed in both heuristic and data driven currently available atrophy models for AD (Frisoni et al., 2010; Vemuri and Jack, 2010; Koval et al., 2018; Young et al., 2018; Archetti et al., 2019). For each region, volumes were obtained averaging the respective volume of the left and right hemisphere, volume of ventricles was obtained as the sum of 3rd and lateral ventricles. Cognitive information was provided by the MMSE score and was used as a proxy in order to verify that the disease model correlated with cognitive decline. Biological data included CSF concentration of $A\beta_{1-42}$ protein and it was used to identify a subpopulation of amyloid-negative healthy subjects defined as those CN subjects from the training set that had an $A\beta_{1-42}$ CSF concentration >192 pg/ml (Shaw et al., 2009). For the training set, $A\beta_{1-42}$ CSF concentration was obtained with Multiplex xMAP Luminex platform with Innogenetic immunoassay kit–based reagents (Kang et al., 2012). For demographic purposes $A\beta_{1-42}$ CSF concentration was collected for the test set subjects as well, but the CSF biomarker was only available for PharmaCog subjects. In this case, $A\beta_{1-42}$ CSF concentration was obtained with Enzyme Linked Immunosorbent Assay (ELISA) (Butler, 2000) which led to different CSF biomarkers distributions with respect to the training set. In order to tackle this issue, $A\beta_{1-42}$ CSF concentrations from PharmaCog were rescaled to match the mean and standard deviation of $A\beta_{1-42}$ distribution of training set subjects. The same cut-off value as the training set was used to define abnormality. As a compensation for inter-cohort demographic variability all volumetric measures for both training and test sets were corrected against the effect of age, sex, education (Gale et al., 2007), *APOE* genotype (Liu et al., 2013) and total intracranial volume (TIV) (Gur et al., 1991; Király et al., 2016) by means of multiple linear regression, and were converted into z-scores with respect to the mean and standard deviation defined by the volumes distribution of the healthy amyloid-negative subjects from the training set. Correction of biomarkers was performed separately for training set and test set.

Modelling

The disease progression model was built using the SuStaIn algorithm (Young et al., 2018), which generalizes the EBM approach (Fonteijn et al., 2012; Young et al., 2015) to allow for subtyping. Traditional EBMs rely on the assumption that it is possible to define a common sequence of events where, in the case of disease models, each event is defined as the value of a biomarker stepping from normality to abnormality. The normality and abnormality of the values are usually defined on the basis of biomarker distributions of healthy and diseased subjects. However, SuStaIn differs from classical EBM models in two main features:

1) The hypothesis of the common event sequence is relaxed in favor of multiple event sequences corresponding to a data-driven number of different disease subtypes that represent different disease trajectories of biomarker change observed in the training set. The optimal number of subtypes is determined using a popular model selection criterion called "Cross Validation Information Criterion" (CVIC) (Gelman et al., 2014).
2) Biomarkers are not treated as binary entities that are either normal or abnormal but all biomarker trajectories are modeled as a succession of z-scores progressing linearly toward abnormality.

Considering such modifications, the disease progression model is then represented by a set of sequences of integer z-scores for each biomarker, which represents the different disease subtypes. For this work z-scores were calculated with respect to the mean and standard deviation defined by the biomarker distribution of the healthy amyloid-negative ADNI subjects.

The maximum number of subtypes was set to 5 and the maximum value of z-scores for each biomarker was set to 3 (Young et al., 2018), meaning that maximum abnormality of each biomarker was reached when the z-score was >= 3.

When the disease progression model is defined, it is possible to outline the subtype that most likely fits any subject as the subtype for which the likelihood of a subject's z-scores projected on the z-score progression is maximized (Young et al., 2018). The subject is then staged on the most likely stage of the z-score progression defined by his or her subtype. The SuStaIn algorithm is publicly available in the form of a python package at the following link: http://europond.eu/software/.

Model Validation and Statistical Analysis

In order to investigate possible similarities with other subtyping methods, correlation between subtypes defined with SuStaIn and subtypes defined on the basis of visual rating scales of regional brain atrophy (Ferreira et al., 2019) was explored. Specifically, the visual scales considered were Scheltens' medial temporal atrophy (MTA) scale (Scheltens et al., 1992), Koedam's scale for Posterior Atrophy (PA) (Koedam et al., 2011) and Pasquier's frontal subscale of global cortical atrophy (GCA-F) (Pasquier et al., 1996; Scheltens et al., 1997).

According to visual ratings, typical AD was defined as abnormal MTA together with abnormal PA and/or abnormal GCA-F. Hippocampal-sparing was characterized by abnormal PA and/or abnormal GCA-F but normal MTA, while minimal atrophy AD was defined as normal scores in MTA, PA, and GCA-F. Limbic-predominant was defined as abnormal MTA alone with normal PA and GCA-F (Ferreira et al., 2017). All the visual ratings were computed automatically by means of the Automatic Visual Ratings of Atrophy (AVRA) tool (Mårtensson et al., 2019).

Further heuristic validation of SuStaIn was tested by exploring correlation of the subjects staging to the cognitive decline measured by means of MMSE.

The transferability of the model to new individuals was tested by subtyping and staging subjects from both the training and test sets on the basis of baseline volumes. Similarities between clinical, demographic, genetic and CSF features of subjects from the training and test sets assigned to different subtypes were explored by means of ANOVA and chi-square tests.

A subset of subjects (502 for the training set and 139 for the test set) were subtyped using 12-months visit biomarkers measurement in order to check the temporal consistency of the subtyping. Predictive capabilities of the model were tested by measuring the area under curve (AUC) of receiver-operator characteristic (ROC) curves obtained from classification of pMCIs and sMCIs from the training and test sets using various combinations of subtype, stage, MMSE and CSF $A\beta_{1-42}$ concentration as predictors in a multivariate logistic model. Statistical differences between ROC curves were tested by means of De Long test (DeLong et al., 1988). All ROC analyses were computed using R (version 3.5.1).

Chi-square and ANOVA tests ($\alpha = 0.05$) were performed in python (version 3.6.9) to test differences between the diagnostic groups and subtypes.

RESULTS

The disease model identified by SuStaIn consisted of three disease subtypes (**Figure 1**). The first disease subtype ("Subtype 1" in the next sections), is characterized by abnormality (Z-score = 1) that can be observed in the ventricles first, then atrophy occurs in the hippocampus and entorhinal cortex, that are also the first regions to show full abnormality (Z-score = 3) alongside amygdala. Interestingly, ventricles are also the last regions to show full abnormality meaning a relatively slow but persistent volumetric expansion process that tracks the disease progression.

The second disease subtype ("Subtype 2" in the next sections) shows an atrophy pattern where abnormality starts in thalamus and pallidum (Z-score = 1). Subsequently, atrophy can be observed in caudate, putamen, insula, precuneus and then fusiform gyrus and middle-temporal cortex and hippocampus which is the first biomarker to become fully abnormal (Z-score = 3). In this subtype, ventricles start expanding later than in Subtype 1. The third subtype ("Subtype 3" in the next sections) shows an atrophy pattern where ventricles become fully abnormal before atrophy starts in almost all the other

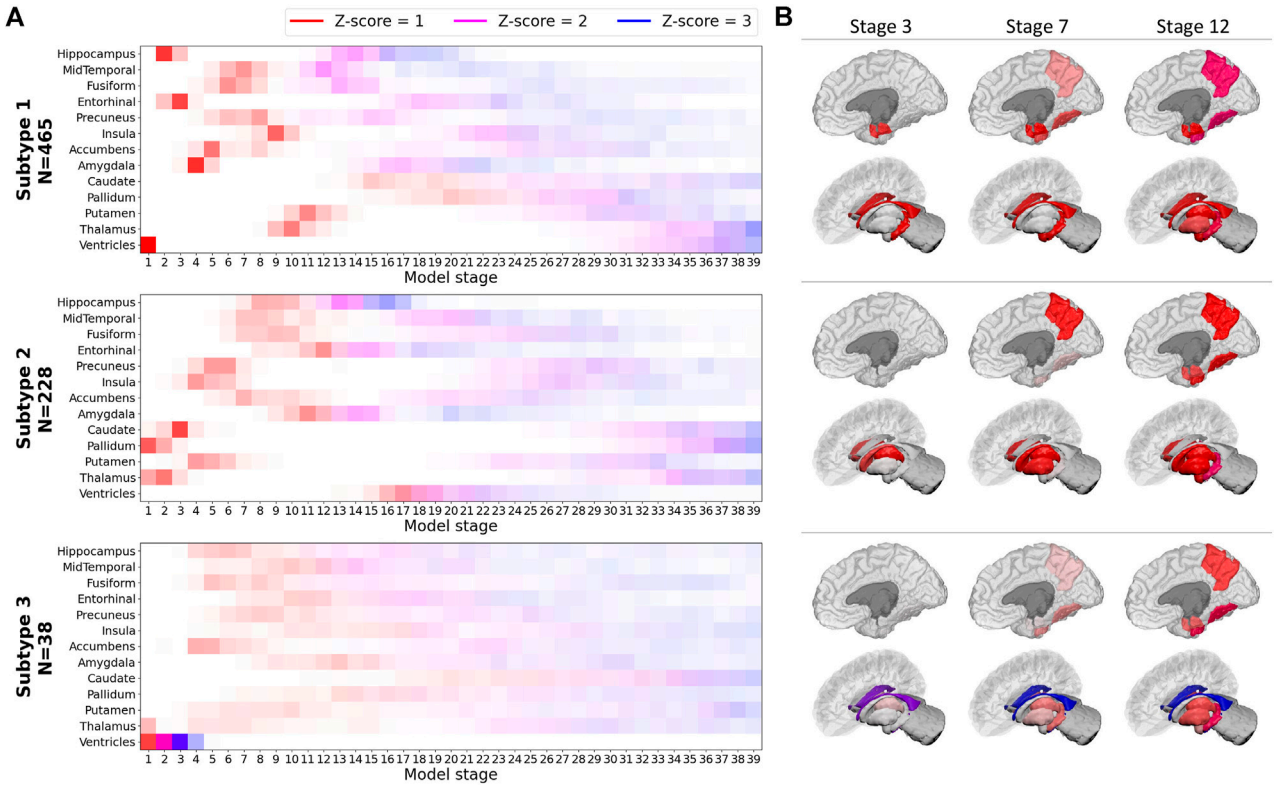


FIGURE 1 | SuStaIn model built on the basis of volumetric biomarkers of the training set.**(A)** Z-score progression patterns for each subtype. Color shades indicate the probability of a Z-score to increment, "N" indicates the number of subjects from the training set assigned to each subtype **(B)** Representations of early stages for each subtype.

regions, for which a less-defined atrophic progression is manifested in comparison to Subtypes 1 and 2.

SuStaIn subtypes were cross linked to AVRA ratings to evaluate whether similarities between subtypes defined by the two methods exist (**Figure 2**). Subtype 1 was mainly characterized by the "Typical AD" atrophy pattern (Ferreira et al., 2019); Subtype 2 showed an equal predominance of the hippocampal-sparing variant; Subtype 3 showed a limbic-predominant subtype. The minimal atrophy subtype (Ferreira et al., 2020) was most consistent with Subtypes 1 and 2. After correcting against effects of sex, age and TIV, relevant differences (p-value for ANOVA <0.05) in volume of hippocampus were observed between subjects from Subtypes 1 and 2 labeled with minimal atrophy according to the AVRA scores (**Figure 3**), with subjects from Subtype 2 exhibiting larger volumes. Subjects with minimal atrophy from Subtype 3 are not reported as they are not enough for statistical significance.

Differences in AVRA visual scores between subtypes were inferred *via* a linear regression model of visual scores vs. model stage (**Supplementary Figure S1**). No relevant subtype differences were observed for GCA. MTA was shown to progress significantly faster for Subtype 2 than Subtypes 1 and 3. Subtype 3 also showed a significantly faster progression of the PA scale. Subjects from each diagnostic category of both training and test sets that were assigned to a specific subtype are shown in **Table 3**. Subjects that were in stage 0 or in the final stage were excluded from the subtyping as these stages are equivalent for each subtype. In each diagnostic group, the majority of subjects were on average assigned to the typical subtype (65% for training set and 82% for the testing set). A minority of the subjects were assigned to the hippocampal sparing subtype, specifically 30% of the training set and 16% for the test set, while only a limited number of subjects for each dataset were assigned to the limbic subtype (5% for the training set and 2% for the test set). For both sets, subjects from each diagnostic category were staged on average at stages that mirror the worsening of their clinical condition (**Table 3**), with the exception of pMCIs and sMCIs from Subtype 3.

Significant differences between subtypes were observed for demographic, clinical, biological and genetic variables (**Table 4**). For each subtype, subjects from all diagnostic categories were considered. In both training and test sets, subjects from Subtype 2 were on average more educated and a larger portion of them were male with respect to subjects from Subtype 1. Similarly, subjects from Subtype 3 had a lower MMSE with respect to Subtype 2. In the training set, where CSF data was widely available, the portion of subjects that had an abnormal $A\beta_{1\text{-}42}$ CSF concentration was significantly lower with respect to the other subtypes. This effect was not observed in the test set for the small number of subjects for which $A\beta_{1\text{-}42}$ is available.

Subtyping consistency of the SuStaIn progression model was tested by comparing subtyping of subjects for which 12-months

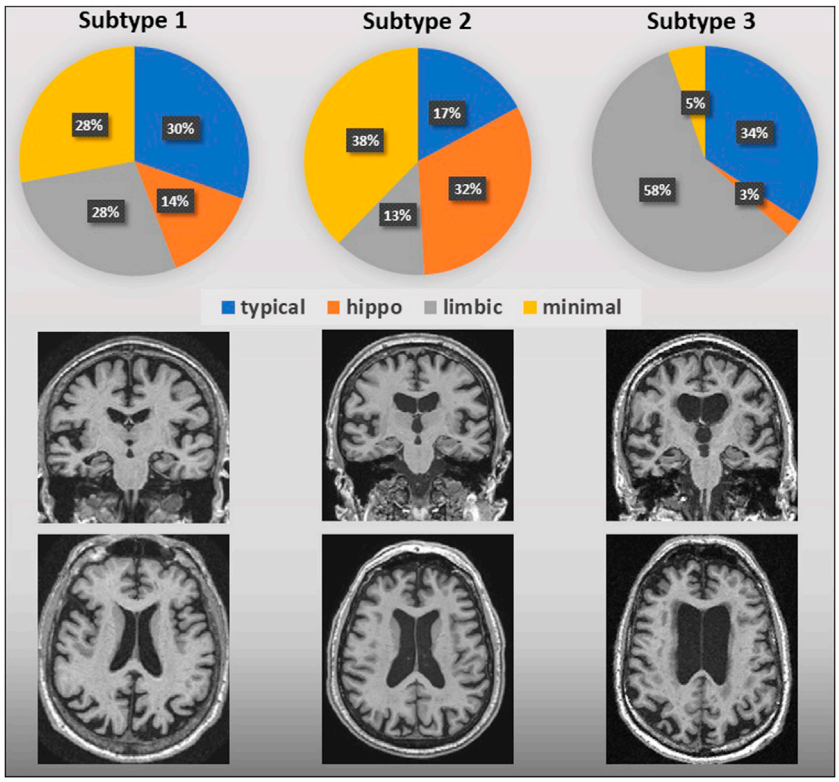


FIGURE 2 | AVRA vs. SuStaIn subtypes of AD. Pie graphs represent the percentage of AVRA subtypes subjects for each SuStaIn subtype. Regional atrophy in AVRA was measured with the MTA, PA and GCA-F scales based on T1-3D weighted images; below, visual examples of the SuStaIn atrophy subtypes are shown.

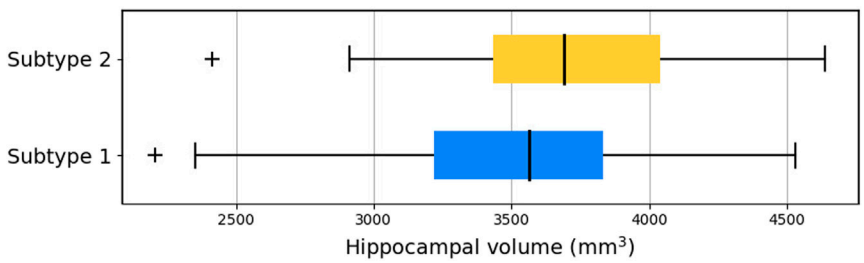


FIGURE 3 | Hippocampal volume of subjects from Subtypes 1 and 2 labeled with minimal atrophy according to AVRA scores. Hippocampal volumes were averaged between right and left hemisphere for simpler representation.

follow-up was available (502 for the training test and 140 for the test set). Few subjects were subtyped to a different group at 12-months follow up (**Figure 4**), with only 11% of training set subjects and 9% of test set subjects assigned to different subtypes. Changes occurred mainly between subtypes 1 and 2 in both training and test sets. For subjects with stable subtype assignment, stage progression was relatively slow in time showing an average progression of 0.8 ± 1.5 stages over the 12-month period.

The disease progression signature defined by Subtype 1 showed good correlation with cognitive performance measured by MMSE (**Figure 5**), with $R^2 = 0.74$ for the training set and $R^2 = 0.82$ for the test set. Similarly, good correlations were registered in Subtype 2 ($R^2 = 0.85$ training set; $R^2 = 0.87$ test set) and Subtype 3 ($R^2 = 0.85$ training set; $R^2 = 0.76$ test set).

Classification of pMCIs and sMCIs, based on subtype and stage retuned ROCs with AUC = 0.67 for the training set and 0.72 for the test set. The combination of subtype and stage with other predictors tracking different aspects of the disease, namely the MMSE and CSF concentration of $A\beta_{1\text{-}42}$ protein, returned a better classification performance than the subtype and stage model alone, with AUC = 0.77 for the training set and AUC = 0.76 for the test set, outperforming also a model that accounts only for MMSE and $A\beta_{1\text{-}42}$ (AUC = 0.72 for the training set and AUC = 0.74 for the test set) and a model that accounts for AVRA

TABLE 3 | Number and percentage of subjects from each diagnostic category assigned to each subtype.

		Subtype 1		Subtype 2		Subtype 3	
		N	**Average Stage**	**N**	**Average Stage**	**N**	**Average Stage**
Training Set	CN	96 (54%)	3 ± 3	74 (41%)	3 ± 3	9 (5%)	4 ± 1
	MCI	243 (62%)	5 ± 4	128 (33%)	5 ± 5	22 (5%)	7 ± 4
	AD	126 (79%)	8 ± 5	26 (16%)	9 ± 6	7 (5%)	11 ± 5
	sMCI	111 (59%)	4 ± 4	67 (36%)	4 ± 4	10 (5%)	8 ± 5
	pMCI	116 (69%)	6 ± 5	44 (26%)	7 ± 6	8 (5%)	7 ± 4
Test Set	CN	303 (86%)	5 ± 4	37 (11%)	5 ± 4	9 (3%)	5 ± 2
	MCI	185 (78%)	7 ± 6	49 (21%)	6 ± 4	3 (1%)	9 ± 2
	AD	41 (95%)	9 ± 7	1 (2.5%)	12	1 (2.5%)	9
	sMCI	83 (68%)	7 ± 6	35 (29%)	5 ± 4	4 (3%)	8 ± 3
	pMCI	32 (84%)	9 ± 6	6 (16%)	11 ± 4	0 (0%)	NA

TABLE 4 | Descriptive statistics of the demographic, clinical, biological and genetic variables of subjects for each subtype

		Age (years)	**Sex (M/F)**	**Education (years)**	**MMSE (raw score)**	**$A\beta_{1-42}$ (positive/negative)**	***APOE*-ε4 (carriers/ non carriers)**
Training Set	Subtype 1	72.5 ± 7.2[a]	48%/52%[a]	15.9 ± 2.7[a]	26.6 ± 2.6[a]	72%/28%[a]	48%/52%
	Subtype 2	73.8 ± 6.7[a]	84%/16%[a,b]	16.4 ± 2.7[a]	27.9 ± 2.0[a,b]	53%/46%[a,b]	44%/56%
	Subtype 3	74.8 ± 6.2	61%/39%[b]	15.9 ± 3.0	26.5 ± 2.6[b]	79%/21[b]	42%/58%
Test Set	Subtype 1	60 ± 24[c]	41%/59%[a]	8.5 ± 5.4[a]	26.7 ± 3.3	3%/10%	8%/14%
	Subtype 2	63 ± 17[b]	64%/36%[a]	10.4 ± 5.6[a]	27.4 ± 2.2[b]	31%/15%	24%/25%
	Subtype 3	74.4 ± 5.7[c,b]	46%/54%	7.9 ± 6.1	25.7 ± 4.7[b]	15%/0%	0%/15%

Values marked with [a]indicate significant differences (p-value < 0.05) between Subtype 1 and Subtype 2 values in the same set; values marked with [c]indicate significant differences (p-value < 0.05) between Subtype 1 and Subtype 3 values in the same set; values marked with [b]indicate significant differences (p-value < 0.05) between Subtype 2 and Subtype 3 values in the same set.

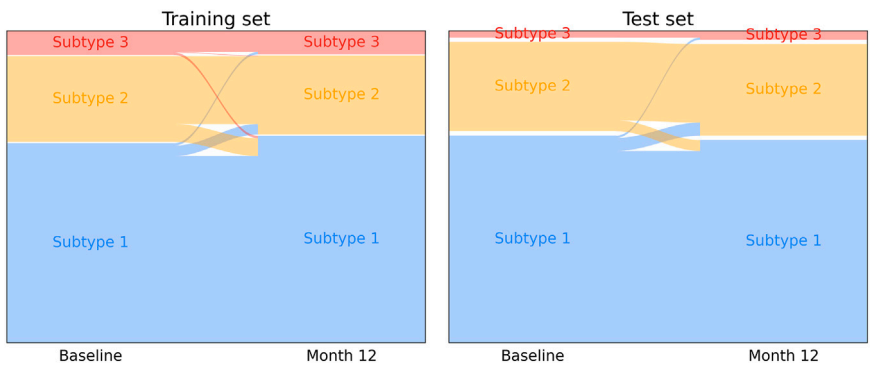


FIGURE 4 | Longitudinal subtype consistency for training set subjects **(left)** and test set subjects **(right)** over a 12-months follow-up period.

subtype, MMSE and $A\beta_{1-42}$ (AUC = 0.72 for the training set, unavailable for the test set). Notably, for each predictor combination no statistically significant differences were observed between ROC curves (**Supplementary Figure S2**) of the training and test sets (p-value of DeLong test >0.05).

DISCUSSION

In this study, we tested the transferability of a SuStaIn AD progression model among clinical data cohorts. The disease progression model trained on volumetric imaging markers from an observational research study estimated three AD-related atrophy patterns. Previously, SuStaIn was only tested on research datasets, such as ADNI and GENetic Frontotemporal dementia Initiative (GENFI) or synthetic data (Young et al., 2018), while in the present study we demonstrated model transferability to clinical cohorts through stable and consistent subtyping.

Subtype 1 mirrored the typical course of AD as supposed in heuristic models and as found in previous EBM and data-driven models (Young et al., 2015; Archetti et al., 2019; Venkatraghavan et al., 2019), according to which hippocampus is one of the earliest regions to show considerable atrophy. This subtype also

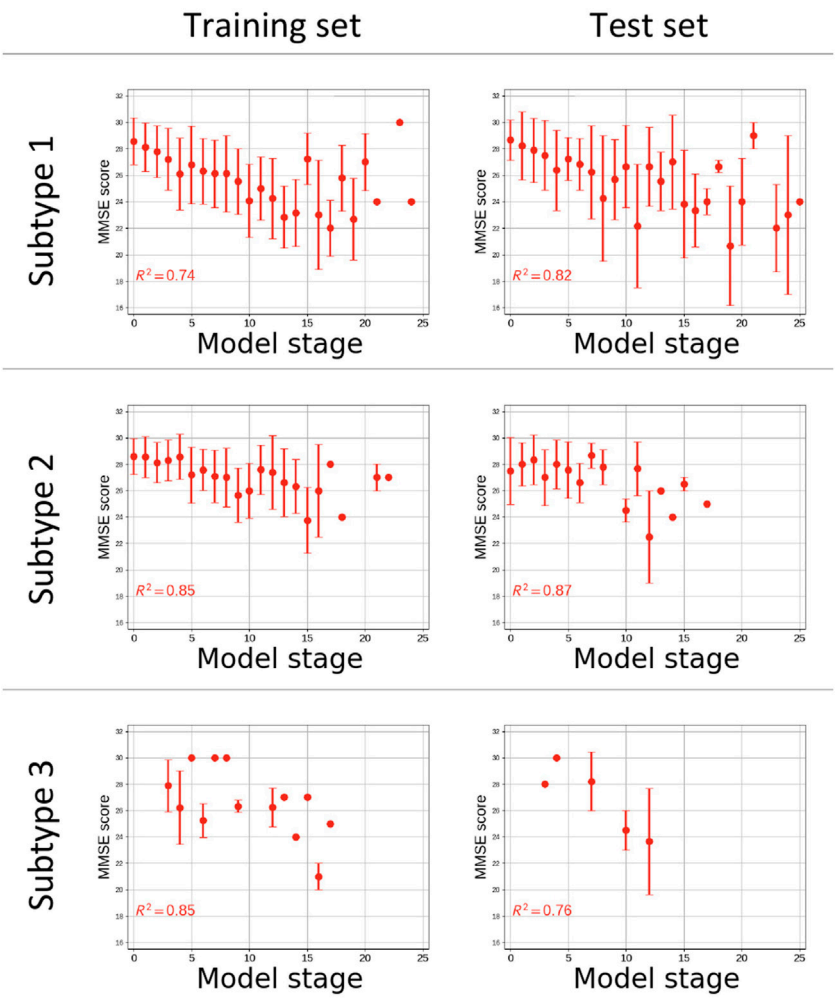


FIGURE 5 | Plot of Cognitive performance measured by Mini Mental State Examination (MMSE) vs. the estimated disease stage subjects from the training **(left)** and test **(right)** sets for each subgroup. Coefficients of determination (R^2) of the linear regression of MMSE score vs. disease stage are reported. The *x*-axes are only reported up to stage 25 of 39 as no subjects were staged beyond.

shares similarities with the typical subtype as defined in the original SuStaIn work (Young et al., 2018) for which hippocampus and amygdala are among the first regions to show atrophy. The correspondence of Subtype 1 with the canonical and most prevalent manifestation of AD (Braak and Braak, 1991), is reinforced by our subject subtyping results, with the majority of subjects assigned to this subtype in both training and test set. In particular, the proportions of AD subjects of the training and testing set assigned to Subtype 1, 79% and 95% respectively, are greater than those from other diagnostic categories. Subtype 1 is also majorly prevalent as assignment of pMCIs, with a proportion of 69% compared to the other diagnostic categories.

Subtype 2 shows similarities with the hippocampal-sparing variant of AD characterized by a relative sparing of the medial temporal lobe as observed in previous works (Murray et al., 2011; Whitwell et al., 2012; Ferreira et al., 2019; Krajcovicova et al., 2019). In this subtype hippocampus starts becoming abnormal after most of the others deep gray matter structures, with loss predominantly focused in the insula, caudate nucleus and parietal cortex. The similarity also extends to the demographic characteristics of this group, that is characterized by a higher prevalence of male subjects as reported in previous works (Ferreira et al., 2020). In this subtype, pallidum, putamen and caudate are among the first regions to show atrophy as observed in the subcortical subtype defined in the original SuStaIn work (Young et al., 2018).

Subtype 3 is characterized by a broader atrophy signature with less distinct ordering than the other subtypes, with the exception of ventricles expansion that was clearly the first marker to become abnormal. In this atypical subtype, atrophy seems to progress simultaneously in most brain regions. Subtype 3 was observed in a minority of subjects when considering our whole cohort. These subjects exhibit similarities with the limbic predominant subtype of AD (Ferreira et al., 2017). Also, Subtype 3 might have some characteristics in common with other subtypes as some subjects

had been labeled as belonging to the typical AD subtype (Ferreira et al., 2017; Persson et al., 2017). Alternatively, it is possible that this group does not reflect a distinct AD subtype but just includes a subgroup of subjects whose ventricles outlie the normal distribution of ventricles in healthy subjects.

The atrophy subtypes of AD have been assessed *via* visual rating scales in several previous studies (Ferreira et al., 2020). AVRA is a method to automatically quantify these visual rating scales, which was used just on ADNI data, therefore it represented the ideal tool to find a correlate between a clinically used subtyping method and the SuStaIn data driven definition performed on our training dataset. We have produced the first comparison of data-driven subtyping results using a disease progression model (SuStaIn) with existing progression-ignorant methods of visual ratings and AVRA. Partial agreement was observed between SuStain and AVRA subtypes on an individual level, and differences may be imputed to the selection of brain regions used to train SuStaIn, that do not cover entirely the same brain region used to assess visual ratings and to a general lack of harmonization of subtyping methods (Mohanty et al., 2020). SuStaIn proved to offer a finer-grained representation of different atrophy patterns as relevant differences in hippocampal volume were observed between subjects from subtypes 1 and 2 that were labeled with minimal atrophy according to the AVRA scores.

The temporal consistency of SuStaIn subtyping was tested on subjects from the training and test sets for which a 12-months follow-up visit was available. The test resulted in excellent consistency with only 10% of subjects receiving a different subtype assignment across different visits. Since disease stage was relatively stable across the 12-months interval for individuals with stable subtype, the excellent subtype consistency was expected.

Once subjects from all subtypes were staged on the respective disease progression sequence, the SuStain stage showed good linear correlation (Perneczky et al., 2006) with general cognitive decline on the MMSE (Tombaugh and McIntyre, 1992) test, particularly for Subtypes 1 and 2, and the ceiling effect that was observed in previous studies (Hoops et al., 2009; Archetti et al., 2019) was not detected, likely due to the absence of early markers of AD in the model, such as CSF markers.

SuStaIn subtype and stage predicted conversion of MCI subjects to AD with an AUC comparable to other novel statistical algorithms (Ramírez et al., 2018; Salvatore et al., 2018). The combination of multiple predictors proved to be key in improving classification performance as classification based on subtype and stage alone or on MMSE and $A\beta_{1\text{-}42}$ alone yielded a lower classification performance. Importantly, classification task performed similarly in the training and test set for each combination of predictors, thus giving a first indication of the transferability of SuStaIn disease models and its use in deep patient phenotypization for future clinical trials as well.

The interpretation of the atrophy subtypes still remains an open issue as solid subtyping ground truth in AD is lacking, since heuristic models such as Jack's (Jack et al., 2010) or Braak's (Braak and Braak, 1991) are more aimed at defining a common disease trajectory rather than detecting different atrophy patterns. Also, the model presented here differs slightly from the AD model presented in the original SuStaIn work (Young et al., 2018), and this difference is provoked by choice of different brain regions as input data for the two models and partially due to the different purpose of this study.

Previous works based on cross sectional models were able to reach better classification performances across a wide range of neurological diseases (Willette et al., 2014; Archetti et al., 2019), but in all cases the models were built ab initio using multi-modal markers accounting for biological features and cognitive scores, while we used CSF and cognitive data only for post-hoc analyses. In the present study, we chose to exclude CSF measurements and cognitive scores because these markers were available only for a small portion of subjects used as test set.

The most important limitation of the present work is the relatively small number of subjects used to train and test the model. The small number of subjects particularly affects the characterization of rarer subtypes, that cannot be modeled as accurately as common subtypes. Also, the small number of subjects considered to assess the predictive value of the model prevented us from assessing with a usual power level measures of sensitivity and specificity for the classification of pMCIs and sMCIs.

An important limitation of the model is the relatively low AUC reached in the classification of pMCIs vs. sMCIs, indeed the AUC could be improved with the inclusion of CSF and cognitive scores for the model building phase rather than using them for post hoc analyses (Archetti et al., 2019), but those biomarkers were excluded from the model building as they should not be important factors in atrophy subtype identification. Moreover, CSF and cognitive scores are more easily affected by inter-cohort and inter-centre harmonization issues (Costa et al., 2017; Delaby et al., 2020) thus requiring a more thorough model validation. Therefore, MRI-only models are more suitable for near-future implementation of SuStaIn-based models in tools for subtype detection in single case-scenarios.

Another key factor affecting the AUCs is the unavailability of the characterization in amnestic and non-amnestic MCI for the major portion of the subjects. The condition of amnestic MCI is a more typical prodromal stage for AD that could provide better classification performances (Cousins et al., 2020). Also, the use of amnestic MCIs for the training process could indeed generate a more accurate disease model that better depicts the transition phase from MCI to dementia.

Future work will concentrate efforts in modeling subtypes using larger and more diverse cohorts, that will allow for a more precise definition of subtypes and for a finer-grade characterization of subjects belonging to each subgroup. Another key factor for an optimal definition of the subtypes is the selection of brain regions, and future work will investigate the optimal choice to obtain a disease model that is descriptive and informant without being redundant and trying to maximize the individual match between AVRA subtypes and SuStaIn subtypes. SuStaIn is a suitable approach to build disease models that include non-imaging markers, and future work will investigate the possibility of defining AD progression subtypes based on CSF markers and cognitive scores coupled with imaging markers, possibly linking subtypes with demographic genetic and lifestyle factors.

There are ongoing efforts to extend this work toward full clinical translation. This includes implementing SuStaIn progression models in user-friendly interfaces, external independent validation studies, and usability assessments from clinicians, all of which form key components of the EuroPOND (http://europond.eu/) and E-DADS initiatives (https://e-dads.github.io/).

CONCLUSIONS

We have demonstrated that a data-driven subtyping model (Young et al., 2018) of Alzheimer's disease progression trained on research-quality MRI (ADNI) is transferable to lower-quality clinical data (PharmaCog, OASIS, ViTA). This is an encouraging result motivated by the expectation that, in the near future, healthcare will increasingly adopt data-driven and ML models in daily clinical practice. Indeed, the validation and generalization of such models on independent datasets is a proof of concept required for their translation from research settings to clinical environments. Open questions remain about the biological mechanisms underpinning Alzheimer's disease subtypes, which will be an important focus of future studies, including ongoing drug-development efforts.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: adni.loni.usc.edu, neugrid2.eu.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. Written informed consent for participation was not required for this study in accordance with the national legislation and the institutional requirements.

CONTRIBUTORS' CONTACT

Alexandra L. Young, alexandra.young@kcl.ac.uk; Neil P. Oxtoby, n.oxtoby@ucl.ac.uk; Daniel Ferreira, daniel.ferreira.padilla@ki.se; Gustav Mårtensson, gustav.martensson@ki.se; Eric Westman, eric.westman@ki.se; Daniel C. Alexander, d.alexander@ucl.ac.uk; Giovanni B. Frisoni, giovanni.frisoni@unige.ch; Alberto Redolfi, aredolfi@fatebenefratelli.eu

AUTHOR CONTRIBUTIONS

DA: conceptualization, investigation, methodology, formal analysis, investigation, validation, writing. ALY: software, methodology. NPO: conceptualization, methodology, project administration. DF: data, resources. GM: data, resources. EW: data, resources. DCA: resources, supervision, project administration. GBF: resources, supervision. AR: supervision, conceptualization, resources, writing - review - editing. All authors contributed to the article and approved the submitted version.

FUNDING

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 666992. This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 634541. ADNI data were funded by the Alzheimer's Disease Neuroimaging Initiative (National Institutes of Health grant U01 AG024904) and Department of Defense Alzheimer's Disease Neuroimaging Initiative (Department of Defense award W81XWH-12-2-0012). The Alzheimer's Disease Neuroimaging Initiative is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals Inc.,; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research and Development LLC; Johnson & Johnson Pharmaceutical Research & Development LLC; Lumosity; Lundbeck; Merck and Co Inc.; Meso Scale Diagnostics LLC; NeuroRx Research; Neuro-track Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. Alzheimer's Disease Neuroimaging Initiative data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California. The investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf. ViTA, and PharmaCog (alias E-ADNI) data used in the preparation of this article were obtained from NeuGRID2 initiative (https://neugrid2.eu) funded by grant 283562 from the European Commission. OASIS was funded by grant P50 AG05681, P01

AG03991, R01 AG021910, P50 MH071616, U24 RR021382, R01 MH56584. NPO is a UKRI Future Leaders Fellow (MR/S03546X/1). NPO and DCA were supported by the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.661110/full#supplementary-material

REFERENCES

Aisen, P. S., Petersen, R. C., Donohue, M. C., Gamst, A., Raman, R., Thomas, R. G., et al. (2010). Clinical Core of the Alzheimer's Disease Neuroimaging Initiative: Progress and Plans. *Alzheimer's Demen.* 6 (3), 239–246. doi:10.1016/j.jalz.2010.03.006

Archetti, D., Ingala, S., Venkatraghavan, V., Wottschel, V., Young, A. L., Bellio, M., et al. (2019). Multi-study Validation of Data-Driven Disease Progression Models to Characterize Evolution of Biomarkers in Alzheimer's Disease. *NeuroImage: Clin.* 24, 101954. doi:10.1016/j.nicl.2019.101954

Braak, H., and Braak, E. (1991). Neuropathological Stageing of Alzheimer-Related Changes. *Acta Neuropathol.* 82 (4), 239–259. doi:10.1007/BF00308809

Butler, J. E. (2000). Enzyme-Linked Immunosorbent Assay. *J. Immunoassay* 21 (2-3), 165–209. doi:10.1080/01971520009349533

Chan, K. Y., Adeloye, D., Asante, K. P., Calia, C., Campbell, H., Danso, S. O., et al. (2019). Tackling Dementia Globally: the Global Dementia Prevention Program (GloDePP) Collaboration. *J. Glob. Health* 9 (2), 020103. doi:10.7189/jogh.09.020103

Costa, A., Bak, T., Caffarra, P., Caltagirone, C., Ceccaldi, M., Collette, F., et al. (2017). The Need for Harmonisation and Innovation of Neuropsychological Assessment in Neurodegenerative Dementias in Europe: Consensus Document of the Joint Program for Neurodegenerative Diseases Working Group. *Alz Res. Ther.* 9, 27. doi:10.1186/s13195-017-0254-x

Cousins, K. A. Q., Irwin, D. J., Wolk, D. A., Lee, E. B., Shaw, L. M. J., Trojanowski, J. Q., et al. (2020). ATN Status in Amnestic and Non-amnestic Alzheimer's Disease and Frontotemporal Lobar Degeneration. *Brain* 143 (7), 2295–2311. doi:10.1093/brain/awaa165

Delaby, C., Teunissen, C. E., Alcolea, D., Amar, E. B., Beaume, A., Bedel, A., et al. (2020). International Initiative for Harmonization of Cerebrospinal Fluid Diagnostic Comments in Alzheimer's Disease. *Alzheimer's Demen.* 16, e047209. doi:10.1002/alz.047209

DeLong, E. R., DeLongClarke-Pearson, D. M. D., and Clarke-Pearson, D. L. (1988). Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves: a Nonparametric Approach. *Biometrics* 44 (3), 837–845. doi:10.2307/2531595

Eshaghi, A., Marinescu, R. V., Young, A. L., Firth, N. C., Prados, F., Jorge Cardoso, M., et al. (2018). Progression of Regional Grey Matter Atrophy in Multiple Sclerosis. *Brain* 141 (6), 1665–1677. doi:10.1093/brain/awy088

Ferreira, D., Hansson, O., Barroso, J., Molina, Y., MachadoHernandez-Cabrera, A., Hernández-Cabrera, J. A., et al. (2017). The Interactive Effect of Demographic and Clinical Factors on Hippocampal Volume: A Multicohort Study on 1958 Cognitively Normal Individuals. *Hippocampus* 27 (6), 653–667. doi:10.1002/hipo.22721

Ferreira, D., Nordberg, A., and Westman, E. (2020). Biological Subtypes of Alzheimer Disease. *Neurology* 94 (10), 436–448. doi:10.1212/WNL.0000000000009058

Ferreira, D., Pereira, J. B., Volpe, G., and Westman, E. (2019). Subtypes of Alzheimer's Disease Display Distinct Network Abnormalities Extending beyond Their Pattern of Brain Atrophy. *Front. Neurol.* 10, 524. doi:10.3389/fneur.2019.00524

Ferreira, D., Verhagen, C., Hernández-Cabrera, J. A., Cavallin, L., Guo, C.-J., Ekman, U., et al. (2017). Distinct Subtypes of Alzheimer's Disease Based on Patterns of Brain Atrophy: Longitudinal Trajectories and Clinical Applications. *Sci. Rep.* 7, 46263. doi:10.1038/srep46263

Firth, N. C., Primativo, S., Brotherhood, E., Young, A. L., Yong, K. X., Crutch, S. J., et al. (2020). Sequences of Cognitive Decline in Typical Alzheimer's Disease and Posterior Cortical Atrophy Estimated Using a Novel Event-based Model of Disease Progression. *Alzheimer's Demen.* 16 (7), 965–973. doi:10.1002/alz.12083

Fischer, P., Jungwirth, S., Krampla, W., Weissgram, S., Kirchmeyr, W., Schreiber, W., et al. (2002). Vienna Transdanube Aging "VITA": Study Design, Recruitment Strategies and Level of Participation. *J. Neural Transm. Suppl.* 62, 105–116. doi:10.1007/978-3-7091-6139-5_11

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., et al. (2012). An Event-Based Model for Disease Progression and its Application in Familial Alzheimer's Disease and Huntington's Disease. *Neuroimage* 60 (3), 1880–1889. doi:10.1016/j.neuroimage.2012.01.062

Frisoni, G. B., Fox, N. C., Jack, C. R., Scheltens, P., and Thompson, P. M. (2010). The Clinical Use of Structural MRI in Alzheimer Disease. *Nat. Rev. Neurol.* 6 (2), 67–77. doi:10.1038/nrneurol.2009.215

Gale, S. D., Baxter, L., Connor, D. J., Herring, A., and Comer, J. (2007). Sex Differences on the Rey Auditory Verbal Learning Test and the Brief Visuospatial Memory Test-Revised in the Elderly: Normative Data in 172 Participants. *J. Clin. Exp. Neuropsychol.* 29 (5), 561–567. doi:10.1080/13803390600864760

Galluzzi, S., Marizzoni, M., Babiloni, C., Albani, D., Antelmi, L., Bagnoli, C., et al. (2016). Clinical and Biomarker Profling of Prodromal Alzheimer's Disease in Workpackage 5 of the Innovative Medicines Initiative PharmaCog Project: a 'European ADNI Study'. *J. Intern. Med.* 279 (6), 576–591. doi:10.1111/joim.12482

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding Predictive Information Criteria for Bayesian Models. *Stat. Comput.* 24, 997–1016. doi:10.1007/s11222-013-9416-2

Gur, R. C., Mozley, P. D., Resnick, S. M., Gottlieb, G. L., Kohn, M., Zimmerman, R., et al. (1991). Gender Differences in Age Effect on Brain Atrophy Measured by Magnetic Resonance Imaging. *Proc. Natl. Acad. Sci.* 88 (7), 2845–2849. doi:10.1073/pnas.88.7.2845

Hoops, S., Nazem, S., Siderowf, A. D., Duda, J. E., Xie, S. X., Stern, M. B., et al. (2009). Validity of the MoCA and MMSE in the Detection of MCI and Dementia in Parkinson Disease. *Neurology* 73 (21), 1738–1745. doi:10.1212/WNL.0b013e3181c34b47

Hwang, J., Kim, C. M., Jeon, S., Lee, J. M., Hong, Y. J., Roh, J. H., et al. (2015). Prediction of Alzheimer's Disease Pathophysiology Based on Cortical Thickness Patterns. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 2, 58–67. doi:10.1016/j.dadm.2015.11.008

Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., et al. (2013). Tracking Pathophysiological Processes in Alzheimer's Disease: an Updated Hypothetical Model of Dynamic Biomarkers. *Lancet Neurol.* 12 (2), 207–216. doi:10.1016/S1474-4422(12)70291-0

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological Cascade. *Lancet Neurol.* 9 (1), 119–128. doi:10.1016/S1474-4422(09)70299-6

Kang, J.-H., Vanderstichele, H., Trojanowski, J. Q., and Shaw, L. M. (2012). Simultaneous Analysis of Cerebrospinal Fluid Biomarkers Using Microsphere-Based xMAP Multiplex Technology for Early Detection of Alzheimer's Disease. *Methods* 56 (4), 484–493. doi:10.1016/j.ymeth.2012.03.023

Király, A., Szabó, N., Tóth, E., Csete, G., Faragó, P., Kocsis, K., et al. (2016). Male Brain Ages Faster: the Age and Gender Dependence of Subcortical Volumes. *Brain Imaging Behav.* 10 (3), 901–910. doi:10.1007/s11682-015-9468-3

Koedam, E. L. G. E., Lehmann, M., Van Der Flier, W. M., Scheltens, P., Pijnenburg, Y. A. L., Fox, N., et al. (2011). Visual Assessment of Posterior Atrophy Development of a MRI Rating Scale. *Eur. Radiol.* 21 (12), 2618–2625. doi:10.1007/s00330-011-2205-4

Koval, I., Schiratti, J.-B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., et al. (2018). Spatiotemporal Propagation of the Cortical Atrophy: Population and Individual Patterns. *Front. Neurol.* 9, 235. doi:10.3389/fneur.2018.00235

Krajcovicova, L., Klobusiakova, P., and Rektorova, I. (2019). Gray Matter Changes in Parkinson's and Alzheimer's Disease and Relation to Cognition. *Curr. Neurol. Neurosci. Rep.* 19 (11), 85. doi:10.1007/s11910-019-1006-z

Liu, C.-C., Kanekiyo, T., Xu, H., Bu, G., and Bu, G. (2013). Apolipoprotein E and Alzheimer Disease: Risk, Mechanisms and Therapy. *Nat. Rev. Neurol.* 9 (2), 106–118. doi:10.1038/nrneurol.2012.263

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *J. Cogn. Neurosci.* 19 (9), 1498–1507. doi:10.1162/jocn.2007.19.9.1498

Mohanty, R., Mrtensson, G., Poulakis, K., Muehlboeck, J.-S., Rodriguez-Vieitez, E., Chiotis, K., et al. (2020). Comparison of Subtyping Methods for Neuroimaging Studies in Alzheimer's Disease: a Call for Harmonization. *Brain Commun.* 2 (2), fcaa192. doi:10.1093/braincomms/fcaa192

Mrtensson, G., Ferreira, D., Cavallin, L., Muehlboeck, J.-S., Wahlund, L.-O., Wang, C., et al. (2019). AVRA: Automatic Visual Ratings of Atrophy from MRI Images Using Recurrent Convolutional Neural Networks. *NeuroImage: Clin.* 23, 101872. doi:10.1016/j.nicl.2019.101872

Murray, M. E., Graff-Radford, N. R., Ross, O. A., Petersen, R. C., Duara, R., and Dickson, D. W. (2011). Neuropathologically Defined Subtypes of Alzheimer's Disease with Distinct Clinical Characteristics: a Retrospective Study. *Lancet Neurol.* 10, 785–796. doi:10.1016/S1474-4422(11)70156-9

Nettiksimmons, J., DeCarli, C., Landau, S., and Beckett, L.Alzheimer's Disease Neuroimaging Initiative (2014). Biological Heterogeneity in ADNI Amnestic Mild Cognitive Impairment. *Alzheimer's Demen.* 10, 511–521. doi:10.1016/j.jalz.2013.09.003

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., et al. (2014). Anatomical Heterogeneity of Alzheimer Disease: Based on Cortical Thickness on MRIs. *Neurology* 83, 1936–1944. doi:10.1212/WNL.0000000000001003

Oxtoby, N. P., Leyland, L., Aksman, L., Thomas, G., Bunting, E., Wijeratne, P., et al. (2021). Sequence of Clinical and Neurodegeneration Events in Parkinson's Disease Progression. *Brain* 144, 975–988. doi:10.1093/brain/awaa461

Oxtoby, N. P., and Alexander, D. C. (2017). Imaging Plus X: Multimodal Models of Neurodegenerative Disease. *Curr. Opin. Neurol.* 30 (4), 371–379. doi:10.1097/WCO.0000000000000460

Oxtoby, N. P., Garbarino, S., Firth, N. C., Warren, J. D., Schott, J. M., and Alexander, D. C.Alzheimer's Disease Neuroimaging Initiative (2017). Data-Driven Sequence of Changes to Anatomical Brain Connectivity in Sporadic Alzheimer's Disease. *Front. Neurol.* 8, 580. doi:10.3389/fneur.2017.00580

Pasquier, F., Leys, D., Weerts, J. G. E., Mounier-Vehier, F., Barkhof, F., and Scheltens, P. (1996). Inter-and Intraobserver Reproducibility of Cerebral Atrophy Assessment on Mri Scans with Hemispheric Infarcts. *Eur. Neurol.* 36 (5), 268–272. doi:10.1159/000117270

Perneczky, R., Wagenpfeil, S., Komossa, K., Grimmer, T., Diehl, J., and Kurz, A. (2006). Mapping Scores onto Stages: Mini-Mental State Examination and Clinical Dementia Rating. *Am. J. Geriatr. Psychiatry* 14 (2), 139–144. doi:10.1097/01.JGP.0000192478.82189.a8

Persson, K., Eldholm, R. S., Barca, M. L., Cavallin, L., Ferreira, D., Knapskog, A.-B., et al. (2017). MRI-assessed Atrophy Subtypes in Alzheimer's Disease and the Cognitive Reserve Hypothesis. *PLoS One* 12 (10), e0186595. doi:10.1371/journal.pone.0186595

Poulakis, K., Ferreira, D., Pereira, J. B., Smedby, Ö., Vemuri, P., and Westman, E. (2020). Fully Bayesian Longitudinal Unsupervised Learning for the Assessment and Visualization of AD Heterogeneity and Progression. *Aging* 12 (13), 12622–12647. doi:10.18632/aging.103623

Ramírez, J., Górriz, J. M., Ortiz, A., Martínez-Murcia, F. J., Segovia, F., Salas-Gonzalez, D., et al. (2018). Ensemble of Random Forests One vs. Rest Classifiers for MCI and AD Prediction Using ANOVA Cortical and Subcortical Feature Selection and Partial Least Squares. *J. Neurosci. Methods* 302, 47–57. doi:10.1016/j.jneumeth.2017.12.005

Redolfi, A., De Francesco, S., Palesi, F., Galluzzi, S., Muscio, C., Castellazzi, G., et al. (2020). Medical Informatics Platform (MIP): A Pilot Study across Clinical Italian Cohorts. *Front. Neurol.* 11, 1021. doi:10.3389/fneur.2020.01021

Salvatore, C., Cerasa, A., and Castiglioni, I.Alzheimer's Disease Neuroimaging Initiative (2018). MRI Characterizes the Progressive Course of AD and Predicts Conversion to Alzheimer's Dementia 24 Months before Probable Diagnosis. *Front. Aging Neurosci.* 10, 135. doi:10.3389/fnagi.2018.00135

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H. C., Vermersch, P., et al. (1992). Atrophy of Medial Temporal Lobes on MRI in "Probable" Alzheimer's Disease and Normal Ageing: Diagnostic Value and Neuropsychological Correlates. *J. Neurol. Neurosurg. Psychiatry* 55, 967–972. doi:10.1136/jnnp.55.10.967

Scheltens, P., Pasquier, F., Weerts, J. G. E., Barkhof, F., and Leys, D. (1997). Qualitative Assessment of Cerebral Atrophy on MRI: Inter- and Intra-Observer Reproducibility in Dementia and Normal Aging. *Eur. Neurol.* 37 (2), 95–99. doi:10.1159/000117417

Shaw, L. M., Vanderstichele, H., Knapik-Czajka, M., Clark, C. M., Aisen, P. S., Petersen, R. C., et al. (2009). Cerebrospinal Fluid Biomarker Signature in Alzheimer's Disease Neuroimaging Initiative Subjects. *Ann. Neurol.* 65, 403–413. doi:10.1002/ana.21610

Sperling, R. A., Aisen, P. S., Beckett, L. A., Bennett, D. A., Craft, S., Fagan, A. M., et al. (2011). Toward Defining the Preclinical Stages of Alzheimer's Disease: Recommendations from the National Institute on Aging-Alzheimer's Association Workgroups on Diagnostic Guidelines for Alzheimer's Disease. *Alzheimer's Demen.* 7 (3), 280–292. doi:10.1016/j.jalz.2011.03.003

ten Kate, M., Ingala, S., Schwarz, A. J., Fox, N. C., Chételat, G., van Berckel, B. N. M., et al. (2018). Secondary Prevention of Alzheimer's Dementia: Neuroimaging Contributions. *Alz Res. Ther.* 10, 112. doi:10.1186/s13195-018-0438-z

Ten Kate, M., Redolfi, A., Peira, E., Bos, I., Vos, S. J., Vandenberghe, R., et al. (2018). MRI Predictors of Amyloid Pathology: Results from the EMIF-AD Multimodal Biomarker Discovery Study. *Alz Res. Ther.* 10 (1), 100. doi:10.1186/s13195-018-0428-1

Tombaugh, T. N., and McIntyre, N. J. (1992). The Mini-Mental State Examination: A Comprehensive Review. *J. Am. Geriatr. Soc.* 40 (9), 922–935. doi:10.1111/j.1532-5415.1992.tb01992.x

Vemuri, P., and Jack, C. R. (2010). Role of Structural MRI in Alzheimer's Disease. *Alz Res. Ther.* 2 (4), 23. doi:10.1186/alzrt47

Venkatraghavan, V., Bron, E. E., Niessen, W. J., and Klein, S.Alzheimer's Disease Neuroimaging Initiative (2019). Disease Progression Timeline Estimation for Alzheimer's Disease Using Discriminative Event Based Modeling. *Neuroimage* 186, 518–532. doi:10.1016/j.neuroimage.2018.11.024

Whitwell, J. L., Dickson, D. W., MurrayWeigard, M. E. M. S., Weigand, S. D., Tosakulwong, N., Senjem, M. L., et al. (2012). Neuroimaging Correlates of Pathologically Defined Subtypes of Alzheimer's Disease: a Case-Control Study. *Lancet Neurol.* 11, 868–877. doi:10.1016/S1474-4422(12)70200-4

Wijeratne, P. A., Young, A. L., Oxtoby, N. P., Marinescu, R. V., Firth, N. C., Johnson, E. B., et al. (2018). An Image-Based Model of Brain Volume Biomarker Changes in Huntington's Disease. *Ann. Clin. Transl Neurol.* 5 (5), 570–582. doi:10.1002/acn3.558

Willette, A. A., Calhoun, V. D., Egan, J. M., and Kapogiannis, D.Alzheimer's Disease Neuroimaging Initiative (2014). Prognostic Classification of Mild Cognitive Impairment and Alzheimer′s Disease: MRI Independent Component Analysis. *Psychiatry Res. Neuroimaging* 224, 81–88. doi:10.1016/j.pscychresns.2014.08.005

Young, A. L., Marinescu, R. V., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., et al. (2018). Uncovering the Heterogeneity and Temporal Complexity of Neurodegenerative Diseases with Subtype and Stage Inference. *Nat. Commun.* 9, 4273. doi:10.1038/s41467-018-05892-0

Young, A. L., Oxtoby, N. P., Oxtoby, N. P., Huang, J., Marinescu, R. V., Daga, P., et al. (2015). Multiple Orderings of Events in Disease Progression. *Process. Med. Imaging* 24, 711–722. doi:10.1007/978-3-319-19992-4_56

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A Data-Driven Model of Biomarker Changes in Sporadic Alzheimer's Disease. *Brain* 137 (9), 2564–2577. doi:10.1093/brain/awu176

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 26 May 2021
doi: 10.3389/fdata.2021.613047

Developing an Explainable Machine Learning-Based Personalised Dementia Risk Prediction Model: A Transfer Learning Approach With Ensemble Learning Algorithms

Samuel O. Danso[1]*, *Zhanhang Zeng*[2], *Graciela Muniz-Terrera*[1] and *Craig W. Ritchie*[1]

[1] Edinburgh Dementia Prevention, Centre for Clinical Brain Sciences, University of Edinburgh Medical School, Edinburgh, United Kingdom, [2] School of Informatics, University of Edinburgh, Edinburgh, United Kingdom

Alzheimer's disease (AD) has its onset many decades before dementia develops, and work is ongoing to characterise individuals at risk of decline on the basis of early detection through biomarker and cognitive testing as well as the presence/absence of identified risk factors. Risk prediction models for AD based on various computational approaches, including machine learning, are being developed with promising results. However, these approaches have been criticised as they are unable to generalise due to over-reliance on one data source, poor internal and external validations, and lack of understanding of prediction models, thereby limiting the clinical utility of these prediction models. We propose a framework that employs a transfer-learning paradigm with ensemble learning algorithms to develop explainable personalised risk prediction models for dementia. Our prediction models, known as *source models,* are initially trained and tested using a publicly available dataset (n = 84,856, mean age = 69 years) with 14 years of follow-up samples to predict the individual risk of developing dementia. The decision boundaries of the best source model are further updated by using an alternative dataset from a different and much younger population (n = 473, mean age = 52 years) to obtain an additional prediction model known as the *target model*. We further apply the SHapely Additive exPlanation (SHAP) algorithm to visualise the risk factors responsible for the prediction at both population and individual levels. The best source model achieves a geometric accuracy of 87%, specificity of 99%, and sensitivity of 76%. In comparison to a baseline model, our target model achieves better performance across several performance metrics, within an increase in geometric accuracy of 16.9%, specificity of 2.7%, and sensitivity of 19.1%, an area under the receiver operating curve (AUROC) of 11% and a transfer learning efficacy rate of 20.6%. The strength of our approach is the large sample size used in training the source model, transferring and applying the "knowledge" to another dataset from a different and undiagnosed population for the early detection and prediction of dementia risk, and the ability to visualise the interaction of the risk factors that drive the prediction. This approach has direct clinical utility.

Keywords: early detection, risk factors, Alzheimer's, personalised dementia risk, explainable AI model, ensemble-based learning

OPEN ACCESS

Edited by:
Viktor Wottschel, Amsterdam University Medical Center, Netherlands

Reviewed by:
Lyduine Collij, Academic Medical Center, Netherlands
Alle Meije Wink, VU University Medical Center, Netherlands

***Correspondence:**
Samuel O. Danso
samuel.danso@ed.ac.uk

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *01 October 2020*
Accepted: *13 April 2021*
Published: *26 May 2021*

Citation:
Danso SO, Zeng Z, Muniz-Terrera G and Ritchie CW (2021) Developing an Explainable Machine Learning-Based Personalised Dementia Risk Prediction Model: A Transfer Learning Approach With Ensemble Learning Algorithms. Front. Big Data 4:613047. doi: 10.3389/fdata.2021.613047

INTRODUCTION

Dementia is the consequence of a number of progressive neurodegenerative diseases with Alzheimer's disease (AD) accounting for ∼60–80% of all types of dementias (Gaugler et al., 2019). AD is considered to be one of the top 10 causes of death, globally. Due to the progressive nature of the disease, people with dementia have different degrees of deterioration in cognition, memory, mental, and other functions (Lyketsos et al., 2002). Moreover, the socioeconomic burden of the disease is estimated to be in the region of one trillion USD per year (World Health Organization, 2017). Dementia has no cure; however, with early detection and diagnosis, it may be possible to delay the onset, which will help reduce the economic burden it currently poses on the society (Prince et al., 2018).

A recent Lancet report has identified modifiable risk factors, which when well-managed could reduce the risk of dementia or delay its onset (Livingston et al., 2020). However, the complexity of the interaction among these risk factors requires computational approaches capable of detecting patterns from these complex interactions to be able to achieve accurate prediction. Meanwhile, machine-learning based approaches have successfully been employed to help identify complex relationships between risk factors and their effect on disease outcomes in various application areas within the care pathway of patients. Examples of such application areas include prediction of pneumonia risk and 30-days readmission in hospital (Caruana et al., 2015), a real-time prediction of patients at the risk of septic shock (Henry et al., 2015), and application of machine learning model in breast screening (Houssami et al., 2017).

Following the above success storeys in the non-dementia domain, numerous attempts are being made to develop machine-learning models for dementia risk prediction. For example, Skolariki et al. (2021) applied machine learning algorithms to predict the likelihood of people with mild cognitive impairment converting to dementia based on features extracted from brain scans. Cui et al. (2019) also applied a recurrent neural network to develop a dementia risk prediction model based on longitudinal features extracted from brain scans. Other studies have also explored features obtained from sources, such as neuropsychological assessments (Barnes et al., 2009; Johnson et al., 2009; Lee et al., 2018; Adam et al., 2020). While these attempts have shown promising results, the prediction algorithms are mostly trained with samples containing diagnosis information and therefore unable to predict beyond the critical window of diagnosis (Prince et al., 2018), making these models ungeneralizable to relatively younger populations (Goerdten et al., 2019). Furthermore, despite these promising results achieved by machine learning-based approaches for dementia, their utility in healthcare settings remains limited partly due to the difficultly in interpreting the outputs of these models (Pellegrini et al., 2018). Interpretable models offer users the confidence and the ability to understand why a certain prediction was made for an individual and the specific underlining factors that led to the prediction. Confidence in how the prediction is made would allow the clinician to communicate this optimally to the patient and intervene. However, lack of confidence on the part of clinicians has resulted in the limited use of powerful machine learning approaches, such as deep learning and ensemble-based learning in developing prediction models for decision support systems in the dementia care pathway. Meanwhile, the complex nature of dementia, which results in complex data structures, makes it imperative to continue to explore these powerful machine learning methods, where traditional approaches, despite their limitations in handling complex data structures (Breiman, 2001), have widely been employed (Goerdten et al., 2019).

We develop and evaluate two ensemble-based interpretable models capable of learning patterns from the complex interactions among risk factors to be able to predict dementia risk at both population and individual levels up to an average of 14 years in advance. Unlike the approaches described above, our final model predicts individual dementia risk based on the parent history of dementia and genetic information about the individual. The prediction models are built using Random Forest (RF) and XGboost algorithms. Briefly, RF like other ensembles of classification and regression trees employs a "divide-and-conquer" strategy in the process of learning by repeatedly partitioning the input data into a number of large classification trees and fitting a prediction model for each tree (Breiman et al., 1984). It then employs the non-parametric bootstrap method (Efron and Tibshirani, 1994) to build a prediction model for each tree. Similarly, the XGBoost also belongs to the family of classification and regression trees and adopts the RF approach to learning. However, XGBoost employs a step-wise, additive approach to sequentially build a prediction model for each tree, while taking into account the difficulties encountered in fitting previous models (Natekin and Knoll, 2013). It is worth noting that RF and XGboost both combine the predictions from weak learners to produce a final model—a process known as "voting." These algorithms have been demonstrated to be powerful when applied to various problems, such as risk prediction of hypoxaemia during general anaesthesia and surgery (Lundberg et al., 2018).

We argue that our proposed approach provides useful and actionable information to assist clinicians and other users in their decision-making process around diagnosis, prognosis, and management. We also believe that this is an important step for machine learning in neurodegenerative disease research and translation to clinical care. Our approach not only significantly improves the ability for the early detection of neurodegenerative disease but also the ability to explain the predictions from accurate and complex models in order to understand drivers of the prediction for important intervention strategies to be developed.

METHODS

Overview of the Research Framework

It is believed that dementia clinically manifests after decades of exposure to risk factors (Ritchie and Ritchie, 2012). Therefore, the aim of this project was to develop a machine learning model capable of predicting the risk of developing dementia decades prior to the onset of the dementia syndrome. To achieve this, the task was formulated as a transfer learning classification

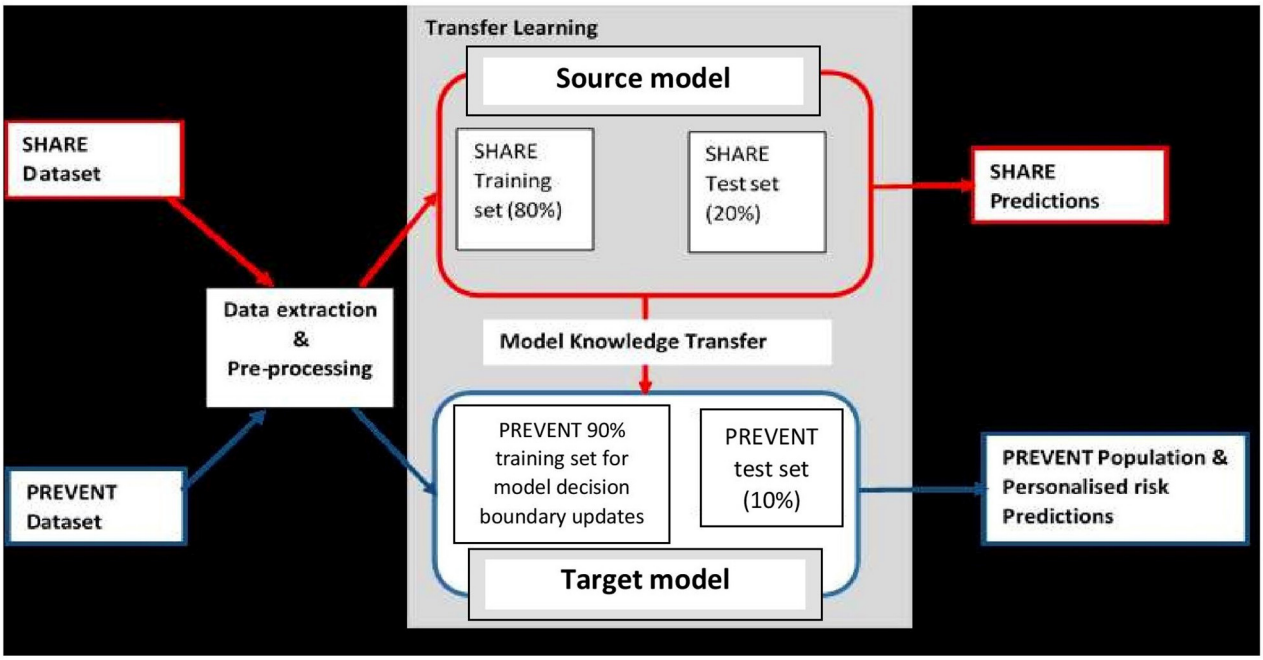


FIGURE 1 | Transfer learning process showing how data extraction and pre-processing procedures are applied to SHARE and PREVENT datasets. A prediction model (Source model) is built using the SHARE dataset with 80% of the data used for training and 20% held-out for testing for SHARE predictions. The Source model is updated with 90% PREVENT training and the updated prediction model (Target model) is applied to PREVENT 10% test set held-out for population as well as personalised risk prediction of dementia.

problem (Pan and Yang, 2009). This made it possible to develop the machine learning prediction model using the data drawn from different populations and applied the model to another population. **Figure 1** illustrates the methodology employed. As the figure shows, unlike traditional machine learning where a model is developed and applied to predict data from the same population, our model was developed using external data source and transferred the knowledge learned from the external population and applied it to data from population of different characteristics. The characteristics of the data sources are discussed in the next section.

Data Description and Preprocessing

The data sources used in developing the models were obtained from the Survey of Health, Ageing, and Retirement in Europe (SHARE) study (Börsch-Supan et al., 2013) and the PREVENT Dementia programme (Ritchie and Ritchie, 2012). While both SHARE and PREVENT projects are related to dementia research, the rationale and aims of each of the studies vary resulting in differences in the datasets. **Table 1** shows a brief description of the datasets. While SHARE population covers 20 European countries with the mean age of 69 years, the PREVENT data, on the other hand, is a relatively younger cohort with the mean age of 52 years drawn from a population limited to the United Kingdom. Further, the SHARE cohort includes individuals with some having been diagnosed with dementia, while the PREVENT cohort contains healthy individuals without a diagnosis of dementia. However, the PREVENT study participants are children of individuals with or without a diagnosed dementia. The study also collects information about the apolipoprotein E (ApoE) genotype of each individual.

Even though both SHARE and PREVENT research programmes have different research aims and objectives, there was a high degree of overlap between the two datasets in terms of data collection. In order to make transfer learning possible, it was important to focus on common data items between the two datasets. **Table 2** shows the categories of common variables found in both datasets. We extracted data records from the SHARE dataset and merged the data of individuals across waves 1–6 which covers the period between 2004 and 2015. Therefore, from the SHARE cohort, it was possible to build a prediction model using a longitudinal dataset of 14 years of follow-up data. The PREVENT dataset on the other hand is the baseline data collected between February 2014 and October 2018.

The difference in data collection protocols used by the studies resulted in structural differences in data. To address these differences, we devised a pre-processing procedure to harmonise the representation of the data items, which were employed as features to train the learning algorithms. All medical history variables were processed to have binary feature representation based on the responses as either condition being present or not present, with a feature value of "1" and "0," respectively. The Body Mass Index (BMI) as per WHO classification was applied to obtain the following four categories: underweight (<18.5 kg/m^2), normal (18.5–24.9 kg/m^2), overweight (25–29.9 kg/m^2), and obese (>30 kg/m^2) with feature values of

"0," "1," "2," and "3," respectively Furthermore, "marital status" had categorical entries ("divorced," "married," "living with spouses," "married," "not living with spouse," "never married," and "registered partnership"), and each of these was separately represented as binary based on the response as either "yes" or "no," with a feature value of "1" and "0," respectively. The International Standard Classification of Education scheme was applied to "education level" variable to have seven categories with feature value representations (0 = none; 1 = first stage of basic education; 2 = lower secondary education or second stage of basic education; 3 = upper secondary education; 4 = post-secondary non-tertiary education; 5 = first stage of tertiary education; and 6 = second stage of tertiary education). The "daily activity" variables had two categories: "vigorous" and "moderate" sports with each having feature value representations (0 = hardly ever or never; 1 = one to three times a month; 2 = once a week; and 3 = more than once a week). We believe that this method of representation provides information on the activity as well as the intensity of the activity, which can be useful for the learning algorithms. The "smoking" variable was also processed to have a binary representation based on the responses with feature values (0 = never smoked and 1 = current or past smoker). Finally, the SHARE dataset contained data on whether a participant had been diagnosed with Alzheimer's disease (AD) and those without a diagnosis. This was therefore used as the class variable for the prediction model feature values representation (Non-AD = no diagnosis; AD = diagnosis of Alzheimer's dementia). However, in the absence of a diagnosis in the PREVENT dataset, and to facilitate the evaluation of our approach, we employed a classification scheme proposed by Ritchie and Ritchie (2012) to group the participants according to parental clinical status and ApoE genotype. Therefore, participants with a parental dementia diagnosis and ApoE 4 genotype were allocated to a "High-Risk" (HR) group as these individuals were considered to be at high risk of dementia. All other participants were allocated to a "Low-Risk" (LR) group. The final distribution of classes is as follows: SHARE dataset, Non-AD (95%) and AD (5%); PREVENT dataset HR (23%) and LR (77%).

TABLE 1 | Characteristics of SHARE and PREVENT datasets.

Data description	SHARE data	PREVENT data
Population	20 European countries	The United Kingdom
Number of samples	84,856	473
Mean age	69	52
Number of years of follow-ups	14 years (2004–2015), 2 years interval on average	Only used baseline data
Class distribution	Diagnosis • Diagnosis of Alzheimer's disease—"AD" (n = 4,157) • No diagnosis of-Alzheimer's disease diagnosis—"non-AD" (n = 80,699)	Parental diagnosis of AD and Apolipoprotein E4 allele (ApoE4) genotype status of individual • Parental diagnosis of AD + ApoE4 status—"High Risk" (n = 109) • No parental diagnosis of AD + No ApoE4 status of individual—"Low Risk" (n = 364)

Building the Prediction Model

We built four ensemble-based prediction models by training RF and XGBoost algorithms. The algorithms were trained by applying a hybrid approach that combines cross-validation and hold out, through a procedure we refer to as *cross-validation with hold out* (Pedregosa et al., 2011). This procedure involved splitting the SHARE data into training and test sets. The training set (D_train), which constituted 80% of the SHARE data, was used to train the algorithms including hyperparameters tuning. The 20% test set (D_eval) was held and used only for the model performance evaluation. Similarly, the PREVENT data was also split into 80% training set (PREV_train) and 20% test set (PREV_eval). The splits were stratified in order to ensure the equal proportion of class representation in both training and test sets. A summary of our cross-validation with hold out training of algorithms procedure is as follows:

- Step 1: We employed a 5-fold cross-validation during training, which randomly split the 80% training set into 5-folds each containing a subset of training (D_train_{1-5}) and validation (D_val_{1-5}) sets.
- Step 2: We applied a set of initial hyperparameters to train the algorithm to obtain five different models using D_train_{1-5} and D_val_{1-5}, to obtain a number of potential hyperparameters from each cross-validation.
- Step 3: We then applied the random search optimization algorithm (Bergstra and Bengio, 2012), to search and choose from a set of potential number of hyperparameters derived

TABLE 2 | The common data items between SHARE and PREVENT datasets used to develop the prediction models.

Data category	Data items
Sociodemographic	• Gender • Age • Education level • Marital status • Had children? • BMI
Self-reported medical history	• Heart attack • Hypertension (high blood pressure) • High cholesterol • Diabetes • Lung disease • Peptic ulcer disease • Parkinson's disease • Emotional disorders • Osteoarthritis
Life style	• Daily activity • Smoking

from Step 2 to obtain the optimal set of hyperparameters based on the evaluation function of the optimization algorithm. **Table 3** shows the set of initial and optimal hyperparameter settings obtained.

- Step 4: Once the optimal hyperparameters are obtained, we then retrained the algorithm using the optimum hyperparameters on the entire training set, D_train.
- Step 5: We applied the procedures in Steps 2–4 for RF and XGBoost to obtain SHARE_RF_pred and SHARE_XGBoost_pred prediction models, respectively.
- Step 6: We evaluated the performance of the prediction models obtained in Step 5 by applying SHARE_XGBoost_pred and SHARE_RF_pred to the hold-out test set (D_eval).
- Step 7: We employed the method proposed by DeLong et al. (1988) to carry out a pairwise comparison of the receiver operating curve (ROC) to compare the performance difference between SHARE_XGBoost_pred and SHARE_RF_pred to determine the best model.
- Step 8: We randomly spit the PREVENT data into 80% training set (PREV_train) and 20% held out test set (PREV_eval). Again, the split was stratified in order to ensure an equal proportion of class representation in both the training and test sets.
- Step 9: We employed a parameter-transfer learning approach as described by Yao and Doretto (2010) to build a target model. This approach assumes that the target shares parameters with the best source model as determined in Step 7. The parameters of the best source model are further updated using the PREV train set. This process adjusted the decision boundaries of the source model to produce PREVENT_target prediction model.
- Step 10: We evaluated the performance of prediction models obtained in Step 9 by applying them to the hold-out test set (PREV_eval).
- Step 11: We trained the XGBoost algorithm using PREV_train and applied procedures into Steps 2–4 to obtain a prediction model (PREVENT_only).
- Step 12: We evaluated the performances of PREVENT_target and PREVENT_only by applying them to the hold-out test set (PREV_eval).
- Step 13: We finally applied the procedures in Step 7 to compare the performance difference between the PREVENT_target and PREVENT_only to determine the best model.

TABLE 3 | Hyperparameter settings for prediction models.

Algorithm	Initial parameters	Optimal hyperparameter settings
Random Forest	n_estimators = range (5, 40), max_features = ['auto', 'sqrt', 'log2'], max_depth = range (10, 25), criterion = [gini, entropy]	Bootstrap = True; ccp_alpha = 0.0; class_weight = None; criterion = entropy; max_depth = 24; max_features = sqrt; max_leaf_nodes = None; max_samples = None; min_impurity_decrease = 0.0; min_impurity_split = None; min_samples_leaf = 1; min_samples_split = 2; min_weight_fraction_leaf = 0.0; n_estimators = 33, n_jobs = None; oob_score = False; random_state = None; verbose = 0; warm_start = False
XGBoost	n_estimators = range (1, 20), max_depth = range (10, 25), learning_rate = [.1,.2,.4,.45,.5,.55,.6], colsample_bytree': [.6,.7,.8,.9, 1], booster = gbtree, min_child_weight = [0.001, 0.003, 0.01]	Objective = multi:softprob; base_score = 0.5; booster = gbtree; colsample_bylevel = 1; colsample_bynode = 1; colsample_bytree = 0.7; gamma = 0; gpu_id = −1; importance_type = gain; interaction_constraints = None; learning_rate = 0.5, max_delta_step = 0; max_depth = 24; min_child_weight = 0.003; missing = nan; monotone_constraints = None; n_estimators = 16; n_jobs = 0; num_parallel_tree = 1; random_state = 0; reg_alpha = 0; reg_lambda = 1; scale_pos_weight = None; subsample = 1; tree_method = None; validate_parameters = False; verbosity = None; num_class = 2

Performance Evaluation

We employed a series of metrics to evaluate the performance of the models based on the D_eval and PREV_eval unseen datasets. As already pointed out, D_eval contained "AD" and "No-AD" which served as the ground truth for the evaluation of SHARE_RF_pred and SHARE_XGBoost_pred models. PREV_eval on the other contained "HR" and "LR" as explained above, and this served as the ground truth for the evaluation of our PREVENT_target and PREVENT_only models. These metrics were primarily based on the following information obtained from the outputs of the prediction models: Refer False Positive (FP), False Negative (FN), True Positive (TP), and True Negative (TN) (Pollack, 1970) for details of these metrics. The comparison of the models was based on geometric accuracy (GA) as expressed in Equation (3) which is derived from Equations (1) and (2) which represent sensitivity and specificity, respectively. GA accounts for both majority and minority class error rates which makes it ideal for imbalanced problems (Kim et al., 2015).

$$Sensitivity = \frac{Number\ of\ TP}{Number\ of\ TP + Number\ of\ FN} \tag{1}$$

$$Specificity = \frac{Number\ of\ TN}{Number\ of\ TN + Number\ of\ FP} \tag{2}$$

$$Geometric\ Accuracy = \sqrt{(Sensitivity * Specificity)} \tag{3}$$

We also employed area under the receiver operating curve (AUROC) to further explore the robustness of our models, given the wide usage of this metric in medical applications (Mandrekar, 2010). Also, as already stated, a significant test was used to examine the performance differences between the prediction models.

Finally, we employed a method proposed by Taylor and Stone (2009) to examine the efficacy of our transfer

learning approach based on a learning ratio as expressed in Equation (4).

$$ratio = \frac{area\ under\ curve\ with\ transfer - area\ under\ curve\ without\ transfer}{area\ under\ curve\ with\ transfer} \quad (4)$$

Feature Importance and Model Interpretability

An important advantage of tree-based algorithms is their ability to provide information on the decisions made around predictions. This information is provided in the form of weights that are assigned to the features as a result of the learning process. The value of weight assigned to a given feature is an indicator of the importance of that feature as determined by the prediction model, which enabled us to examine how each feature was ranked by the prediction models.

We further applied the SHapley Additive exPlanation (SHAP) algorithm to explore the interactions between the features (Lundberg et al., 2018). Briefly, the algorithm is inspired by game theory, where the interaction between features is considered as a "team" of features, with each feature being a member of the team responsible for driving the overall risk. An instance of the interaction between the features registers a set of predicted values produced by the prediction model. These values serve as input for the SHAP algorithm to generate another set of values known as "impact values." The SHAP values provide a dynamic view of the effects of the interaction between the features to determine the probability of risk and the role of each feature on the individual level. Furthermore, the SHAP algorithm offers the possibility to compare an individual predicted risk probability with a baseline prediction, which is the average predicted probability known as the "base value."

RESULTS

Model Performance Analyses

Figure 2 shows the confusion matrix of the results obtained when SHARE_RF_pred (**Figure 2A**) and SHARE_XGBoost_pred (**Figure 2B**) models were applied to 20% of SHARE unseen test set. The figure also shows the results when PREVENT_target (**Figure 2C**) and PREVENT_only (**Figure 2D**) models were applied to 20% of PREVENT unseen test set. **Table 4** further shows a summary of the performances obtained. As seen from the table, SHARE_XGBoost achieves a GA of 87%, specificity of 99%, sensitivity of 76%, and AUROC of 96%. In comparison, SHARE_RF_pred achieves a GA of 85%, specificity of 99%, sensitivity of 73%, and AUROC of 94%. **Figure 3A** shows an AUROC curve comparison between SHARE_RF_pred and SHARE_XGBoost, with SHARE_XGBoost showing a marginal difference in the performance between the two models. A pairwise comparison of the AUROC scores between the two prediction models demonstrates a significant difference in performance ($P < 0.0001$, 95% Confidence Interval: 0.01–0.02), suggesting SHARE_XGBoost as the best performing model.

Again, as seen from **Table 4**, PREVENT_target achieves a GA of 56.5%, specificity of 84.7%, sensitivity of 38.1%, and AUROC of 63%. In comparison, PREVENT_only achieves a GA of 39.6%, specificity of 82.0%, sensitivity of 19%, and AUROC of 51%. **Figure 3B** shows an AUROC curve comparison between PREVENT_target and PREVENT_only, with PREVENT_target showing a marginal difference in performance between the two models. Even though a pairwise comparison of the AUROC scores between PREVENT_target and PREVENT_only, no significant difference in performance is observed ($P = 0.2166$, 95% Confidence Interval: 0.07–0.325), the PREVENT_target model outperformed PREVENT_only model across all the performance metrics as shown in **Table 4**. There is an increase in the sensitivity of 19.1%, specificity of 2.7%, GA of 16.9%, AUROC of 11%, and a transfer-learning rate of 20.6%.

Feature Importance Analysis and Interpretability of Personalised Risk Prediction

Even though RF and XGboost are both considered ensemble-based algorithms, the learning strategy tends to differ as briefly discussed. From that score, we examine how both models assessed the importance of the features used in training the models. **Figures 4A,B** depict a comparison between SHARE_RF_pred and SHARE_XGBoost_pred prediction models on how features were ranked based on the weights assigned. As shown by **Figures 4A,B**, while significant similarities in the ranking of the features exist between the two models, some striking differences can also be observed. For example, the ranking of the top seven features of both RF and XGBoost appear to be in the same order, with '"age" being the most important feature followed by "moderate sport," "education," "vigorous sports," "BMI," "hypertension," and "esmoked." Some differences in rankings were observed. Where RF ranks "gender" and "emotional disorders" as the 8th and 9th most important features, XGBoost ranks "high cholesterol" and "osteoarthritis," respectively. Additionally, RF ranks "widowed" as the 10th most important feature, whereas XGBoost ranks "diabetes" as the 10th most important feature, and ranks "widowed" as one of the least important features (ranked 18th).

Similarly, a comparison between PREVENT_only and PREVENT_target shows how these prediction models ranked the features as shown in **Figures 4C,D**, respectively. Again, while there appear to be some overlaps in the order of feature rankings between the models, some differences can also be observed. For example, "age" remains the most important feature among the two models. A close examination of the top 10 features of the models show some differences in the order of rankings. For example, while PREVENT_only ranks "divorced," and "no_children" among the top 10, PREVENT_target also ranks "BMI" and "gender" among the top 10, but ranks "divorced," and "no_children" in the 11th and 13th positions, respectively. Even though these differences in feature rankings can be observed between these two models, the difference is not statistically significant. However, because our PREVENT_target demonstrated some marginal increase in the performance over PREVENT_only, our analysis will be based on the output of PREVENT_target model. A further comparison of the order

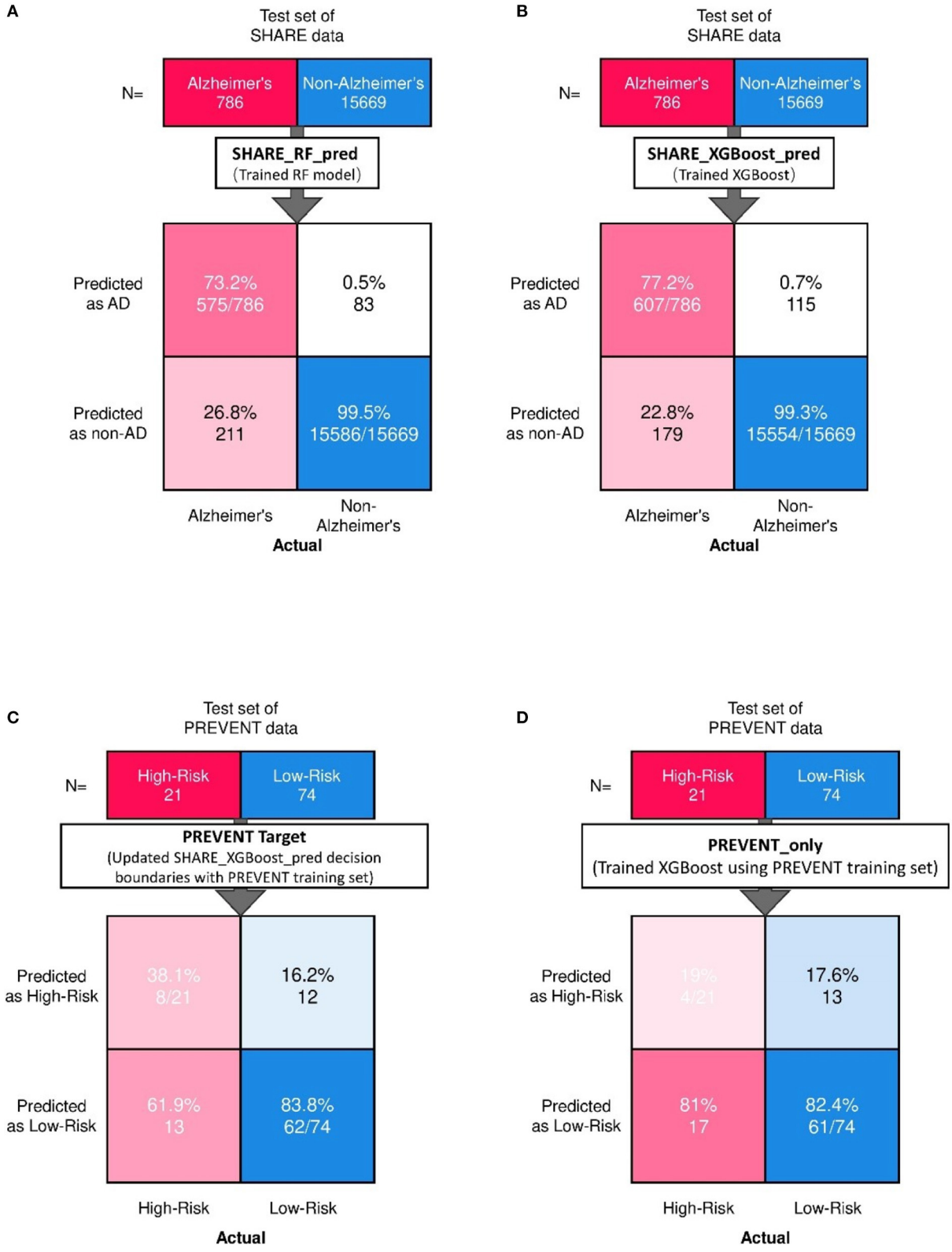


FIGURE 2 | Confusion matrix showing the prediction results from unseen 20% of SHARE test data as predicted by **(A)** Random Forest **(A,B)** XGBoost models. Also showing are the prediction results from 20% unseen PREVENT test data as predicted by **(C)** Updated SHARE_XGBoost_pred decision boundaries with PREVENT training set and **(D)** Trained XGBoost using PREVENT training set.

TABLE 4 | Summary of prediction models on the unseen test set.

Model	Sensitivity (%)	Specificity (%)	Geometric Accuracy (%)	AUROC (%)	*P*-value	Transfer learning efficacy (%)
SHARE_RF_pred	73	99	85	94	$P < 0.0001$	N/A
SHARE_XGBoost_pred	*76 (+3%)	*99 (0%)	*87 (2%)	*96 (2%)		
PREVENT_target	**38.1 (+19.1%)	**84.7 (+2.7%)	**56.5 (+16.9%)	**63 (+11%)	$P = 0.2166$	20.6%
PREVENT_only	19.0	82.0	39.6	51		

**Performance comparison in relation to SHARE_RF_pred.*
***Performance comparison in relation to PREVENT_only.*

of rankings of features between SHARE_XGBoost_pred as the source model and our PREVENT_target as the target model also shows 70% overlap among the top 10 features as ranked by both the models. The differences observed include: "emotional_disorders," "hypertension," and "diabetes" ranked among the top 10 by SHARE_XGBoost_pred, but ranked by PREVENT_target model at 12th, 14th, and 21st positions, respectively.

Furthermore, we examined the performance of the models at individual levels. **Figure 5** shows the visualisation of SHAP values of four randomly selected prediction outputs when SHARE_XGBoost_pred was applied to SHARE unseen test set. **Figure 5A** shows an individual with AD and correctly predicted by the model, with the probability of 80%. **Figure 5B** shows an individual with AD which is incorrectly predicted as a non-AD with the probability of 6%. **Figure 5C** shows an individual without AD predicted as AD with the probability of 66%. **Figure 5D** also shows an individual without AD and correctly predicted as a Non-AD with the probability of 4%. The figures also show the risk factors that drive each of the probabilities, with red indicating risk factors and blue suggesting protective factors. For example, **Figure 5A** shows a 69-year-old woman correctly predicted to be living with AD with the probability of 80%. While smoking, vigorous sports, education, BMI, and osteoarthritis appear to be playing a role in the prediction, the lack of moderate sports appears to be the most important risk factors as determined by the colour (red) and the length of the bar allocated to each risk factor. In contrast, as **Figure 5B** shows, age and the fact that the person engages in moderate sports appear to have significant impact on the prediction, which resulted in a relatively low risk of probability of 6%. Similarly, age and moderate sports appear to have a significant impact on the prediction of probabilities in both **Figures 5C,D**. However, while moderate sports appear to be protective for the individual as shown in **Figure 5C**, the relatively older age (80 years) and the lack of education appear to be the risk factors that have a significant impact on the prediction resulting in the probability of 66% of AD. In contrast, the individual shown in **Figure 5D** is relatively young and engages in moderate as well as vigorous sports, which appear to be the proactive factors driving the prediction with a relatively low probability of 4% risk of AD.

Examining our target model at the individual level, **Figure 6** shows randomly selected outputs when PREVENT_target model was applied to PREVENT unseen test set. **Figure 6A** shows a low-risk individual predicted as a high-risk with the probability of 70%. **Figure 6B** shows a high-risk individual correctly predicted with the probability of 7%. **Figure 6C** shows a high-risk individual predicted as low-risk with the probability of 19%. **Figure 6D** is also a low-risk individual correctly predicted as low-risk with the probability of 27%. As the figures show, while age appears to be the most protective factor for all the individuals, the lack of vigorous sports, relatively low education, and BMI appear to be the risk factors with the greatest impact. A closer look at **Figure 6A** shows a 60-year-old individual who has no education and lacks physical activity and therefore predicted by the model to be at high risk despite having been allocated to the low-risk group. Similarly, **Figure 6B** shows a 52-year-old individual belonging to the high-risk group and correctly predicted by the model with a probability of 63%. In this figure, individual age is the most protective factor, while education ($3 =$ upper secondary level) and having a healthy weight ($BMI = 1$) appear to be risk factors. This may suggest that higher education may be critical for individuals with an APOE e4 gene and a parental history of dementia, compared to individuals without that fall outside the high-risk group.

DISCUSSION

This study developed an ensemble-based machine-learning model to predict Alzheimer's dementia risk at both population and individual levels based on the data drawn from two populations with different characteristics. Our models were built using large heterogeneous data drawn from a population of 20 European countries with up to 14 years of follow-up data. Our best model achieves high-performance accuracy, obtaining an AUROC score of 96% on the unseen test set. The decision boundaries of the best model were further updated through transfer learning. The update was done using data from a different population with different dementia risk profiles to produce a target model. The target model achieves an AUROC score of 63% and a transfer learning efficacy rate of 20%. It is also able to visualise the risk as well as protective factors that are responsible for the prediction at both population and individual levels.

To the best of our knowledge, this is the first approach that employs transfer learning with ensembles to develop dementia risk prediction models and visualisation of risk factors from an undiagnosed population in mid-life. Although numerous computational approaches have been developed, these methods

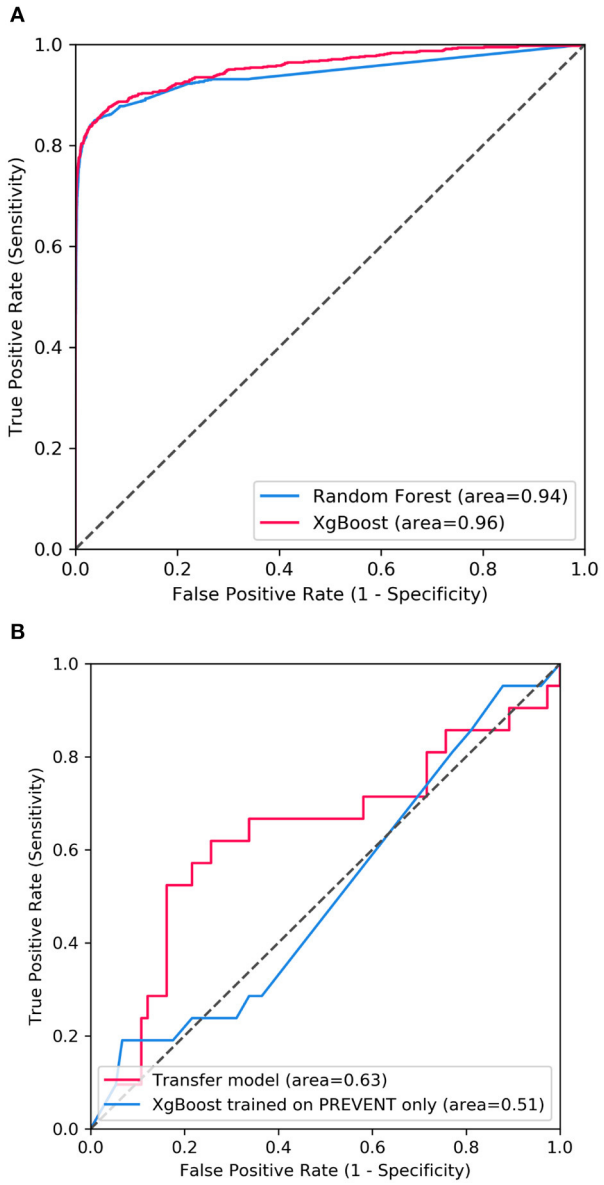


FIGURE 3 | Showing ROC curves with AUROC scores of **(A)** the performance difference between Random Forest and XGBoost prediction models when applied to 20% SHARE unseen test set and **(B)** the performance compassion between XGBoost model updated with PREVENT training set (Transfer model) and XGBoost trained with PREVENT only (PREVENT only model) and applied to 20% PREVENT unseen test set.

have been limited in terms of sample size and the over-reliance on a homogenous sample for validation (Goerdten et al., 2019). van Maurik et al. (2019) attempted to address this issue by combining data from older adults in different populations across Europe and North America to develop dementia-risk prediction models for people with mild cognitive impairment. They employed traditional statistical modelling approaches and biomarkers, such as cerebrospinal fluid and imaging data to develop the prediction models. While we are unable to compare our proposed approach to that of van Maurik et al. (2019) due to differences in data used, it would be interesting to compare the performance of the two modelling approaches on the same dataset in the future.

Even though the relative differences in feature rankings between the models may be hard to interpret relative to their importance in predicting the dementia risk, and given that XGBoost outperforms RF as our significant test suggests, it would be reasonable to conclude that the feature rankings of XGBoost model could be more accurate and therefore reliable. The prediction models developed here identified risk factors that agree with previous literature. We demonstrate this by examining the top 10 features as ranked by the XGboost prediction models. Numerous studies have concluded that age remains the single biggest risk factor (Song et al., 2014). This is consistent with our model, ranking age to be the most important risk factor. Even though age is considered a non-modifiable risk factor, the Lancet commission report on dementia prevention by Livingston et al. (2020) identified a number of risk factors which when modified could reduce the risk of dementia by 40%. The report identified less education, hypertension, hearing impairment, smoking, obesity, depression, physical inactivity, diabetes, and infrequent social contact as potentially modifiable risk factors. Seventy percent of these risk factors were ranked among the top 10 by the study's prediction model as shown in **Figure 4**.

Furthermore, the interaction effects identified by the study's models are also in accordance with the existing evidence. For example, low education level is known to account for up to 8% and physical inactivity accounts for up to 3% of the dementia risk (Livingston et al., 2017). Again, both education and physical activity are associated with cognitive reserves and improvement in mental functions, suggesting that these could act as protective factors (Sharp and Gatz, 2011). Therefore, poorly educated individuals with a sedentary lifestyle could have an increased risk of dementia. This phenomenon is consistent with what is observed in **Figures 5**, **6**. As **Figure 5A** demonstrates, the relatively low education and low levels of physical activity (moderate/vigorous sports) were the two major risk factors among the (non-age) other risk factors that increased the risk of dementia up 80% of this individual. This is consistent with what is observed in **Figure 6A** which shows an individual considered to be at low risk but due to lack of education and physical activity, the risk profile of this individual is predicted with 70% probability, with age being the only protective factor.

While the majority of the top 10 risk factors ranked by the study's prediction model were part of those identified by the recent Lancet Commission report, there are a few that appear to be playing a major role in the risk prediction but not currently part of the report. **Figure 6B** demonstrates the effect of emotional disorder on the risk of dementia at the individual level. Again, while age and physical activity remain significant protective factors, emotional disorder appears to be playing a significant role in the 7% risk of Alzheimer's Dementia for this individual. Therefore, any intervention in the emotional health of this participant chosen for illustrative purposes could further reduce their risk. This approach is exactly what is envisaged in the Brain Health Clinics being developed across Europe (Frisoni et al., 2020) based on a consensus led by our group in how to change clinical services for dementia prevention (Ritchie et al.,

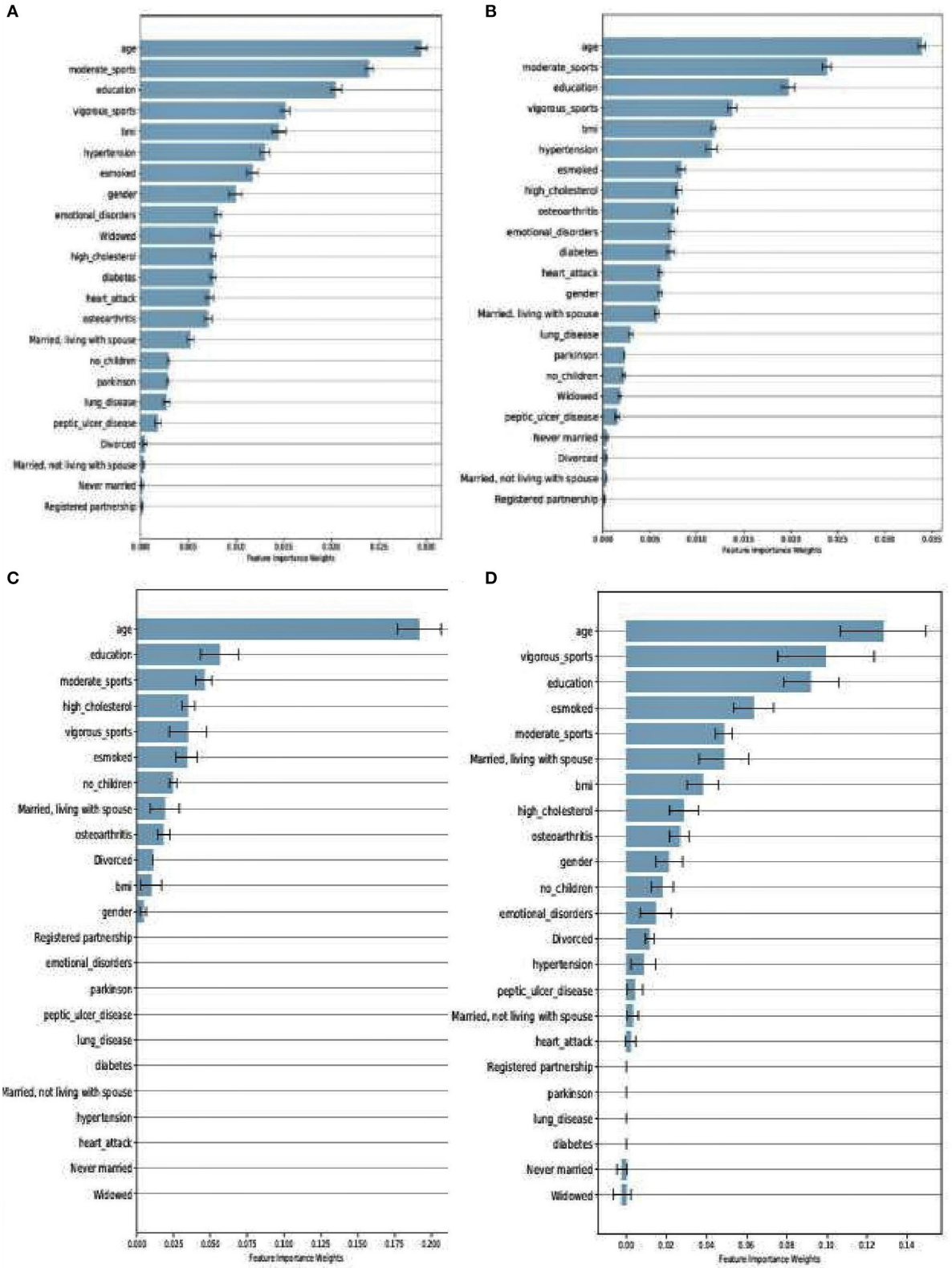


FIGURE 4 | Feature importance as ranked by the weights derived from SHARE_RF_pred **(A)** and SHARE_XGBoost_pred **(B)** prediction models that were trained using SHARE dataset. It also shows the ranking of features of PREVENT only **(C)** and PREVENT target **(D)**.

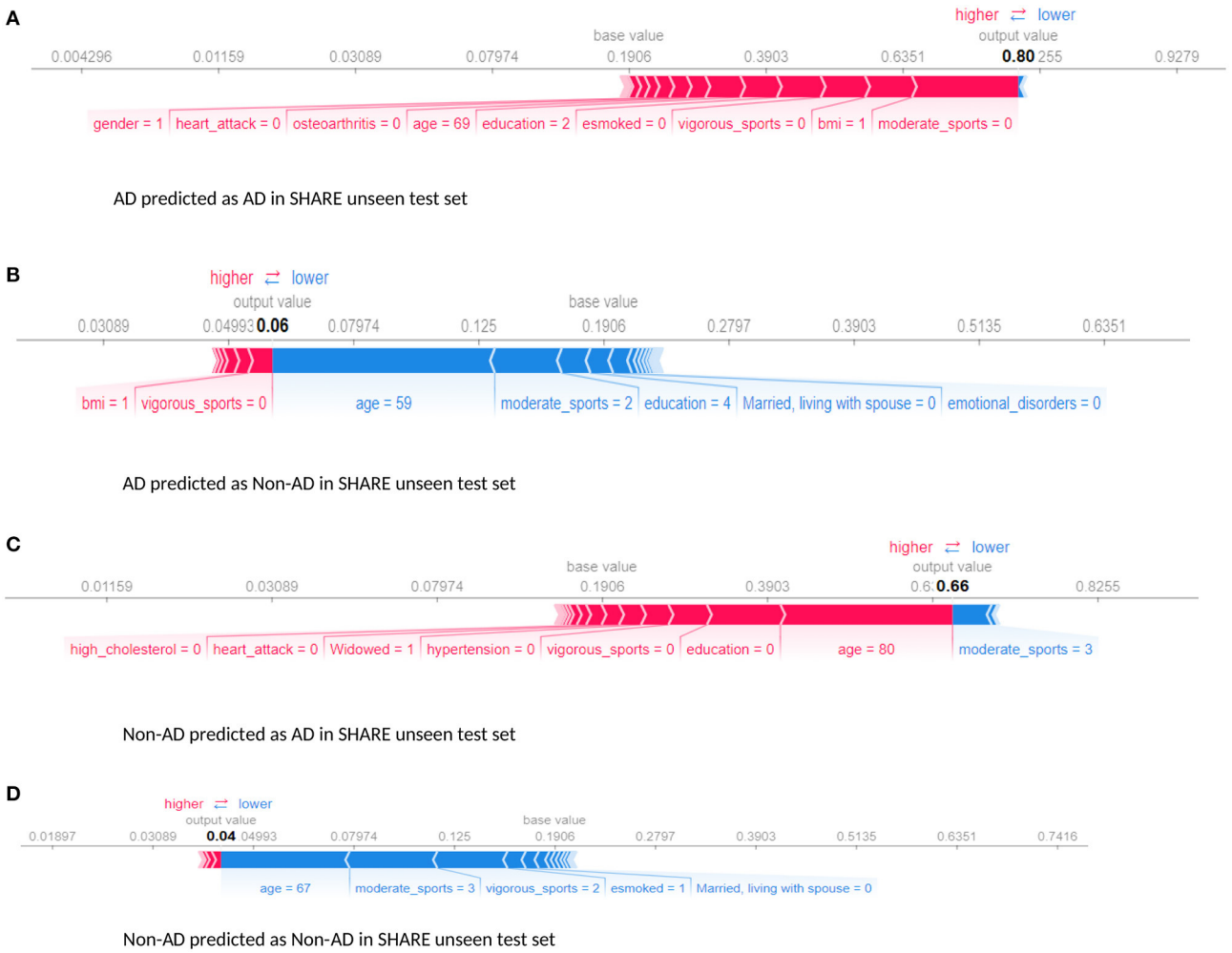


FIGURE 5 | Force plot showing the effect of SHAP values on the interaction of features and the overall prediction at the individual level. This shows examples of prediction outputs with taken from SHARE as predicted by the SHARE model. Features in red show risk factors pushing up the overall probability while those in blue are protective factors pushing down the probability. **(A)** Shows SHARE participants predicted to have AD with 85% probability. **(B)** Shows SHARE participant diagnosed to have AD but has been predicted by the model to be Non-AD with 6% probability. **(C)** Shows a Non-AD participant predicted as AD with 63% probability. **(D)** Shows as Non-AD participant predicted as Non-AD with 4% probability. Feature labels are: **esmoked** (0 = never smoked); **emotional disorders** (0 = no); **hypertension** (0 = no); **osteroarthritis** (1 = yes); **high cholesterol** (0 = no); **heart attack** (0 = no); **education** (2 = lower secondary education or second stage of basic education; 3 = upper secondary education); **moderate sports** (0 = hardly ever, or never, 1 = one to three times a month); **vigorous sports** (0 = hardly ever, or never, 1 = one to three times a month); **no children** (0 = no children); **widowed** (1 = yes); **BMI** [1 = under weight (<18.5)]; **married, living with spouse** (0 = no) and **gender** (1 = male).

2017). This is based on collecting data from these Brain Health Clinics to support Real World machine learning approaches and using these algorithms to support the development of personalised prevention plans driven by early disease detection and comprehensive risk profiling.

Even though the performance of the study's prediction model demonstrates a potential clinical utility, we do acknowledge that it would benefit from further development and validation. Firstly, it would be beneficial to evaluate the effect of additional data sources derived from biological samples and neuroimaging on the overall performance of the study's model as well as the effect of the interactions of additional features at both population and individual levels. Secondly, further validation of the model using data from non-research settings is crucial. The dataset used in training the model is obtained from research settings, which is considered to be of high quality due to the strict data collection protocols that are used in these settings. Thirdly, the problem of imbalanced data and the ability to develop accurate prediction models that account for these problems are major challenges (Khalilia et al., 2011). However, RF and XGBoost have consistently been shown to have the capacity to handle imbalanced challenges due to the strategy employed in learning. For example, Facal et al. (2019) compared the performance of number learning algorithms,

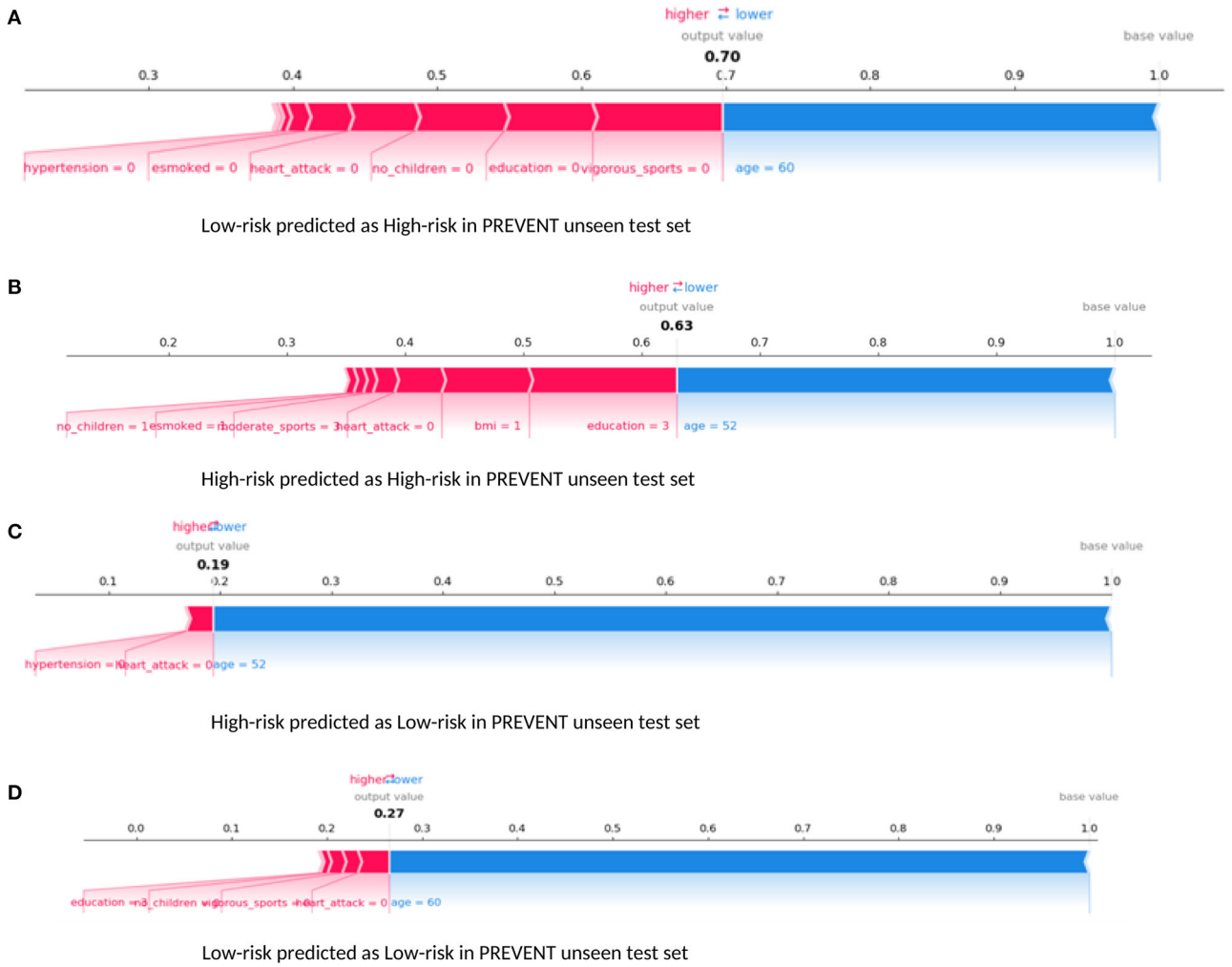


FIGURE 6 | Force plot showing the effect of SHAP values on the interaction of features and the overall prediction at the individual level. This shows examples of prediction outputs PREVENT target models. Features in red show risk factors pushing up the overall probability while those in blue are protective factors pushing down the probability. **(A)** Shows PREVENT participant assigned to the Low-risk group but has been predicted by the model to be High-risk with 70% probability. **(B)** Shows a High-risk participant predicted as High-risk with 63% probability. **(C)** Shows a High-risk participant predicted as Low-risk with 19% probability. **(D)** Shows Low-risk participant predicted as Low-risk with 27% probability. Feature labels are: **esmoked** (0 = never smoked); **emotional disorders** (0 = no); **hypertension** (0 = no); **osteroarthritis** (1 = yes); **high cholesterol** (0 = no); **heart attack** (0 = no); **education** (2 = lower secondary education or second stage of basic education; 3 = upper secondary education); **moderate sports** (0 = hardly ever, or never, 1 = one to three times a month); **vigorous sports** (0 = hardly ever, or never, 1 = one to three times a month); **no children** (0 = no children); **widowed** (1 = yes); **BMI** [1 = under weight (<18.5)]; **married, living with spouse** (0 = no) and **gender** (1 = male).

including RF and XGBoost, to predict mild cognitive impairment to dementia conversion with highly skewed class distribution, and XGBoost demonstrated superior performance over the rest of the algorithms and outperforming RF, which is consistent with the study's findings. Nevertheless, the study's model may benefit from incorporating some of the numerous imbalanced data techniques discussed by Fernández et al. (2018) in the processing pipeline as part of future work. Lastly, all missing data were removed from the training set as part of the pre-processing step, which may have led to loss of data. This approach is not ideal and sub-optimal particularly when dealing with longitudinal datasets with long follow-up periods as well as real-world datasets, which mostly have a high prevalence of missing data. Therefore, approaches to handling missing data such as those described by Buck (1960) could potentially be explored.

Even though the study's source model achieved a relatively good performance, the performance of our target model could be better. The 63% AUROC score and a transfer learning efficacy rate of 20% achieved by the study's target model could be attributed to the limited sample used to update the decision boundaries of the study's source model. This could be considered

a limitation, and therefore a bigger sample size will be required to further update and evaluate the model.

CONCLUSION

Drawing on the transfer learning paradigm of artificial intelligence, we developed ensemble-based models capable of predicting Alzheimer's dementia onset in a relatively younger population up to 14 years in advance of the mean in the training set with promising results. The models not only predict dementia risk but also provide a visualisation of the interactions between risk factors to determine those driving the risk prediction at the individual level. The complex nature of dementia requires powerful machine learning models to be able to learn complex patterns from the interactions between risk factors, and the study's proposed model achieves this with reasonable accuracy. While some of the risk factors identified are well-documented, our model further identified less suspected risk factors that appear to be significant in driving the risk of AD. We believe that with further development and validation, our prediction model has the potential to support the early detection for appropriate interventions to be developed to prevent dementia.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analysed in this study. This data can be found at: DOIs: 10.6103/SHARE.w1.710, 10.6103/SHARE.w2.710, 10.6103/SHARE.w3.710, 10.6103/SHARE.w4.710, 10.6103/SHARE.w5.710, 10.6103/SHARE.w6.710, 10.6103/SHARE.w7.710. PREVENT WL210v1.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by PREVENT Dementia Programme Consortium. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

SD designed the study. SD and ZZ carried out the experiments and analysed the results. All authors contributed to drafting the manuscript.

FUNDING

This work was supported by The PREVENT dementia programme funded by grants from the UK Alzheimer's Society (Grant Numbers: 178 and 264), the U.S. Alzheimer's Association (Grant Number: TriBEKa-17–519007), and philanthropic donations.

ACKNOWLEDGMENTS

This paper used data from SHARE Waves 1, 2, 3, 4, 5, 6, and 7 (DOIs: 10.6103/SHARE.w1.710, 10.6103/SHARE.w2.710, 10.6103/SHARE.w3.710, 10.6103/SHARE.w4.710, and 10.6103/SHARE.w5.710, 10.6103/SHARE.w6.710, 10.6103/SHARE.w7.710) (see Börsch-Supan et al., 2013 for methodological details). The SHARE data collection has been funded by the European Commission through FP5 (QLK6-CT-2001-00360), FP6 (SHARE-I3: RII-CT-2006-062193, COMPARE: CIT5-CT-2005-028857, SHARELIFE: CIT4-CT-2006-028812), FP7 (SHARE-PREP: GA N°211909, SHARE-LEAP: GA N°227822, SHARE M4: GA N°261982, DASISH: GA N°283646), and Horizon 2020 (SHARE-DEV3: GA N°676536, SHARE-COHESION: GA N°870628, SERISS: GA N°654221, SSHOC: GA N°823782) and by DG Employment, Social Affairs & Inclusion. Additional funding from the German Ministry of Education and Research, the Max Planck Society for the Advancement of Science, the U.S. National Institute on Ageing (U01_AG09740-13S2, P01_AG005842, P01_AG08291, P30_AG12815, R21_AG025169, Y1-AG-4553-01, IAG_BSR06-11, OGHA_04-064, HHSN271201300071C), and from various national funding sources is gratefully acknowledged (see www.share-project.org).

REFERENCES

Adam, H., Dyer, A. H., Murphy, C., Lawlor, B., Kennelly, S. P., Segurado, R., et al. (2020). Cognitive outcomes of long-term benzodiazepine and related drug (BDZR) use in people living with mild to moderate Alzheimer's disease: results from NILVAD. *J. Am. Med. Direct. Assoc.* 21, 194–200. doi: 10.1016/j.jamda.2019.08.006

Barnes, D. E., Covinsky, K. E., Whitmer, R. A., Kuller, L. H., Lopez, O. L., and Yaffe, K. (2009). Predicting risk of dementia in older adults: the late-life dementia risk index. *Neurology* 73, 173–179. doi: 10.1212/WNL.0b013e3181a81636

Bergstra, J., and Bengio, Y. (2012). Random search for hyper-parameter optimization. *J. Machine Learn. Res.* 13, 281–305.

Börsch-Supan, A., Brandt, M., Hunkler, C., Kneip, T., Korbmacher, J., Malter, F., et al. (2013). Data resource profile: the Survey of Health, Ageing and Retirement in Europe (SHARE). *Int. J. Epidemiol.* 42, 992–1001. doi: 10.1093/ije/dyt088

Breiman, L. (2001). Statistical modeling: the two cultures (with comments and a rejoinder by the author). *Stat. Sci.* 16, 199–231. doi: 10.1214/ss/1009213726

Breiman, L., Friedman, J., Stone, C. J., and Olshen, R. A. (1984). *Classification and Regression Trees*. CRC Press.

Buck, S. F. (1960). A method of estimation of missing values in multivariate data suitable for use with an electronic computer. *J. R. Stat. Soci. B* 22, 302–306.

Caruana, R., Lou Y., Gehrke J., Koch P., Sturm M., and Elhadad, N. (2015). "Intelligible models for healthcare: predicting pneumonia risk and hospital 30-day readmission," in *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (Sydney), 1721–1730. doi: 10.1145/2783258.2788613

Cui, R., Liu, M., and Alzheimer's Disease Neuroimaging Initiative (2019). RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. *Computer. Med. Imaging Graph.* 73, 1–10. doi: 10.1016/j.compmedimag.2019.01.005

DeLong, E. R., DeLong, D. M., and Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics* 44, 837–845. doi: 10.2307/2531595

Efron, B., and Tibshirani, R. J. (1994). *An Introduction to the Bootstrap*. New York, NY: Chapman & Hall.

Facal, D., Valladares-Rodriguez, S., Lojo-Seoane, C., Pereiro, A. X., Anido-Rifon, L., and Juncos-Rabadán, O. (2019). Machine learning approaches to studying the role of cognitive reserve in conversion from mild cognitive impairment to dementia. *Int. J. Geriatr. Psychiatry* 34, 941–949. doi: 10.1002/gps.5090

Fernández, A., García, S., Galar, M., Prati, R. C., Krawczyk, B., and Herrera, F. (2018). *Learning From Imbalanced Data Sets*. Berlin: Springer.

Frisoni, G. B., Molinuevo, J. L., Altomare, D., Carrera, E., Barkhof, F., Berkhof, J., et al. (2020). Precision prevention of Alzheimer's and other dementias: Anticipating future needs in the control of risk factors and implementation of disease-modifying therapies. *Alzheimer's Dement*. 16, 1457–1468. doi: 10.1002/alz.12132

Gaugler, J., James, B., Johnson, T., Marin, A., and Weuve, J. (2019). 2019 Alzheimer's disease facts and figures. *Alzheimers Dementia* 15, 321–387. doi: 10.1016/j.jalz.2019.01.010

Goerdten, J., Cukić, I., Danso, S. O., Carrière, I., and Muniz-Terrera, G. (2019). Statistical methods for dementia risk prediction and recommendations for future work: a systematic review. *Alzheimer Dementia Transl. Res. Clin. Intervent*. 5, 563–569. doi: 10.1016/j.trci.2019.08.001

Henry, K. E., Hager, D. N., Pronovost, P. J., and Saria, S. (2015). A targeted real-time early warning score (TREWScore) for septic shock. *Sci. Transl. Med.* 7, 299–309. doi: 10.1126/scitranslmed.aab3719

Houssami, N., Lee, C. I., Buist, D. S., and Tao, D. (2017). Artificial intelligence for breast cancer screening: opportunity or hype? *Breast* 36, 31–33. doi: 10.1016/j.breast.2017.09.003

Johnson, D. K., Storandt, M., Morris, J. C., and Galvin, J. E. (2009). Longitudinal study of the transition from healthy aging to Alzheimer disease. *Arch. Neurol.* 66, 1254–1259. doi: 10.1001/archneurol.2009.158

Khalilia, M., Chakraborty, S., and Popescu, M. (2011). Predicting disease risks from highly imbalanced data using random forest. *BMC Med. Informatics Decision Making* 11:51. doi: 10.1186/1472-6947-11-51

Kim, M. J., Kang, D. K., and Kim, H. B. (2015). Geometric mean based boosting algorithm with over-sampling to resolve data imbalance problem for bankruptcy prediction. *Expert Systems Appl.* 42, 1074–1082. doi: 10.1016/j.eswa.2014.08.025

Lee, S., Zhou, X., Gao, Y., Vardarajan, B., Reyes-Dumeyer, D., Rajan, K. B., et al. (2018). Episodic memory performance in a multi-ethnic longitudinal study of 13,037 elderly. *PLoS ONE* 13:e0206803. doi: 10.1371/journal.pone.0206803

Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., et al. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. *Lancet* 396, 413–446. doi: 10.1016/S0140-6736(20)30367-6

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., et al. (2017). Dementia prevention, intervention, and care. *Lancet* 390, 2673–2734. doi: 10.1016/S0140-6736(17)31363-6

Lundberg, S. M., Nair, B., Vavilala, M. S., Horibe, M., Eisses, M. J., Adams, T., et al. (2018). Explainable machine-learning predictions for the prevention of hypoxaemia during surgery. *Nat. Biomed. Eng.* 2, 749–760. doi: 10.1038/s41551-018-0304-0

Lyketsos, C. G., Lopez, O., Jones, B., Fitzpatrick, A. L., Breitner, J., and DeKosky, S. (2002). Prevalence of neuropsychiatric symptoms in dementia and mild cognitive impairment: results from the cardiovascular health study. *JAMA* 288, 1475–1483. doi: 10.1001/jama.288.12.1475

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *J. Thoracic Oncol.* 5, 315–1316. doi: 10.1097/JTO.0b013e3181ec173d

Natekin, A., and Knoll, A. (2013). Gradient boosting machines, a tutorial. *Front. Neurorob.* 7:21. doi: 10.3389/fnbot.2013.00021

Pan, S. J., and Yang, Q. (2009). A survey on transfer learning. *IEEE Trans. Knowl. Data Eng.* 22, 1345–1359. doi: 10.1109/TKDE.2009.191

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., et al. (2011). Scikit-learn: machine learning in Python. *J. Machine Learn. Res.* 12, 2825–2830.

Pellegrini, E., Ballerini, L., Hernandez, M. D. C. V., Chappell, F. M., González-Castro, V., Anblagan, D., et al. (2018). Machine learning of neuroimaging for assisted diagnosis of cognitive impairment and dementia: a systematic review. *Alzheimer Dementia Diagnosis Assessment Dis. Monitor*. 10, 519–535. doi: 10.1016/j.dadm.2018.07.004

Pollack, I. (1970). A nonparametric procedure for evaluation of true and false positives. *Behav. Res. Methods Instrument*. 2, 155–156. doi: 10.3758/BF03209289

Prince, M., Bryce, R., and Ferri, C. (2018). *World Alzheimer Report 2011. The Benefits of Early Diagnosis and Intervention*. Alzheimer's Disease International. Available online at: www.alz.co.uk/research/WorldAlzheimerReport2011.pdf

Ritchie, C. W., and Ritchie, K. (2012). The PREVENT study: a prospective cohort study to identify mid-life biomarkers of late-onset Alzheimer's disease. *BMJ Open* 2:e001893. doi: 10.1136/bmjopen-2012-001893

Ritchie, K., Ropacki, M., Albala, B., Harrison, J., Kaye, J., Kramer, J., et al. (2017). Recommended cognitive outcomes in preclinical Alzheimer's disease: consensus statement from the European Prevention of Alzheimer's Dementia project. *Alzheimer Dementia* 13, 186–195. doi: 10.1016/j.jalz.2016.07.154

Sharp, E. S., and Gatz, M. (2011). The relationship between education and dementia an updated systematic review. *Alzheimer Dis. Assoc. Disord.* 25:289. doi: 10.1097/WAD.0b013e318211c83c

Skolariki, K., Terrera, G. M., and Danso, S. O. (2021). Predictive models for mild cognitive impairment to Alzheimer's disease conversion. *Neural Regen. Res.* 16, 1766–1767. doi: 10.4103/1673-5374.306071

Song, J., Lee, W. T., Park, K. A., and Lee, J. E. (2014). Association between risk factors for vascular dementia and adiponectin. *BioMed Res. Int*. 2014:261672. doi: 10.1155/2014/261672

Taylor, M. E., and Stone, P. (2009). Transfer learning for reinforcement learning domains: a survey. *J. Mach. Learn. Res.* 10, 1633–1685.

van Maurik, I. S., Vos, S. J., Bos, I., Bouwman, F. H., Teunissen, C. E., Scheltens, P., et al. (2019). Biomarker-based prognosis for people with mild cognitive impairment (ABIDE): a modelling study. *Lancet Neurol.* 18, 1034–1044. doi: 10.1016/S1474-4422(19)30283-2

World Health Organization (2017). *Global Action Plan on the Public Health Response to Dementia*. Geneva. 2017–2025.

Yao, Y., and Doretto, G. (2010). "Boosting for transfer learning with multiple sources," in *Proceedings of 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition* (IEEE), 1855–1862. doi: 10.1109/CVPR.2010.5539857

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

frontiers in Big Data

ORIGINAL RESEARCH
published: 30 July 2021
doi: 10.3389/fdata.2021.577164

Differences Between MR Brain Region Segmentation Methods: Impact on Single-Subject Analysis

W. Huizinga[1], D. H. J. Poot[1], E. J. Vinke[2,3], F. Wenzel[4], E. E. Bron[1], N. Toussaint[5], C. Ledig[6], H. Vrooman[1], M. A. Ikram[3], W. J. Niessen[1,7], M. W. Vernooij[2,3] and S. Klein[1]*

[1]Biomedical Imaging Group Rotterdam, Department of Radiology & Nuclear Medicine and Medical Informatics, Erasmus MC, Rotterdam, Netherlands, [2]Department of Radiology & Nuclear Medicine, Erasmus MC, Rotterdam, Netherlands, [3]Department of Epidemiology, Erasmus MC, Rotterdam, Netherlands, [4]Philips Research Hamburg, Hamburg, Germany, [5]School of Biomedical Engineering, King's College London, London, United Kingdom, [6]Biomedical Image Analysis Group, Department of Computing, Imperial College London, London, United Kingdom, [7]Quantitative Imaging Group, Department of Imaging Physics, Faculty of Applied Sciences, Delft University of Technology, Delft, Netherlands

For the segmentation of magnetic resonance brain images into anatomical regions, numerous fully automated methods have been proposed and compared to reference segmentations obtained manually. However, systematic differences might exist between the resulting segmentations, depending on the segmentation method and underlying brain atlas. This potentially results in sensitivity differences to disease and can further complicate the comparison of individual patients to normative data. In this study, we aim to answer two research questions: 1) to what extent are methods interchangeable, as long as the same method is being used for computing normative volume distributions and patient-specific volumes? and 2) can different methods be used for computing normative volume distributions and assessing patient-specific volumes? To answer these questions, we compared volumes of six brain regions calculated by five state-of-the-art segmentation methods: Erasmus MC (EMC), FreeSurfer (FS), geodesic information flows (GIF), multi-atlas label propagation with expectation–maximization (MALP-EM), and model-based brain segmentation (MBS). We applied the methods on 988 non-demented (ND) subjects and computed the correlation (PCC-v) and absolute agreement (ICC-v) on the volumes. For most regions, the PCC-v was good (>0.75), indicating that volume differences between methods in ND subjects are mainly due to systematic differences. The ICC-v was generally lower, especially for the smaller regions, indicating that it is essential that the same method is used to generate normative and patient data. To evaluate the impact on single-subject analysis, we also applied the methods to 42 patients with Alzheimer's disease (AD). In the case where the normative distributions and the patient-specific volumes were calculated by the same method, the patient's distance to the normative distribution was assessed with the z-score. We determined the diagnostic value of this z-score, which showed to be consistent across methods. The absolute agreement on the AD patients' z-scores was high for regions of thalamus and putamen. This is encouraging as it indicates that the studied methods are interchangeable for these regions. For regions such as the hippocampus, amygdala, caudate nucleus and accumbens, and globus pallidus, not all method combinations showed a high ICC-z. Whether two methods are

OPEN ACCESS

Edited by:
Holger Fröhlich, University of Bonn, Germany

Reviewed by:
Alex Pagnozzi, Commonwealth Scientific and Industrial Research Organisation (CSIRO), Australia
Alberto Redolfi, Centro San Giovanni di Dio Fatebenefratelli (IRCCS), Italy

***Correspondence:**
W. Huizinga
wykehuizinga@gmail.com

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *28 June 2020*
Accepted: *21 May 2021*
Published: *30 July 2021*

Citation:
Huizinga W, Poot DHJ, Vinke EJ, Wenzel F, Bron EE, Toussaint N, Ledig C, Vrooman H, Ikram MA, Niessen WJ, Vernooij MW and Klein S (2021) Differences Between MR Brain Region Segmentation Methods: Impact on Single-Subject Analysis. Front. Big Data 4:577164. doi: 10.3389/fdata.2021.577164

indeed interchangeable should be confirmed for the specific application and dataset of interest.

Keywords: brain region segmentation, subcortical, comparison study, normative modeling, magnetic resonance imaging

1 INTRODUCTION

Quantitative imaging biomarkers are biological features that can be measured using medical images. They are of interest for diagnosis when changes in these features are due to disease. In the case of traumatic brain injury or neurodegenerative disease, typical valuable quantitative imaging biomarkers are brain region volumes (Zagorchev et al., 2015; Ledig et al., 2015; Scheltens et al., 2002). A well-known example is the volume of the hippocampus. A relatively low volume may indicate the presence of Alzheimer's disease (AD)' (Convit et al., 1997; Jack et al., 1999; den Heijer et al., 2006). To determine if a patient deviates significantly, one can compare it to the so-called normative data (Brewer, 2009; Ziegler et al., 2014; Marquand et al., 2016). Normative data are acquired in a reference population, and they are used as baseline distribution for a measurement, against which an individual measurement can be compared. Normative data may incorporate covariates such as age or gender, when the distribution is expected to vary significantly as a function of these variables. Well-known examples are head-circumference-for-age, height-for-age, weight-for-age, and weight-for-height norms, provided by the WHO (de Onis et al., 2006), for detecting abnormal growth in children. The dependency on age is also the case for volumetric magnetic resonance (MR) brain images. Brewer (2009) proposed using quantile curves as a function of age as normative data for volumetric MR measurements.

Volumetric MR measurements are acquired by segmenting the brain into its different tissue types and regions of interest. The manual segmentation of a brain image is a time-consuming task, which has to be performed by an expert and is therefore too expensive and impractical for a clinical setting (Brewer (2009)). To automatically obtain brain region volumes from MRI brain data, numerous fully automated brain segmentation methods have been proposed in the literature. Each method relies on different techniques to segment either the full brain or a specific region. We can subdivide the methods that are based on prior probability maps (Fischl et al., 2002), statistical shape and appearance models (Babalola et al., 2008a; Patenaude et al., 2011; Wenzel et al., 2018), multi-atlas registration and labeling (Bron et al., 2014; Cardoso et al., 2015; Ledig et al., 2015; Murphy et al., 2014; Wang et al., 2014; Wolz et al., 2010; van der Lijn et al., 2008), deep learning approaches (Bao and Chung, 2018; Shakeri et al., 2016; de Brébisson and Montana, 2015), and other (Hammers et al., 2009; Corso et al., 2007; Morra et al., 2008; Tue et al., 2008). Each method aims to segment the brain as accurately as possible where manual segmentation serves as the gold standard.

Various comparison studies have been performed with regard to automated brain segmentation methods. Grimm et al. (2015) assessed the differences in amygdalar and hippocampal volume resulting from Freesurfer (Fischl et al., 2002), VBM8 (VBM[1]), and manual segmentation. They concluded that volumes computed with VBM8 and Freesurfer V5.0 were comparable, and systematic and proportional differences were mainly due to different definitions of anatomic boundaries. They concluded that large differences can still exist even with high correlation coefficients. Morey et al. (2009) also compared amygdalar and hippocampal volumes but using methods such as FSL/FIRST 4.0.1[2], Freesurfer 4.0.5 (Fischl et al., 2002), and manual segmentation. They concluded that for the hippocampus, Freesurfer was more similar to manual segmentation in terms of volume difference, overlap, and correlation. For the amygdala, FIRST represented the shape more accurately than Freesurfer. Babalola et al. (2008b) compared four different state-of-the-art algorithms for automatic segmentation of subcortical structures in MR brain images and evaluated spatial overlap, distance, and volumetric measures: classifier fusion and labeling (Aljabar et al., 2007), profile active appearance models (Babalola et al., 2007), Bayesian appearance models (Patenaude et al., 2011), and expectation–maximization–based segmentation using a dynamic brain atlas (Murgasova et al., 2006). They concluded that all four methods perform on par with recently published methods. One of their evaluating methods (Aljabar et al., 2007) performed significantly better than the other three methods according to their evaluation. Perlaki et al. (2017) compared the segmentation accuracy of the caudate nucleus and putamen between FSL/FIRST (version FSL's build: 507) and Freesurfer (versions 4.5 and 5.3) by studying the Dice coefficient, and absolute and relative volume difference. They also measured consistency and absolute agreement. They concluded that for caudate segmentation, FIRST and Freesurfer 4.5 and 5.3 performed similarly, but for putaminal segmentation, FIRST was superior to Freesurfer 5.3.

The impact, however, of using different methods on the analyses of individual patients within a normative modeling framework is still unknown. This is relevant when volumetric MR data are used to generate normative distributions for both research and clinical use. In this study, we therefore aim to answer two research questions: 1) to what extent are methods interchangeable, as long as the same method is being used for deriving normative volume distributions and patient-specific volumes? and 2) can different methods be used for deriving normative volume distributions and patient-specific volumes? To answer these questions, we evaluated five state-of-the-art segmentation methods (Bron et al., 2014; Wenzel et al., 2018;

[1]http://dbm.neuro.uni-jena.de/wordpress/vbm/
[2]https://fsl.fmrib.ox.ac.uk/fsl/fslwiki

Cardoso et al., 2015; Ledig et al., 2015; Fischl et al., 2002; Ikram et al., 2015).

2 MATERIAL AND METHODS

2.1 Data

To derive the normative distributions as a function of age, we applied the brain region segmentation methods to a subset of the population-based Rotterdam Scan Study, a prospective longitudinal study among community-dwelling subjects aged 45 years and older (Ikram et al., 2015). This subset is uniformly distributed over age and consists of 988 T1w MR brain images from non-demented (ND) (425 male, age = 68.1 ± 13.0 years). The total sample size of the Rotterdam Scan Study is larger: as of July 2015, a total of 12,174 brain MR scans have been obtained on the research scanner in over 5,800 individuals (Ikram et al., 2015). The 988 subjects form a subset with uniform age distribution (433 male, age = 68.3 ± 13.0 (mean ± std)). We adopted this dataset from Huizinga et al. (2018). All brain images were acquired on a single 1.5T MRI system (GE Healthcare, US). The T1w imaging protocol was a 3-dimensional fast radiofrequency spoiled gradient recalled acquisition with an inversion recovery pre-pulse sequence (Ikram et al., 2015). The images were reconstructed to a voxel size of $0.5 \times 0.5 \times 0.8\,\mathrm{mm}^3$, and the number of voxels in each dimension was $512 \times 512 \times 192$.

In addition, we used the brain images of 42 (25 male, age = 81.9 ± 4.9 years) patients with AD at the time of the MRI scan from the same imaging study. Different MR acquisition protocols may lead to different image contrasts, and since most automated methods are—partly or entirely—driven by the contrast in the image; this may influence the segmentation results. To rule out possible differences of the segmentation due to the acquisition protocol, the methods were applied to the same images, all acquired with the same acquisition protocol (Ikram et al. (2015)).

2.2 Brain Segmentation Methods

We applied five previously proposed brain segmentation methods to the imaging data. The following five segmentation methods, explained in detail later, were evaluated:

1. Multi-atlas registration combined with tissue segmentation for cortical regions, developed at Erasmus MC (EMC), the Netherlands;
2. Freesurfer 5.1 (FS), developed at the Athinoula A. Martinos Center for Biomedical Imaging at Massachusetts General Hospital, United States of America;
3. Geodesic information flows (GIF), developed at University College London, United Kingdom;
4. Multi-atlas label propagation with expectation–maximization–based refinement (MALP-EM), developed at Imperial College London, United Kingdom; and
5. Model-based brain segmentation (MBS), developed at Philips Research Hamburg, Germany.

The regions segmented by each method are shown in **Table 1**. Later, a short description of each method is given.

2.2.1 EMC

This method combines multi-atlas registration and voxel-wise tissue segmentation for cortical regions, and hippocampus and amygdala. Probabilistic tissue segmentations are obtained on the image to be segmented using the unified tissue segmentation method (Ashburner and Friston, 2005) of SPM8 (Statistical Parametric Mapping, London, United Kingdom). Thirty labeled T1-weighted MR brain images are used as atlas images (Gousias et al., 2008; Hammers et al., 2003). The atlas images are registered to the subjects' image using a rigid, affine, and non-rigid transformation model consecutively, and a mutual information-based similarity measure. The subjects' images are corrected for inhomogeneities to improve registrations using the N3 algorithm (Tustison et al., 2010). Labels are fused using a majority voting algorithm (Heckemann et al., 2006). For the cortical regions, as well as hippocampus and amygdala, the label-map is combined with the tissue map such that the brain region volumes are determined on gray matter voxels only. For subcortical regions, the volumes are determined with a multi-atlas segmentation only as the probabilistic tissue segmentation for these regions is inaccurate. A more detailed description of this method can be found in Bron et al. (2014).

2.2.2 FS

Freesurfer is widely used neuroimaging software developed by the Laboratory for Computational Neuroimaging at the Athinoula A. Martinos Center for Biomedical Imaging at Massachusetts General Hospital. It has many applications, but in this work, we use the brain region segmentation method described in Fischl et al. (2002). The method defines the problem of segmentation using a Bayesian approach in which the probability is estimated of a segmentation, given the observed image. First, the image is transformed into the atlas space with an affine transformation. Manually labeled atlas images provide the prior spatial information of the brain regions. The final segmentation is estimated by combining this spatial information with the intensity distribution of each brain region in the individual image. (For more detailed information about this method, we refer the reader to Fischl et al. (2002).) In our experiments, we used FS version 5.1. The user is able to use his own atlas, however, we used the atlas provided by FS. This method is publicly available[3].

2.2.3 GIF

This method is atlas-based and uses the geodesic path of a spatially variant graph to propagate the atlas labels (Cardoso et al., 2015). The atlas image database contains 130 T1-weighted MR brain images of cognitively normal participants from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study and 35 T1-weighted MR brain images from 30 young controls of the OASIS database

[3]http://freesurfer.net/

TABLE 1 | Characteristics of each method. The input format of each method is a 3D NIFTI file.

Method	References	Used reference data	Method of segmentation	# Regions	Region description
EMC	Bron et al. (2014)	Hammers et al. (2003), Gousias et al. (2008)	Multi-atlas segmentation with majority voting for label fusion	83	Subcortical regions, cortical regions, ventricles, corpus callosum, substantia nigra, lobes, brain stem, and cerebellum
FS	Fischl et al. (2002)	Fischl et al. (2002)	Multi-atlas segmentation with a Bayesian approach for label assignment	261	Subcortical regions, cortical regions, ventricles, lobes, optic chiasm, ventral diencephalon, lesions, vessels, corpus callosum, choroid plexus, brain stem, and cerebellum
GIF	Cardoso et al. (2015)	Petersen et al. (2010), Marcus et al. (2007) and Neuromorphometrics4	Multi-atlas segmentation with heat-kernel–weighted label fusion	144	Subcortical regions, cortical regions, ventricles, optic chiasm, ventral diencephalon, lesions, vessels, lobes, brain stem, and cerebellum
MALP-EM	Ledig et al. (2015)	Marcus et al. (2007) and Neuromorphometrics4	Multi-atlas segmentation with label refinement using prior information	138	Subcortical regions, cortical regions, ventricles, lobes, brain stem, and cerebellum
MBS	Wenzel et al. (2018)	Petersen et al. (2010), an Alzheimer's disease study at the Lahey Clinic, Burlington, MA	Model-based segmentation using a pretrained shape-constrained deformable surface model	56	Subcortical regions, ventricles, corpus callosum, fornix, septum pellucidum, lobes, brain stem, pons, and cerebellum

EMC is the method Erasmus MC by Bron et al. (2014), FS is the method FreeSurfer by Fischl et al. (2002), GIF is the method geodesic information flows by Cardoso et al. (2015), MALP-EM is the method multi-atlas label propagation with expectation–maximization–based refinement by Ledig et al. (2015), and MBS is the method model-based segmentation by Wenzel et al. (2018).

(Marcus et al., 2007). The labeled images are made publicly available by Neuromorphometrics[4] under academic subscription, as part of the MICCAI 2012 Grand Challenge on label fusion. First, each atlas image is registered to the individual image using a non-rigid transformation. A morphological distance of this image to each atlas image is estimated using the displacement field resulting from the image registration and the intensity similarity. The segmentation is estimated by fusing the labels of the morphologically closest atlas images. (For more details about this method, we refer the reader to Cardoso et al. (2015).) This method is publicly available[5].

2.2.4 MALP-EM

Like EMC, this method also combines multi-atlas registration and voxel-wise tissue segmentation. The atlas database of this method consists of 35 manually annotated T1-weighted MR brain images of 30 subjects of the OASIS database, which are also part of the atlas images of the GIF method (see **Section 2.2.3**). The atlas images of these 30 subjects are transformed to the space of the image that is to be segmented. These transformations are obtained via a non-rigid image registration approach (Heckemann et al., 2010). The subjects' brains are extracted using the method proposed in Heckemann et al. (2015). The resulting 30 label images are fused, and a probabilistic map of each brain region is obtained. The labels are refined using expectation–maximization (EM) (Leemput et al., 1999), a brain tissue segmentation technique based on the image intensities. (More details can be found in Ledig et al. (2015).) In our experiments, we used MALP-EM version 1.2. This method is publicly available[6].

2.2.5 MBS

The MBS method is based on the model-based brain segmentation presented in Wenzel et al. (2018). The model is shape-constrained and represented by a triangulated mesh of fixed topology. Shape variations are modeled by principal component analysis of manually annotated meshes of a set of training images, resulting in a point distribution model (PDM) with a mean mesh and shape modes (Cootes et al., 1992). To segment a new image, the mean mesh is placed within the image by a generalized Hough transform compensating global translation and translation. Subsequently, the mean mesh is adapted by a global affine transformation and then region-specific affine transformations by adding weighted shape modes. The global and local affine transform parameters and the mode weights are estimated using a boundary detection based, for example, on the local intensity gradient and a penalization component regularizing the mesh shape, including the PDM. Finally, in a deformable deformation step, triangles can adapt individually, leading to a close match of the model surface with the image boundaries.

A database of 96 3T scans following the MP-RAGE acquisition protocol, split over three vendors (GE, Siemens, and Philips) served as training data. These scans have been randomly selected from the ADNI study ($n = 87$) and an Alzheimer's disease study at the Lahey Clinic, Burlington, MA ($n = 9$). Ground truth delineations mostly followed structure definitions of the CMA guidelines,[7] with two exceptions: (1) lateral thalamus borders

[4]http://neuromorphometrics.com/

[5]http://cmicti g.cs.ucl.ac.uk/niftyweb/program.php?p=GIF

[6]https://github.com/ledigchr/MALPEM

[7]https://web.archive.org/web/20180226014735/http://www.cma.mgh.harvard.edu/manuals/

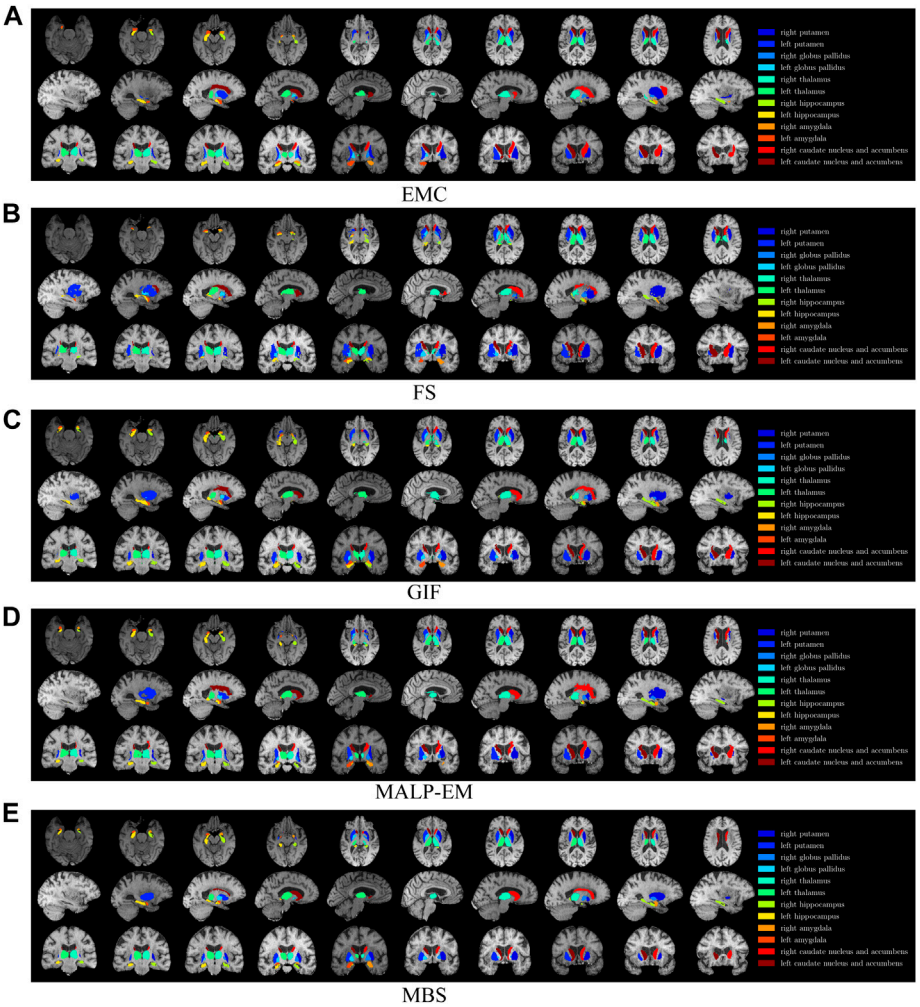


FIGURE 1 | T1w MR brain image from one of the subjects, with a colored overlay of the brain regions analyzed in this work, segmented with all methods. Slices in the axial direction are shown in the top row, slices in the saggital direction are shown in the middle row, and slices in the coronal direction are shown in the bottom row. The legend on the right side shows the regions and their corresponding colors in the overlay. Note that only for this visualization, the segmentations were registered to the MNI space; some differences might be due to imperfections of this registration.

follow image contrast, which may deviate from the CMA description, and (2) hippocampus annotations follow the EADC-ADNI harmonized protocol[8] (Boccardi et al., 2015a; Boccardi et al., 2015b). The training data and procedure are extensively described in Wenzel et al. (2018).

2.3 Regions of Interest

The set of brain regions in which each image is segmented differs per method. In this study, we focus on the following $S = 6$ regions: hippocampus, amygdala, caudate nucleus and accumbens, putamen, thalamus, and globus pallidus. **Figure 1** shows an example image of an ND subject with the analyzed brain regions in colored overlay. In the analysis, the volumes of the regions in the left hemisphere and the right hemisphere were summed.

For all methods except MBS, the volume of the caudate nucleus was added to the accumbens volume because MBS already segments these as a single region.

2.4 Outlier Detection

Segmentation errors may occur due to bad image quality, pathology, or other method-related problems. These errors could lead to outliers in the volume data and may influence the statistics excessively. We therefore remove them from the volume data prior to the statistical analyses.

The segmentations of the ND subjects were not visually inspected as this would be too time-consuming. Method failures, that is, when the software pipeline did not result in a segmentation for the image, were excluded. On the remaining

[8]http://www.hippocampal-protocol.net/SOPs/index.php

images, outliers were defined as having an absolute z-score higher than 5.0, derived with the population mean and standard deviation. Note that a z-score >5.0 does not necessarily imply a failed segmentation. We chose an absolute z-score of >5.0, instead of the typical value of 3.0 because we wanted to include as much of the normal population as possible to generate the normative data, but we did not want to contaminate the normative data with unrealistic volumes. The segmentations of the AD patients were visually inspected, and obviously failed regions were excluded.

2.5 Statistical Analyses

In the analyses, two scenarios are considered: 1) both the normative volume distribution and the patient-specific volumes are calculated by the same method, and 2) the normative volume distribution and the patient-specific volumes are calculated by different methods. The requirements for two methods to yield comparable results under scenario 1) are given as follows:

i) a high correlation on the absolute volumes, measured with the Pearson's correlation coefficient (PCC) and referred to as PCC-v;
ii) a high absolute agreement on the patient's distances relative to the normative distribution, that is, a high absolute agreement on the patients' z-scores, measured with the intraclass correlation coefficient (ICC) and referred to as ICC-z.

The requirements for two methods to yield comparable results under scenario 2) are given as follows:

i) a high absolute agreement on the absolute volumes, measured with the intraclass correlation coefficient (ICC) and referred to as ICC-v;
ii) a high absolute agreement on the patients' z-scores, measured with the intraclass correlation coefficient (ICC) and referred to as ICC-z.

For scenario 2), requirement i naturally results in requirement ii. The requirements for scenario 2) are stricter than those for scenario 1). If in scenario 1), an offset or scaling is present in the volumes of different methods, the resulting patient's z-score will be the same because the same method is used for comparing the patient to the normative distribution. However in scenario 2), absolute agreement on the volumes is necessary, that is, no offset or scaling is allowed for comparing the patient to the normative distribution as an offset or scaling will affect the patient's z-score. The next sections describe how the normative distribution was established, how the correlation and absolute agreement are measured, and, in the case of scenario 1), how the diagnostic value of the z-scores was assessed.

2.5.1 Normative Distribution Fitting

We fit an age-dependent normative distribution with the previously proposed LMS method (Cole and Green (1991)). This method assumes that the data are standard and normally distributed after applying the Yeo–Johnson transformation (Yeo and Johnson (2000)). The method estimates the λ–parameter of this transformation (L), the median (M), and coefficient of variation (S) for the appropriate volume at each age. With these three parameters, z-scores can be computed at each age. The smoothness of the resulting iso–z-score curves is influenced by the degrees of freedom δ, a user-defined parameter. In our experiments, we set the smoothness parameter δ to a value of 2. We used R-package VGAM for fitting these iso–z-score curves (Yee, 2010). The value of the brain region volume may also be influenced by other covariates than age, for example, gender and height. We correct for these covariates in the fitting procedure.

2.5.2 Correlation and Absolute Agreement

To verify if scenario 1) is applicable, we first measure the correlation of the volumes calculated by the methods, with the Pearson's correlation coefficient (PCC). We refer to these correlations as PCC-v. This coefficient is invariant for an offset and scaling of the data.

To verify if scenario 2) is applicable, we compute the absolute agreement on the volumes, which was measured with the intraclass correlation coefficient (ICC). The type of ICC to be chosen depends on the problem at hand. McGraw and Wong (1996) give an overview of the possible ICCs. For the presented experiments, ICC(A,1) is the appropriate absolute agreement measure (McGraw and Wong, 1996). Let X be an $n \times k$ matrix where each column contains the measurements of a single method and each row contains the measurements of a single subject, then ICC(A,1) is given by McGraw and Wong (1996) is given as follows:

$$\mathrm{ICC}(\mathrm{A},1) = \frac{\mathrm{MS_R}(X) - \mathrm{MS_E}(X)}{\mathrm{MS_R}(X) + (k-1)\mathrm{MS_E}(X) + \frac{k}{n}(\mathrm{MS_C}(X) - \mathrm{MS_E}(X))}, \tag{1}$$

where $\mathrm{MS_R}(X)$ is the mean square for rows, $\mathrm{MS_C}(X)$ is the mean square for columns, and $\mathrm{MS_E}(X)$ is the mean square error, which is defined as follows:

$$\mathrm{MS_E}(X) = \frac{1}{(n-1)(k-1)} \sum_{i,j=1}^{nk} \left[X_{ij} - \overline{X}_i - \overline{X}_j + \overline{X}\right]^2, \tag{2}$$

where $\overline{X}_i = \frac{1}{k}\sum_{j=1}^{k} X_{ij}$, $\overline{X}_j = \frac{1}{n}\sum_{i=1}^{n} X_{ij}$, and $\overline{X} = \frac{1}{nk}\sum_{i,j=1}^{nk} X_{ij}$. We refer to the absolute agreement on the volumes as ICC-v. The absolute agreement is maximal (1.0) when the measurements are exactly the same. When one or more measurements deviate, the absolute agreement is no longer 1.0 and drops according to how large the deviation is. A systematic error causing an offset in the measurements with a magnitude of, for example, the population standard deviation would lower the absolute agreement to ~0.67. Or a scaling of the data by a factor of 1.2 would lower the absolute agreement to ~0.7. The higher the ICC-v, the more reasonable it is to interchange methods.

We report all possible pairwise method combinations of PCC-v and ICC-v for $M = 5$ methods for each of the S brain regions. Since the correlation and absolute agreement are determined with symmetric measures, we present PCC-v and ICC-v of the

methods in a single 5×5 table, for each of the analyzed brain regions.

2.5.3 Absolute Z-Score Agreement

To further assess the applicability of scenario 1), we also computed the absolute agreement on the AD patient z-scores with ICC(A,1). We indicated these values with ICC-z. We present ICC-z on AD subjects with PCC-v for ND subjects (see **Section 2.5.2**) in the same table, to facilitate their comparison.

2.6 AUC

To estimate how well the AD patient z-scores discriminate between normative volumes and patient-specific volumes in scenario 1), we determine the area under the receiver operating characteristic curve (AUC) of the z-score. The z-score was computed, as described in **Section 2.5.1**. The expected z-scores for the AD patients are <0, since we expect their brain structure volume to be lower than normal. We therefore define the AUC as the probability that a randomly chosen ND subject will have a higher z-score than a randomly chosen AD patient. The higher the AUC, the better will be the discrimination between AD patients and ND subjects. Since not every region is a known discriminative biomarker for AD, it is not necessarily expected that the AUC is high for each region. The hippocampus and amygdala are known to be discriminative biomarkers for AD, so for these regions, a high AUC is expected. For the computation of the AUC, only ND subjects within the age range of the AD patients, [71, 91] years, were included. A 95% confidence interval was computed by bootstrapping the z-scores 1,000 times.

3 RESULTS

We used the following rating scale for PCC-v, ICC-v, and ICC-z, adopted from the rules of thumb in Mukaka (2012):

- Poor: < 0.5
- Fair: $0.5 - 0.7$
- Good: $0.7 - 0.9$
- Excellent: > 0.9

3.1 Outlier Detection

Method FS failed for nine ND subjects, either by not finishing the segmentation pipeline or by giving a zero volume output for some of the analyzed brain regions. Visual inspection of the MRI scans of these subjects did not show pathology or severe artifacts that would clearly explain failure. The method EMC failed for one ND subject, which was due to the failure of the brain extraction tool (Smith (2002)), which is used at the beginning of the pipeline. The remainder of the methods provided a segmentation for all images. The number of outliers per region and method on the remaining 978 subjects is reported in **Table 2** Two T1w images of AD patients were excluded due to large scanning or motion artifacts. The number of failed segmentations per region and method in the remaining 40 images is shown in **Table 3**. In one image, there was a large lesion in the frontal lobe, affecting the segmentation of the caudate nucleus and accumbens of all methods. In one other image, the method MBS failed to segment the putamen and globus pallidus correctly.

3.2 Volume Distributions

Table 4 shows the mean and standard deviation of the volumes of the ND subjects for each method and region. We performed a one-way ANOVA test, which showed that the p-values for each brain structure is $p < 0.05$, indicating that the volume distributions differ significantly between the methods. A multiple comparison post hoc analysis was done with the Tukey test. This test showed a limited number of non-significant differences, namely, the amygdala for methods EMC vs. GIF, the thalamus for methods FS vs. GIF and FS vs. MBS, and, finally, the putamen for methods FS vs. GIF. All other pairwise differences were statistically significant. The hippocampus volume of methods EMC and GIF deviates substantially from the other methods. The method EMC deviates due to a different definition of the hippocampus in the atlases that are used by the methods. The Hammers' atlas (Hammers et al. (2003), Gousias et al. (2008)), used by the method EMC, defines the posterior border of the hippocampus such that the hippocampus tail is not included in the definition, whereas the other methods include the hippocampus tail. The method GIF deviates because it generally delineates the hippocampus in a larger volume. These same methods have a smaller average globus pallidus volume than the other methods. Visual inspection on a representative subset showed that these methods delineated a smaller globus pallidus. Methods MALP-EM and MBS calculated a smaller amygdala than the other methods.

Figure 2 shows the normative brain structure volume distribution fitted on 978 ND subjects, visualized in iso-z-score lines, for each method and brain structure. The red scatters show the volumes of the 40 AD patients, segmented with the same method as the normative distribution (scenario 1).

3.3 Correlation and Absolute Agreement

Table 5 present PCC-v and ICC-v for each pairwise combination of the five methods. For most regions, PCC-v was good (≥ 0.75) and was excellent for the region thalamus ($0.91 - 0.97$) and good to excellent for the putamen ($0.88 - 0.96$).

For the three smallest structures, the hippocampus, amygdala and globus pallidus, ICC-v was generally poor, with some exceptions. The combination MALP-EM–MBS scored relatively high on ICC-v compared to the other method combinations. Visual inspection on a representative subset showed that the delineated hippocampus, amygdala, and globus pallidus for MALP-EM and MBS was similar in shape, explaining the good ICC-v. For the amygdala, the combination GIF–EMC also showed a good ICC-v. The three larger structures, the caudate nucleus and accumbens, thalamus, and putamen, showed generally higher ICC-vs. Visual inspection showed that their shape was, on average, more similar, possibly due to the less irregular shape of these regions than the smaller regions. Some method combinations

TABLE 2 | Number of outliers in the ND subjects per method for each brain region. The outliers were defined as having an absolute z-score > 5.0, derived with the population mean and standard deviation. The ten subjects that failed in the in the postprocessing were not included. As the outliers of the methods may overlap, the last column of the tables indicates the number of subjects included in the statistical analysis.

	EMC	FS	GIF	MALP-EM	MBS	TOTAL N
Hippocampus	0	0	0	0	0	978
Amygdala	0	1	1	0	0	976
Caudate nucleus and accumbens	2	1	0	2	0	975
Thalamus	0	1	0	0	0	977
Putamen	0	2	0	1	0	976
Globus pallidus	0	0	0	0	0	978

EMC is Erasmus MC by Bron et al. (2014), FS is FreeSurfer by Fischl et al. (2002), GIF is geodesic information flows by Cardoso et al. (2015), MALP-EM is multi-atlas label propagation with expectation–maximization–based refinement by Ledig et al. (2015), and MBS is model-based segmentation by Wenzel et al. (2018).

TABLE 3 | Number of rejected segmentations in the AD subjects per method for each brain region, determined by visual inspection. The two subjects that failed in the postprocessing were not included. As the outliers of the methods may overlap, the last column of the tables indicates the number of subjects included in the statistical analysis.

	EMC	FS	GIF	MALP-EM	MBS	Total N
Hippocampus	0	0	0	0	0	40
Amygdala	0	0	0	0	0	40
Caudate nucleus and accumbens	1	1	1	1	1	39
Thalamus	0	0	0	0	0	40
Putamen	0	0	0	0	1	39
Globus pallidus	0	0	0	0	1	39

EMC is Erasmus MC by Bron et al. (2014), FS is FreeSurfer by Fischl et al. (2002), GIF is geodesic information flows by Cardoso et al. (2015), MALP-EM is multi-atlas label propagation with expectation–maximization–based refinement by Ledig et al. (2015), and MBS is model-based segmentation by Wenzel et al. (2018).

TABLE 4 | Mean (standard deviation) of brain region volumes in mm^3 for the ND subjects.

	Hippocampus	Amygdala	Caudate nucleus and accumbens	Thalamus	Putamen	Globus pallidus
EMC	3,652 (494)	2,289 (320)	8,428 (1,265)	11,926 (1,637)	8,049 (1,139)	1897 (281)
FS	7,533 (1,166)	2,664 (402)	7,995 (1,154)	12,328 (1,614)	9,008 (1,338)	2,834 (480)
GIF	8,766 (906)	2,284 (269)	7,882 (1,059)	12,581 (1,333)	9,014 (1,090)	1735 (207)
MALP-EM	5,723 (862)	1887 (299)	7,640 (1,568)	13,678 (1,654)	7,427 (1,218)	2,472 (349)
MBS	6,052 (782)	1775 (243)	7,280 (895)	12,422 (1,451)	7,746 (977)	2,561 (304)

EMC is the method Erasmus MC by Bron et al. (2014), FS is the method FreeSurfer by Fischl et al. (2002), GIF is the method geodesic information flows by Cardoso et al. (2015), MALP-EM is the method multi-atlas label propagation with the expectation–maximization–based refinement by Ledig et al. (2015), and MBS is the method model-based segmentation by Wenzel et al. (2018).

showed poor ICC-v values for these larger regions, for example, MBS—EMC and MBS–MALP-EM for the caudate nucleus and accumbens, and GIF–MALP-EM for the putamen. MALP-EM–MBS also had a fair PCC-v for the regions caudate nucleus and accumbens; however, the other combinations showed a good PCC-v, indicating that the low ICC-v can mainly be explained by a volume offset and/or scaling.

3.4 Absolute Z-Score Agreement

Table 6 shows ICC-z in the lower left triangle. In the upper-right triangle, PCC-v of the ND subjects is showed again, for easy comparison. ICC-z was good to excellent for regions thalamus (0.75 – 0.94) and putamen (0.83 – 0.96), fair to good for regions hippocampus (0.56 – 0.81), amygdala (0.65 – 0.88), and globus pallidus (0.50 – 0.72), and fair to excellent for the caudate nucleus and accumbens (0.51 – 0.96). The two method combinations with the lowest PCC-v of the caudate nucleus and accumbens, MBS–EMC and MBS–MALP-EM, also have the lowest ICC-z. This is also the case for the globus pallidus, where combinations EMC–FS and MALP-EM–FS have the lowest PCC-v and the lowest ICC-v.

3.5 AUC

Table 7 shows the AUC for each method and brain region. The highest AUC was achieved for the hippocampus (on average 0.79) and amygdala (on average 0.78), demonstrating their involvement in AD. For the thalamus and putamen, the AUC was > 0.5 for all methods, indicating that these regions are also affected by AD. For the method GIF, the AUC of regions thalamus and globus pallidus were high compared to the other methods. The methods FS, MBS, and GIF had comparable thalamus volumes for the ND subjects, but the AD thalamus volumes segmented by GIF were, on average, 120 mm^3

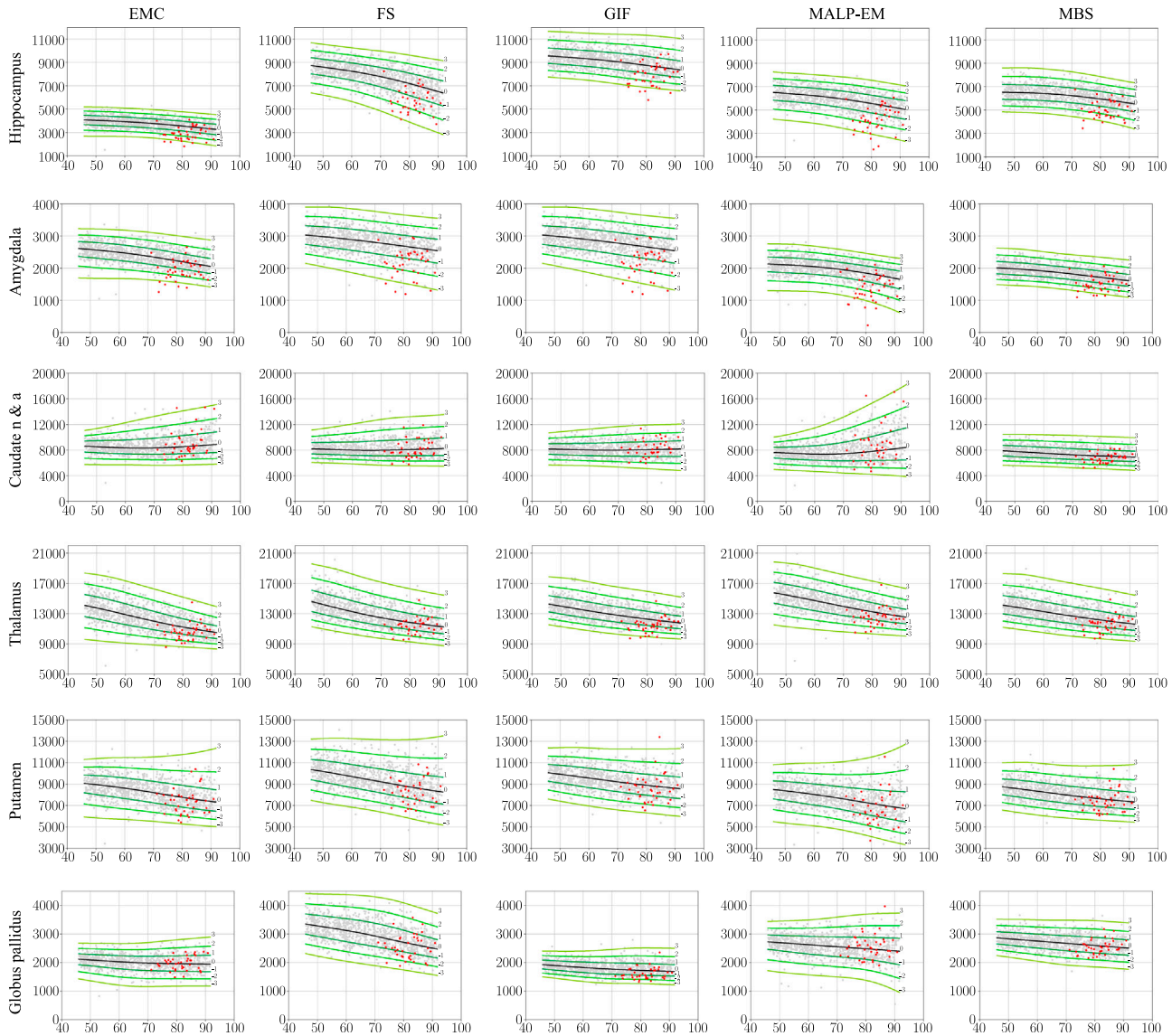


FIGURE 2 | Normative brain structure volume distribution fitted on 978 ND subjects, visualized in iso-z-score lines from −3 to 3. All volumes are given in mm^3 as a function of age [y]. The columns show volumes of each method, and the rows show the volumes per brain structure. The light gray scatters show the volumes of the ND subjects, and the red scatters show the volumes of the 40 AD patients, segmented with the same method as the normative distribution (scenario 1). The distribution was corrected for gender and height and is shown here for males of height 170 cm. EMC is the method Erasmus MC by Bron et al. (2014), FS is the method FreeSurfer by Fischl et al. (2002), GIF is the method geodesic information flows by Cardoso et al. (2015), MALP-EM is the method multi-atlas label propagation with expectation–maximization–based refinement by Ledig et al. (2015), and MBS is the method model-based segmentation by Wenzel et al. (2018). The caudate nucleus and accumbens was shortened to caudate n & a for visualization purposes.

lower than those segmented by MBS and 50 mm^3 lower than those segmented by FS. The methods EMC and GIF had comparable globus pallidus volumes for the ND subjects, but for AD subjects, the volumes segmented by GIF were, on average, 320 mm^3 lower than those segmented by EMC.

3.6 Computational Efficiency

All methods were executed on a Linux Sun Grid Engine (SGE) computing cluster with eight computing nodes, each having multiple cores. All methods, except FS, provide an option for using multiple cores. This is especially efficient for methods that use multi-atlas registration, where the registrations of the subjects in the atlas database can run in parallel. In practice, the method GIF had the longest computation time, despite the usage of multiple cores. This was mainly due to the non-rigid image registrations of the 165 images in the atlas database. The method MBS was most efficient, needing only a few minutes to segment all 56 regions in a brain image on a single core. Except for MALP-EM, needing 33 GB of RAM per brain image, the memory usage of the

TABLE 5 | PCC-v (upper-right triangle) and ICC-v (lower-left triangle) of ND volumes.

(a) Hippocampus

PCC-v / ICC-v	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.77	0.85	0.82	0.85
FS	0.05	1.00	0.77	0.80	0.76
GIF	0.03	0.44	1.00	0.83	0.87
MALP-EM	0.13	0.30	0.12	1.00	0.79
MBS	0.10	0.33	0.14	0.73	1.00

(b) Amygdala

PCC-v / ICC-v	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.80	0.77	0.79	0.80
FS	0.51	1.00	0.82	0.73	0.80
GIF	0.76	0.47	1.00	0.74	0.81
MALP-EM	0.43	0.20	0.37	1.00	0.75
MBS	0.29	0.15	0.27	0.68	1.00

(c) Caudate nucleus and accumbens

PCC-v / ICC-v	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.93	0.89	0.92	0.72
FS	0.87	1.00	0.93	0.86	0.80
GIF	0.79	0.92	1.00	0.82	0.85
MALP-EM	0.78	0.80	0.75	1.00	0.55
MBS	0.44	0.62	0.71	0.46	1.00

(d) Thalamus

PCC-v / ICC-v	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.92	0.93	0.96	0.91
FS	0.89	1.00	0.93	0.95	0.93
GIF	0.83	0.90	1.00	0.95	0.97
MALP-EM	0.61	0.71	0.73	1.00	0.95
MBS	0.86	0.92	0.96	0.71	1.00

(e) Putamen

PCC-v / ICC-v	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.91	0.94	0.93	0.93
FS	0.69	1.00	0.91	0.88	0.90
GIF	0.69	0.89	1.00	0.88	0.96
MALP-EM	0.81	0.50	0.45	1.00	0.87
MBS	0.88	0.54	0.55	0.82	1.00

(f) Globus pallidus

PCC-v / ICC-v	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.59	0.76	0.74	0.80
FS	0.13	1.00	0.71	0.66	0.75
GIF	0.61	0.10	1.00	0.77	0.83
MALP-EM	0.27	0.46	0.16	1.00	0.75
MBS	0.22	0.56	0.13	0.72	1.00

EMC is the method Erasmus MC by Bron et al. (2014), FS is the method FreeSurfer by Fischl et al. (2002), GIF is the method geodesic information flows by Cardoso et al. (2015), MALP-EM is the method multi-atlas label propagation with expectation–maximization–based refinement by Ledig et al. (2015), and MBS is the method model-based segmentation by Wenzel et al. (2018).

methods was modest (≤8 GB) for the hardware in modern computers.

4 DISCUSSION

We evaluated the correlation and absolute agreement on regional volumes computed with different automated brain segmentation methods, and the impact of the volume differences between these methods on single-subject analysis in a normative modeling framework. We evaluated two scenarios: 1) The normative volume distributions and the patient-specific volumes were calculated by the same method, and 2) the normative volume distributions was calculated by a different method than the patient-specific volumes. To this end, we applied five state-of-the-art automated brain segmentation methods on the T1w MR brain images of 988 ND subjects, and 42 AD patients acquired with the same MR acquisition protocol.

The PCC-v showed that the volumes of all regions correlated well, indicating that volume differences between methods in ND subjects are mainly due to systematic differences, such as the usage of different atlases and region definitions. The ICC-v however was generally low, especially

TABLE 6 | PCC-v of the ND volumes (upper-right triangle) and ICC-z of AD volume z-scores (lower-left triangle). The ICC-z is computed according to scenario 1.

(a) Hippocampus

PCC-v / ICC-z	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.77	0.85	0.82	0.85
FS	0.61	1.00	0.77	0.80	0.76
GIF	0.69	0.56	1.00	0.83	0.87
MALP-EM	0.72	0.61	0.69	1.00	0.79
MBS	0.79	0.57	0.78	0.81	1.00

(b) Amygdala

PCC-v / ICC-z	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.80	0.77	0.79	0.80
FS	0.80	1.00	0.82	0.73	0.80
GIF	0.85	0.70	1.00	0.74	0.81
MALP-EM	0.78	0.65	0.80	1.00	0.75
MBS	0.82	0.67	0.88	0.71	1.00

(c) Caudate nucleus and accumbens

PCC-v / ICC-z	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.93	0.89	0.92	0.72
FS	0.85	1.00	0.93	0.86	0.80
GIF	0.87	0.90	1.00	0.82	0.85
MALP-EM	0.96	0.83	0.85	1.00	0.55
MBS	0.58	0.78	0.69	0.51	1.00

(d) Thalamus

PCC-v / ICC-z	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.92	0.93	0.96	0.91
FS	0.85	1.00	0.93	0.95	0.93
GIF	0.88	0.75	1.00	0.95	0.97
MALP-EM	0.93	0.91	0.88	1.00	0.95
MBS	0.88	0.88	0.88	0.94	1.00

(e) Putamen

PCC-v / ICC-z	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.91	0.94	0.93	0.93
FS	0.86	1.00	0.91	0.88	0.90
GIF	0.93	0.95	1.00	0.88	0.96
MALP-EM	0.90	0.85	0.88	1.00	0.87
MBS	0.89	0.89	0.96	0.83	1.00

(f) Globus pallidus

PCC-v / ICC-z	EMC	FS	GIF	MALP-EM	MBS
EMC	1.00	0.59	0.76	0.74	0.80
FS	0.58	1.00	0.71	0.66	0.75
GIF	0.60	0.79	1.00	0.77	0.83
MALP-EM	0.69	0.50	0.62	1.00	0.75
MBS	0.72	0.69	0.71	0.60	1.00

EMC is the method Erasmus MC by Bron et al. (2014), FS is the method FreeSurfer by Fischl et al. (2002), GIF is the method geodesic information flows by Cardoso et al. (2015), MALP-EM is the method multi-atlas label propagation with expectation–maximization–based refinement by Ledig et al. (2015), and MBS is the method model-based segmentation by Wenzel et al. (2018).

for the smaller regions, including the hippocampus, amygdala, and globus pallidus. The low ICC-v indicates that the methods cannot be interchanged in a normative modeling framework and scenario 2) is not applicable. This also becomes visually clear from **Figure 2**, when comparing the location of the red dots across graphs in a row.

The ICC-z, with which the agreement on the AD patient position relative to the normative distribution was measured in the case of scenario 1), was good to excellent for the thalamus and putamen, which also showed a good to excellent PCC-v. The other four regions showed lower ICC-z, indicating that different methods would result in different AD patient positions relative to the normative distribution, even when the normative distribution was computed using the same method as the patient data. A low PCC-v also seemed to result in a low ICC-z. A high PCC-v however does not necessarily result in a high ICC-z. This may indicate that brain morphology changes because AD affects each method differently.

The AUC, with which the z-score discrimination between the patient and normative volumes was measured in the case of scenario 1), was relatively high for the regions hippocampus and amygdala for all methods, demonstrating the involvement of these regions in AD. For the method GIF, the thalamus volume showed to be a better discriminator for AD than the hippocampus volume, which is unexpected, as this region is not known for its involvement in AD, and the other methods did not show such a high AUC for the thalamus. A possible explanation is that the method GIF is more affected than the

TABLE 7 | AUC (95% confidence interval) for all regions, where the volumes of the normative distribution and the AD patients were generated by the same method (scenario 1).

	EMC	FS	GIF	MALP-EM	MBS
Hippocampus	0.78 (0.68,0.87)	0.83 (0.75,0.89)	0.73 (0.64,0.82)	0.80 (0.72,0.88)	0.80 (0.71,0.88)
Amygdala	0.77 (0.67,0.85)	0.81 (0.73,0.88)	0.76 (0.67,0.85)	0.82 (0.74,0.89)	0.73 (0.63,0.83)
Caudate nucleus and accumbens	0.52 (0.42,0.62)	0.56 (0.46,0.66)	0.47 (0.37,0.57)	0.49 (0.40,0.60)	0.63 (0.54,0.73)
Thalamus	0.68 (0.58,0.77)	0.63 (0.54,0.71)	0.76 (0.66,0.84)	0.69 (0.60,0.78)	0.66 (0.56,0.75)
Putamen	0.62 (0.52,0.72)	0.61 (0.51,0.71)	0.62 (0.52,0.72)	0.63 (0.53,0.73)	0.61 (0.51,0.71)
Globus pallidus	0.50 (0.41,0.60)	0.63 (0.52,0.74)	0.71 (0.61,0.81)	0.47 (0.38,0.56)	0.58 (0.49,0.68)

EMC is the method Erasmus MC by Bron et al. (2014), FS is the method FreeSurfer by Fischl et al. (2002), GIF is the method geodesic information flows by Cardoso et al. (2015), MALP-EM is the method multi-atlas label propagation with expectation–maximization–based refinement by Ledig et al. (2015), and MBS is the method model-based segmentation by Wenzel et al. (2018).

other methods by the brain morphology change due to AD, such as larger ventricles.

Several limitations of this study can be highlighted. First, the segmented results rely strongly on the atlas that was used by the method. As was shown with the hippocampus, differences in volume may be largely explained by the atlas and how the region was defined. For this reason, operationalized and quantitated landmark differences to help a Delphi panel converge on a set of landmarks on the hippocampus and provided a set of manually segmented images for training models for automatic hippocampus segmentation. In this study however, we considered the atlas a part of the method, and we did not study specific atlas-related volume differences. Second, the number of AD patients was limited, which limits the generalization of the conclusions drawn from these results. In future studies, a higher number of AD patients should be used to generalize the study results. Third, we used images that were acquired on a single 1.5 T scanner with the same acquisition protocol. This allowed us to study the effect of differences in segmentation methods, while not considering the confounding effect of differences in acquisition protocols. Future research should investigate how differences in acquisition protocols influence the comparison of individual patients to normative data and to study the generalizability of our results in more heterogeneous datasets. Previously, tools have been developed to cope with volumetric differences due to scanning artifacts. The effectiveness of these tools can be tested using our research setup with normative data. Finally, we limited our study to five automatic segmentation methods. Many more have been previously proposed, and it remains an active area of research, particularly since the rise of deep learning techniques (Bao and Chung, 2018; Shakeri et al., 2016). These methods may achieve higher accuracy and precision, and therefore, the AUC of the AD patient z-scores may increase. Future studies should therefore also include deep learning–based approaches.

4.1 Conclusion

In this study, we aimed to answer two research questions: 1) to what extent are methods interchangeable, as long as the same method is being used for computing normative volume distributions and patient-specific volumes? and 2) can different methods be used for generating normative volume distributions and patient-specific volumes? Based on the absolute agreement results on the volume data of 988 non-demented subjects, we conclude that it is essential that the same method is used to generate normative volume distributions and patient-specific volumes. For most regions, the correlation was good (>0.75), indicating that volume differences between methods in ND subjects are mainly due to systematic differences. When the same method is being used for generating normative and patient data, we found that the agreement on the AD patient's position relative to the normative distribution (ICC-z) was high for the regions thalamus and putamen. Our results are encouraging as they indicate that the studied methods are interchangeable for these regions. For the regions hippocampus, amygdala, caudate nucleus and accumbens, and globus pallidus, not all method combinations showed a high ICC-z. Whether two methods are indeed interchangeable should be confirmed for the specific application and dataset of interest.

DATA AVAILABILITY STATEMENT

The datasets presented in this article are not readily available because of restrictions based on privacy regulations and informed consent of the participants. Requests should be directed toward the management team of the Rotterdam Study (secretariat.epi@erasmusmc.nl), which has a protocol for approving data requests.

ETHICS STATEMENT

The Rotterdam Study has been approved by the Medical Ethics Committee of the Erasmus MC (registration number MEC 02.1015) and by the Dutch Ministry of Health, Welfare and

Sport (Population Screening Act WBO, license number 1071272-159521-PG). The Rotterdam Study has been entered into the Netherlands National Trial Register (NTR; www.trialregister.nl) and into the WHO International Clinical Trials Registry Platform (ICTRP; www.who.int/ictrp/network/primary/en/) under shared catalogue number NTR6831. All participants provided written informed consent to participate in the study and to have their information obtained from treating physicians.

AUTHOR CONTRIBUTIONS

All authors designed the study. WN, MV, and SK provided supervision. FW, EB, NT, and CL provided methodology and assisted in performing data analyses. WH collected and analyzed the results. All authors interpreted the results and drafted the manuscript. All authors critically revised the manuscript for important intellectual content. WH had full access to all the data in the study and takes responsibility for the integrity of the data and the accuracy of the data analysis.

FUNDING

The research leading to these results has received funding from the European Union Seventh Framework Programme FP7/2007 - 2013, project VPH-DARE@IT (Grant Agreement No: 601055) and from the European Union's Horizon 2020 research and innovation programme, project EuroPOND (Grant Agreement No: 666992).

REFERENCES

Aljabar, P., Heckemann, R., Hammers, A., Hajnal, J., and Rueckert, D. (2007). "Classifier Selection Strategies for Label Fusion Using Large Atlas Databases," in *Medical Image Computing and Computer-Assisted Interventions - MICCAI 2007. Lecture Notes in Computer Science, Vol 4791*. Editors N. Ayache, S. Ourselin, and A. Maeder (Berlin, Heidelberg: Springer-Verlag), 523–531.

Ashburner, J., and Friston, K. J. (2005). Unified Segmentation. *NeuroImage* 26, 839–851. doi:10.1016/j.neuroimage.2005.02.018

Babalola, K., Petrovic, V., Cootes, T., Taylor, C., Twining, C., Williams, T., et al. (2007). "Automated Segmentation of The Caudate Nuclei Using Active Appearance Models," in *3D Segmentation In The Clinic: A Grand Challenge. MICCAI 2007* (Berlin, Heidelberg: Springer-Verlag), 57–64.

Babalola, K. O., Cootes, T. F., Twining, C. J., Petrovic, V., and Taylor, C. (2008a). "3d Brain Segmentation Using Active Appearance Models and Local Regressors," in *Medical Image Computing and Computer-Assisted Interventions - MICCAI 2008. Lecture Notes in Computer Science, Vol 5241*. Editors D. Metaxas, L. Axel, G. Fichtinger, and G. Székely (Berlin, Heidelberg: Springer-Verlag), 401–408. doi:10.1007/978-3-540-85988-8_48

Babalola, K. O., Patenaude, B., Aljabar, P., Schnabel, J., Kennedy, D., Crum, W., et al. (2008b). "Comparison and Evaluation of Segmentation Techniques for Subcortical Structures in Brain MRI," in *Medical Image Computing and Computer-Assisted Interventions - MICCAI 2008. Lecture Notes in Computer Science, Vol 5241*. Editors D. Metaxas, L. Axel, G. Fichtinger, and G. Székely (Berlin, Heidelberg: Springer-Verlag), 409–416. doi:10.1007/978-3-540-85988-8_49

Bao, S., and Chung, A. C. S. (2018). Multi-scale Structured CNN with Label Consistency for Brain MR Image Segmentation. *Computer Methods Biomech. Biomed. Eng. Imaging Visualization* 6, 113–117. doi:10.1080/21681163.2016.1182072

Boccardi, M., Bocchetta, M., Morency, F. C., Collins, D. L., Nishikawa, M., Ganzola, R., et al. (2015a). Training Labels for Hippocampal Segmentation Based on the EADC-ADNI Harmonized Hippocampal Protocol. *Alzheimers Dement.* 11 (2), 175–183.

Boccardi, M., Bocchetta, M., Ganzola, R., Robitaille, N., Redolfi, A., Duchesne, S., et al. (2015b). Operationalizing Protocol Differences for EADC-ADNI Manual Hippocampal Segmentation. *Alzheimers Dement.* 2 (11), 184–194.

Brewer, J. B. (2009). Fully-automated Volumetric MRI with Normative Ranges: Translation to Clinical Practice. *Behav. Neurol.* 21, 21–28. doi:10.1155/2009/616581

Bron, E. E., Steketee, R. M. E., Houston, G. C., Oliver, R. A., Achterberg, H. C., Loog, M., et al. (2014). Diagnostic Classification of Arterial Spin Labeling and Structural MRI in Presenile Early Stage Dementia. *Hum. Brain Mapp.* 35, 4916–4931. doi:10.1002/hbm.22522

Cardoso, M. J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., et al. (2015). Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion. *IEEE Trans. Med. Imaging* 34, 1976–1988. doi:10.1109/tmi.2015.2418298

Chupin, M., Hammers, A., Liu, R. S., Colliot, O., Burdett, J., Bardinet, E., et al. (2009). Automatic Segmentation of the hippocampus and the Amygdala Driven by Hybrid Constraints: Method and Validation. *NeuroImage* 46, 749–761. doi:10.1016/j.neuroimage.2009.02.013

Cole, T. J., and Green, P. J. (1991). Smoothing Reference Centile Curves: the LMS Method and Penalized Likelihood. *Stat. Med.* 11, 1305–1319. doi:10.1002/sim.4780111005

Convit, A., De Leon, M. J., Tarshish, C., De Santi, S., Tsui, W., Rusinek, H., et al. (1997). Specific Hippocampal Volume Reductions in Individuals at Risk for Alzheimer's Disease. *Neurobiol. Aging* 18, 131–138. doi:10.1016/s0197-4580(97)00001-8

Cootes, T. F., Taylor, C. J., Cooper, D. H., and Graham, J. (1992). "Training Models of Shape from Sets of Examples," in BMVC92, September, 1992 (London, UK: Springer), 9–18. doi:10.1007/978-1-4471-3201-1_2

Corso, J. J., Tu, Z., Yuille, A., and Toga, A. (2007). Segmentation of Sub-cortical Structures by the Graph-Shifts Algorithm. *Inf. Process. Med. Imaging* 20, 183–197. doi:10.1007/978-3-540-73273-0_16

de Brébisson, A., and Montana, G. (2015). "Deep Neural Networks for Anatomical Brain Segmentation," in 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW), June 7–12, 2015 (New York, USA: IEEE), 20–28.

de Onis, M., Onyango, A., Borghi, E., Siyam, A., and Pinol, A. (2006). *WHO Child Growth Standards: Length/height-For-Age, Weight-For-Age, Weight-For-Length, Weight-Forheight and Body Mass index-for-age: Methods and Development*. Geneva, Switzerland: Tech. rep., WHO Department of Health and Nutrition.

den Heijer, T., Geerlings, M. I., Hoebeek, F. E., Hofman, A., Koudstaal, P. J., and Breteler, M. M. B. (2006). Use of Hippocampal and Amygdalar Volumes on Magnetic Resonance Imaging to Predict Dementia in Cognitively Intact Elderly People. *Arch. Gen. Psychiatry* 63, 57–62. doi:10.1001/archpsyc.63.1.57

Fischl, B., Salat, D. H., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., et al. (2002). Whole Brain Segmentation. *Neuron* 33, 341–355. doi:10.1016/s0896-6273(02)00569-x

Gousias, I. S., Rueckert, D., Heckemann, R. A., Dyet, L. E., Boardman, J. P., Edwards, A. D., et al. (2008). Automatic Segmentation of Brain MRIs of 2-Year-Olds into 83 Regions of Interest. *Neuroimage* 40, 672–684. doi:10.1016/j.neuroimage.2007.11.034

Grimm, O., Pohlack, S., Cacciaglia, R., Winkelmann, T., Plichta, M. M., Demirakca, T., et al. (2015). Amygdalar and Hippocampal Volume: A Comparison between Manual Segmentation, Freesurfer and VBM. *J. Neurosci. Methods* 253, 254–261. doi:10.1016/j.jneumeth.2015.05.024

Hammers, A., Allom, R., Koepp, M. J., Free, S. L., Myers, R., Lemieux, L., et al. (2003). Three-dimensional Maximum Probability Atlas of the Human Brain,

with Particular Reference to the Temporal Lobe. *Hum. Brain Mapp.* 19, 224–247. doi:10.1002/hbm.10123

Heckemann, R. A., Hajnal, J. V., Aljabar, P., Rueckert, D., and Hammers, A. (2006). Automatic Anatomical Brain MRI Segmentation Combining Label Propagation and Decision Fusion. *NeuroImage* 33, 115–126. doi:10.1016/j.neuroimage.2006.05.061

Heckemann, R. A., Keihaninejad, S., Aljabar, P., Rueckert, D., Hajnal, J. V., Hammers, A., et al. (2010). Improving Intersubject Image Registration Using Tissue-Class Information Benefits Robustness and Accuracy of Multi-Atlas Based Anatomical Segmentation. *NeuroImage* 51, 221–227. doi:10.1016/j.neuroimage.2010.01.072

Heckemann, R. A., Ledig, C., Gray, K. R., Aljabar, P., Rueckert, D., Hajnal, J. V., et al. (2015). Brain Extraction Using Label Propagation and Group Agreement: Pincram. *PLoS One* 10, e0129211. doi:10.1371/journal.pone.0129211

Huizinga, W., Poot, D. H. J., Vernooij, M. W., Roshchupkin, G. V., Bron, E. E., Ikram, M. A., et al. (2018). A Spatio-Temporal Reference Model of the Aging Brain. *NeuroImage* 169, 11–22. doi:10.1016/j.neuroimage.2017.10.040

Ikram, M. A., van der Lugt, A., Niessen, W. J., Koudstaal, P. J., Krestin, G. P., Hofman, A., et al. (2015). The Rotterdam Scan Study: Design Update 2016 and Main Findings. *Eur. J. Epidemiol.* 30, 1299–1315. doi:10.1007/s10654-015-0105-7

Jack, C. R., Petersen, R. C., Xu, Y. C., O'Brien, P. C., Smith, G. E., Ivnik, R. J., et al. (1999). Prediction of Ad with Mri-Based Hippocampal Volume in Mild Cognitive Impairment. *Neurology* 52, 1397. doi:10.1212/wnl.52.7.1397

Ledig, C., Heckemann, R. A., Hammers, A., Lopez, J. C., Newcombe, V. F. J., Makropoulos, A., et al. (2015). Robust Whole-Brain Segmentation: Application to Traumatic Brain Injury. *Med. Image Anal.* 21, 40–58. doi:10.1016/j.media.2014.12.003

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *J. Cogn. Neurosci.* 19, 1498–1507. doi:10.1162/jocn.2007.19.9.1498

Marquand, A. F., Rezek, I., Buitelaar, J., and Beckmann, C. F. (2016). Understanding Heterogeneity in Clinical Cohorts Using Normative Models: Beyond Case-Control Studies. *Biol. Psychiatry* 80, 552–561. doi:10.1016/j.biopsych.2015.12.023

McGraw, K. O., and Wong, S. P. (1996). Forming Inferences about Some Intraclass Correlation Coefficients. *Psychol. Methods* 1, 30–46. doi:10.1037/1082-989x.1.1.30

Morey, R. A., Petty, C. M., Xu, Y., Pannu Hayes, J., Wagner, H. R., Lewis, D. V., et al. (2009). A Comparison of Automated Segmentation and Manual Tracing for Quantifying Hippocampal and Amygdala Volumes. *NeuroImage* 45, 855–866. doi:10.1016/j.neuroimage.2008.12.033

Morra, J. H., Tu, Z., Apostolova, L. G., Green, A. E., Avedissian, C., Madsen, S. K., et al. (2008). Validation of a Fully Automated 3D Hippocampal Segmentation Method Using Subjects with Alzheimer's Disease Mild Cognitive Impairment, and Elderly Controls. *NeuroImage* 43, 59–68. doi:10.1016/j.neuroimage.2008.07.003

Mukaka, M. M. (2012). Statistics Corner: A Guide to Appropriate Use of Correlation Coefficient in Medical Research. *Malawi Med. J.* 24, 69–71.

Murgasova, M., Dyet, L., Edwards, D., Rutherford, M., Hajnal, J. V., and Rueckert, D. (2006). "Segmentation of Brain MRI in Young Children," in *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2006. Lecture Notes in Computer Science, Vol 4190*. Editors R. Larsen, M. Nielsen, and J. Sporring (Berlin, Heidelberg: Springer). doi:10.1007/11866565_84

Murphy, S., Mohr, B., Fushimi, Y., Yamagata, H., and Poole, I. (2014). "Fast, Simple, Accurate Multi-Atlas Segmentation of the Brain," in *Biomedical Image Registration. WBIR 2014. Lecture Notes in Computer Science, Vol. 8545*. Editors S. Ourselin and M. Modat (Cham: Springer), 1–10. doi:10.1007/978-3-319-08554-8_1

Patenaude, B., Smith, S. M., Kennedy, D. N., and Jenkinson, M. (2011). A Bayesian Model of Shape and Appearance for Subcortical Brain Segmentation. *NeuroImage* 56, 907–922. doi:10.1016/j.neuroimage.2011.02.046

Perlaki, G., Horvath, R., Nagy, S. A., Bogner,.Doczi, P. T., Doczi, T., Janszky, J., et al. (2017). Comparison of Accuracy between FSL's FIRST and Freesurfer for Caudate Nucleus and Putamen Segmentation. *Sci. Rep.* 7, 2418. doi:10.1038/s41598-017-02584-5

Petersen, R. C., Aisen, P. S., Beckett, L. A., Donohue, M. C., Gamst, A. C., and Harvey, D. J. (2010). Alzheimer's Disease Neuroimaging Initiative (ADNI): Clinical Characterization. *NeuroI.* 3 (74), 201–209.

Scheltens, P., Fox, N., Barkhof, F., and De Carli, C. (2002). Structural Magnetic Resonance Imaging in the Practical Assessment of Dementia: beyond Exclusion. *Lancet Neurol.* 1, 13–21. doi:10.1016/s1474-4422(02)00002-9

Shakeri, M., Tsogkas, S., Ferrante, E., Lippe, S., Kadoury, S., Paragios, N., et al. (2016). "Sub-cortical Brain Structure Segmentation Using F-CNN's," in 2016 IEEE 13th International Symposium on Biomedical Imaging (ISBI), April 9–16, 2016 (New York, USA: IEEE), 269–272.

Smith, S. M. (2002). Fast Robust Automated Brain Extraction. *Hum. Brain Mapp.* 17, 143–155. doi:10.1002/hbm.10062

Tu, Z., Narr, K. L., Dollar, P., Dinov, I., Thompson, P. M., and Toga, A. W. (2008). Brain Anatomical Structure Segmentation by Hybrid Discriminative/generative Models. *IEEE Trans. Med. Imaging* 27, 495–508. doi:10.1109/TMI.2007.908121

Tustison, N. J., Avants, B. B., Cook, P. A., Yuanjie Zheng, Y., Egan, A., Yushkevich, P. A., et al. (2010). N4ITK: Improved N3 Bias Correction. *IEEE Trans. Med. Imaging* 29, 1310–1320. doi:10.1109/tmi.2010.2046908

van der Lijn, F., den Heijer, T., Breteler, M. M. B., and Niessen, W. J. (2008). Hippocampus Segmentation in MR Images Using Atlas Registration, Voxel Classification, and Graph Cuts. *NeuroImage* 43, 708–720. doi:10.1016/j.neuroimage.2008.07.058

Van Leemput, K., Maes, F., Vandermeulen, D., and Suetens, P. (1999). Automated Model-Based Tissue Classification of MR Images of the Brain. *IEEE Trans. Med. Imaging* 18, 897–908. doi:10.1109/42.811270

Wang, J., Vachet, C., Rumple, A., Gouttard, S., Ouziel, C., Perrot, E., et al. (2014). Multi-atlas Segmentation of Subcortical Brain Structures via the AutoSeg Software Pipeline. *Front. Neuroinform.* 8, 7. doi:10.3389/fninf.2014.00007

Wenzel, F., Meyer, C., Stehle, T., Peters, J., Siemonsen, S., Thaler, C., et al. (2018). Rapid Fully Automatic Segmentation of Subcortical Brain Structures by Shape-Constrained Surface Adaptation. *Med. Image Anal.* 46, 146–161. doi:10.1016/j.media.2018.03.001

Wolz, R., Aljabar, P., Hajnal, J. V., Hammers, A., and Rueckert, D. (2010). LEAP: Learning Embeddings for Atlas Propagation. *NeuroImage* 49, 1316–1325. doi:10.1016/j.neuroimage.2009.09.069

Yee, T. (2010). The VGAM Package for Categorical Data Analysis. *J. Stat. Softw.* 32, 1–34. doi:10.18637/jss.v032.i10

Yeo, I.-K., and Johnson, R. (2000). A New Family of Power Transformations to Improve Normality or Symmetry. *Biometrika* 87, 954–959. doi:10.1093/biomet/87.4.954

Zagorchev, L., Meyer, C., Stehle, T., Wenzel, F., Young, S., Peters, J., et al. (2015). Differences in Regional Brain Volumes Two Months and One Year after Mild Traumatic Brain Injury. *J. Neurotrauma* 33, 29–34. doi:10.1089/neu.2014.3831

Ziegler, G., Ridgway, G. R., Dahnke, R., and Gaser, C.for the Alzheimer's Disease Neuroimaging Initiative (2014). Individualized Gaussian Process-Based Prediction and Detection of Local and Global gray Matter Abnormalities in Elderly Subjects. *NeuroImage* 97, 333–348. doi:10.1016/j.neuroimage.2014.04.018

Conflict of Interest: Author WN is co-founder, scientific advisor, and shareholder of Quantib BV. Author FW is employed by Philips Research.

The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Big Data

ORIGINAL RESEARCH
published: 05 August 2021
doi: 10.3389/fdata.2021.662200

A Multi-Study Model-Based Evaluation of the Sequence of Imaging and Clinical Biomarker Changes in Huntington's Disease

Peter A. Wijeratne[1,2], Eileanoir B. Johnson[2], Sarah Gregory[2], Nellie Georgiou-Karistianis[3], Jane S. Paulsen[4], Rachael I. Scahill[2], Sarah J. Tabrizi[2] and Daniel C. Alexander[1]*

[1]Centre for Medical Image Computing, Department of Computer Science, University College London, London, United Kingdom, [2]Huntington's Disease Research Centre, Department of Neurodegenerative Disease, University College London, Queen Square Institute of Neurology, London, United Kingdom, [3]Monash Institute of Cognitive and Clinical Neurosciences, School of Psychological Sciences, Faculty of Nursing, Medicine, and Health Sciences, Monash University Clayton Campus, Clayton, VIC, Australia, [4]Departments of Neurology and Psychiatry, Carver College of Medicine, University of Iowa, Iowa City, IA, United States

Understanding the order and progression of change in biomarkers of neurodegeneration is essential to detect the effects of pharmacological interventions on these biomarkers. In Huntington's disease (HD), motor, cognitive and MRI biomarkers are currently used in clinical trials of drug efficacy. Here for the first time we use directly compare data from three large observational studies of HD (total $N = 532$) using a probabilistic event-based model (EBM) to characterise the order in which motor, cognitive and MRI biomarkers become abnormal. We also investigate the impact of the genetic cause of HD, cytosine-adenine-guanine (CAG) repeat length, on progression through these stages. We find that EBM uncovers a broadly consistent order of events across all three studies; that EBM stage reflects clinical stage; and that EBM stage is related to age and genetic burden. Our findings indicate that measures of subcortical and white matter volume become abnormal prior to clinical and cognitive biomarkers. Importantly, CAG repeat length has a large impact on the timing of onset of each stage and progression through the stages, with a longer repeat length resulting in earlier onset and faster progression. Our results can be used to help design clinical trials of treatments for Huntington's disease, influencing the choice of biomarkers and the recruitment of participants.

Keywords: huntington's disease, biomarkers, disease progression model, multi-study investigation, clinical staging

OPEN ACCESS

Edited by:
Enrico Capobianco, University of Miami, United States

Reviewed by:
Jessica A. Turner, Georgia State University, United States
Pekka Ruusuvuori, University of Turku, Finland

****Correspondence:***
Peter A. Wijeratne
p.wijeratne@ucl.ac.uk

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *31 January 2021*
Accepted: *07 July 2021*
Published: *05 August 2021*

Citation:
Wijeratne PA, Johnson EB, Gregory S, Georgiou-Karistianis N, Paulsen JS, Scahill RI, Tabrizi SJ and Alexander DC (2021) A Multi-Study Model-Based Evaluation of the Sequence of Imaging and Clinical Biomarker Changes in Huntington's Disease. Front. Big Data 4:662200. doi: 10.3389/fdata.2021.662200

INTRODUCTION

The development of disease modifying treatments for Huntington's disease (HD), a fatal neurodegenerative condition, has taken remarkable steps in recent years. There are a wide range of clinical trials attempting to validate a treatment for HD currently ongoing, including trials testing antisense oligonucleotide and micro RNA therapies (Rodrigues et al., 2020). As we move towards larger Phase III clinical trials, it is imperative that both patient recruitment and endpoint selection are targeted to ensure trials have high sensitivity to detect the efficacy of pharmacological interventions. In order to tailor cohorts and clinical trial endpoints for different therapeutic targets, we require a detailed understanding of candidate biomarkers in HD.

Onset of HD symptoms typically begins in mid-life, with individual genetic burden determining a large amount of variance in the timing of disease onset (Bates et al., 2015).

It is clear that imaging and fluid biomarkers are sensitive to disease-related change many years prior to symptom onset (Tabrizi et al., 2009; Byrne et al., 2018), although the exact timing and order of these changes is still being studied. Imaging biomarkers that measure atrophy in regional brain volume show some of the largest effect sizes in both pre-manifest HD (PreHD) and manifest HD compared to other biomarker candidates, particularly in subcortical structures (Tabrizi et al., 2012; Tabrizi et al., 2013). Clinical markers assessing motor symptoms and cognitive decline typically exhibit disease-related change later than imaging biomarkers, but are currently used as primary endpoints since they have a more direct relationship with the clinical benefit of a therapy. However, when moving into large phase III trials it is important to select endpoints that relate closely to the disease stage of the patients, and biomarkers that are likely to be the most sensitive to change during this time.

Disease progression models can reveal disease-related changes at the group and individual levels directly from observed data (Oxtoby and Alexander, 2017). Here we focus on the event-based model (EBM), which infers the order in which biomarkers become abnormal from cross-sectional data. We have previously applied the EBM in HD to reveal a sequence of regional brain volume changes in the TRACK-HD study, a large multi-site study of HD (Wijeratne et al., 2018). We demonstrated that three subcortical structures (the putamen, caudate and pallidum) were the first to become abnormal, followed by regions of the insula, CSF spaces, and amygdala. We have also applied the EBM to reveals the sequence of mixed biofluid, imaging and clinical changes in the HD-CSF study, a smaller single-site cohort study of HD (Byrne et al., 2018; Rodrigues et al., 2020).

However, these analyses were performed separately, and no direct comparison was made between studies to determine which features and findings were consistent. The analysis we present here is the first cross-study EBM analysis performed in HD (or any other disease), using data from the three largest imaging cohort studies in HD: TRACK-HD, PREDICT-HD and IMAGE-HD (Paulsen et al., 2008; Tabrizi et al., 2013; Poudel et al., 2015). We also add commonly used phenotypic cognitive and motor markers to the analysis to compare the stage at which these become abnormal across cohorts. Furthermore, we investigate the impact of genetic burden, as measured by cytosine-adenine-guanine (CAG) repeat length, on progression through the sequence of events. We therefore provide new information on the consistency of measurable imaging and clinical biomarker changes across differing study designs and individual-level genetic information, which has direct relevance to the design of multi-centre clinical trials in HD.

MATERIALS AND METHODS

Cohorts

Participants from the PREDICT-HD, TRACK-HD and IMAGE-HD studies with MRI data collected at three time-points (study baseline plus two follow-ups) on the same scanner were included in the study. All scans underwent visual quality control (QC) prior to inclusion, after which there were 284 participants from four centres in TRACK-HD; 171 participants from 20 centres in PREDICT-HD; and 77 participants from one centre in IMAGE-HD. We note that no participants underwent any disease modifying treatment during data collection. **Table 1** shows the demographic, clinical and cognitive data at baseline for all cohorts and groups. As noted previously (Wijeratne et al., 2020), there are differences between the groups in a number of criteria due to different recruitment strategies.

TRACK-HD Study

Data for TRACK-HD were collected at four centres; Leiden, London, Paris and Vancouver between 2008–2011 (Tabrizi et al., 2013). HD gene-carriers were recruited from HD clinics and were required to have a CAG of ≥40. At baseline, 123 controls, 120 PreHD participants and 123 HD participants were recruited. PreHD participants were required to have a burden of pathology score > 250 (calculated as [age x (CAG-35.5)] (Langbehn et al., 2004), and a UHDRS Total Motor Score (UHDRS-TMS) (Huntington Study Group, 1996) of less than five, indicating minor motor symptoms. Manifest HD participants were required to have a diagnostic confidence level (DCL) of four and a Total Functional Capacity of seven or more, as measured by the UHDRS TFC (Huntington Study Group, 1996). 3T T1-weighted scans were acquired from four scanners (two Siemens, two Philips). The parameters for Siemens were TR = 2200 ms, TE = 2.2 ms FOV = 28 cm, matrix size = 256 × 256, 208. For Philips TR = 7.7 ms, TE = 3.5 ms, FOV = 24 cm, matrix size = 242 × 224, 164. The acquisition was sagittal to cover the whole-brain. There was a slice thickness of 1mm, with no gap between slices. These acquisition protocols were validated for multi-site use. The study was approved by the local ethics committees, and written informed consent was obtained from each participant.

PREDICT-HD Study

Participants were recruited at 33 global centres, with most participants either PreHD or healthy controls (Paulsen et al., 2008). All participants were required to have had genetic testing (CAG ≥ 39 repeats) independent of the research study. PREDICT-HD recruited a total of 1,013 PreHD and 301 gene-negative controls between 2001 and 2012. Participants were excluded from the study at enrolment if there was a diagnosis of HD or evidence of an unstable illness, alcohol or drug abuse, a history of special education or central nervous system disease, a pacemaker or metallic implants, anti-psychotic medications prescribed in the previous 6 months or use of phenothiazine-derivative anti-emetic medication for 3 months or more. MRI acquisition parameters for the PREDICT-HD scanners included in this analysis are provided in (Wijeratne et al., 2020). The study was reviewed and approved by institutional review boards at all study and data processing sites. Participants underwent informed consent procedures and signed consents for both participation and to allow de-identified research data to be sent to collaborative institutions for analysis.

IMAGE-HD Study

IMAGE-HD was a single-centre study which recruited control, PreHD and manifest HD participants (Poudel et al., 2015). Gene

TABLE 1 | Demographic data for the PREDICT-HD, TRACK-HD and IMAGE-HD participants at baseline. Acronyms used: HC = healthy control, PRE = preHD, HD = manifest HD, P = PREDICT, T = TRACK, I = IMAGE. TIV = Total Intracranial volume, TMS = UHDRS Total Motor Score, DCL = Diagnostic Confidence Level, TFC = UHDRS Total Functional Capacity, DBS = Disease Burden Score, SDMT = Symbol Digit Modalities Test, SWRT = Stroop Word Reading Test. A value of "-" indicates that the data were not available.

	HC_P	HC_T	HC_I	PRE_P	PRE_T	PRE_I	HD_P	HD_T	HD_I
Age	45.1 ± 10.9	46.3 ± 10.4	43.3 ± 13.6	41.8 ± 11.0	41.2 ± 8.9	39.3 ± 8.2	46.5 ± 10.7	48.5 ± 9.3	53.0 ± 7.9
Sex	25:11	58:42	17:5	85:47	55:49	15:13	3:0	43:37	7:19
TIV (l)	2.07 ± 0.2	2.12 ± 0.22	2.14 ± 0.23	2.01 ± 0.19	2.15 ± 0.22	2.05 ± 0.19	1.89 ± 0.08	2.09 ± 0.19	2.15 ± 0.28
CAG	20.44 ± 3.5	—	—	42.4 ± 2.7	43.0 ± 2.3	42.7 ± 2.0	43.3 ± 4.2	43.8 ± 3.0	42.9 ± 2.1

carriers had a CAG of ≥ 39 repeats, and PreHD and manifest HD participants were allocated to each group based on their UHDRS-TMS, with those having a score of five or less included in the PreHD group and participants with a score of greater than five included in the manifest HD group. 108 participants were recruited at baseline, with imaging data available for 31 PreHD, 31 manifest HD and 29 control participants. Data were collected using a Siemens Magnetom Tim Trio 3T scanner with a 32 channel head coil. T1-weighted images were acquired with 192 slices, 0.9 mm slice thickness, 0.8 mm × 0.8 mm in-plane resolution, TE = 2.59 ms, TR = 1900 ms, flip angle = 9°. The study was approved by the Monash University and Melbourne Health Human Research Ethics Committees and informed written consent was obtained from each participant prior to testing in accord with the Helsinki Declaration.

Image Analysis

Structural MRI for each participant at baseline plus two follow-ups were analysed. T_1-weighted MRI data at 3T were used from the TRACK-HD and IMAGE-HD datasets, and at 1.5T (N = 136) and 3T (N = 35) from the PREDICT-HD dataset. For each dataset longitudinal registrations were performed on each participant via SPM12 using MATLAB version 2012b. The serial longitudinal registration pipeline was applied to all participants with data from three consecutive timepoints using default settings (Ashburner and Ridgway, 2012). This registration process resulted in an average scan for each participant along with Jacobean deformation maps. For every participant, the average scan was parcellated into 156 regions using the Geodesic Information Flows (GIF) software (Cardoso et al., 2015). Each region was then multiplied by Jacobian deformation maps to create a volumetric map for every region for every time-point.

Bilateral regions were combined across hemispheres as there is little evidence of hemispheric differences in HD atrophy (Minkova et al., 2017; Minkova et al., 2018). To enable interpretation of our results, we included a subset of biomarkers in this analysis based on HD pathology. These were the putamen, caudate, pallidum, lateral ventricles and global white matter. Total intracranial volume was calculated as the sum of cerebrospinal fluid (CSF), cortical gray matter, deep gray matter, and white matter (WM). All scans, registrations and segmentations underwent visual QC to remove scans due to poor quality defacing that was conducted on the MRI scans, or failures in registration and segmentation, or due to other pathology.

Other Variables

To facilitate further comparison among the three studies, three additional measures of phenotypic progression from the Unified Huntington's Disease Rating Scale (UHDRS) that were available from all three cohorts were included. The UHDRS Total Motor Score (TMS) was used to measure motor symptoms (Huntington Study Group, 1996). Two cognitive scores from the UHDRS—the symbol digit modalities test (SDMT) (Smith, 1991) and stroop word reading test (SWRT) (MacLeod, 1991)—were used as cognitive outcome measures, and CAG repeat length was used to quantify approximate lifetime genetic burden.

Covariates

All imaging and clinical variables were adjusted for covariates (age, sex, site) by regressing against the HC samples in each study separately. In addition, the imaging variables in the PREDICT-HD cohort were adjusted for field strength; the imaging variables in all studies were adjusted for total intracranial volume; and the clinical variables in all studies were adjusted for level of education.

Event-Based Model of Disease Progression

We use the event-based model (EBM; Fonteijn et al., 2012; Young et al., 2014) to infer the sequence of imaging and clinical biomarker changes in each study cohort. The EBM defines disease progression as an ordered sequence of abnormality events, which correspond to the transition of a biomarker from a healthy to abnormal state. To infer the most likely sequence of events across the population, the EBM fits healthy and abnormal distributions for each marker separately and makes the assumption of monotonic biomarker change. This assumption is reasonable for many biomarkers in progressive diseases, and in particular the imaging and clinical markers we use in this analysis.

Here we use non-parametric kernel density estimate mixture models (Firth et al., 2020) to fit the healthy and abnormal biomarker distributions, as they are more flexible than Gaussian mixture models. We fit these models to baseline data from the TRACK-HD cohort, as it provides the best sampling of HC (i.e., healthy) and HD (i.e., abnormal) groups (**Supplementary Figure S1** for the distributions and fits). We then use these mixture models to infer the most likely sequence, S, for each study separately using their respective baseline cohorts, and estimate the uncertainty in the sequence ordering using Markov chain Monte Carlo sampling of the model posterior. After inferring S, we can obtain a model-based disease stage by calculating

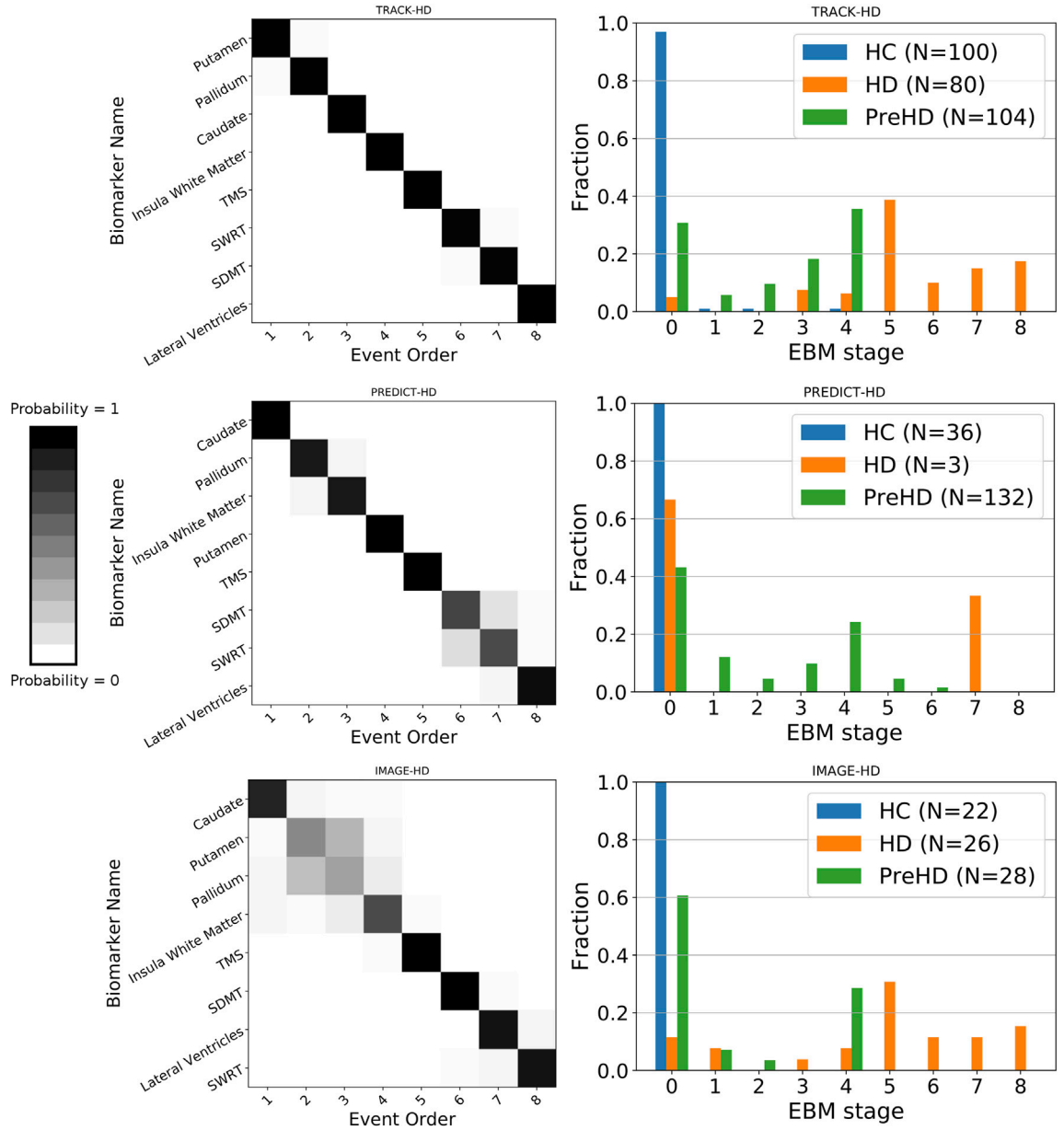


FIGURE 1 | Left column: positional variance diagrams showing the estimated order of regional brain volume and clinical marker abnormality events in PreHD and manifest HD patients at baseline, from the TRACK-HD, PREDICT-HD and IMAGE-HD cohorts separately. The heatmaps indicate the magnitude of the probability of the ordering; dark diagonal boxes indicate strong event ordering, and lighter indicate possible event permutations with strength proportional to the off-diagonal boxes. **Right column:** individual-level disease stage for each group in each cohort, predicted by the EBM sequence fit to each cohort separately.

the likelihood distribution over all stages for a given individual. We then take the maximum likelihood stage as the inferred individual-level stage.

Statistical Models of Progression

To interrogate the relationship between EBM stage and genetic burden, as specified by an individual's CAG repeat length, we build polynomial mixed effects regression models. Specifically, we regress the inferred individual-level EBM stage against age at each time-point (not just baseline) for each CAG group separately, with individual-level random intercepts. Instead of taking the maximum likelihood EBM stage, here we take the weighted average stage, as it accommodates uncertainty in the staging; as such, the stage is a continuous measure We construct both linear and quadratic mixed effects models for each CAG group,

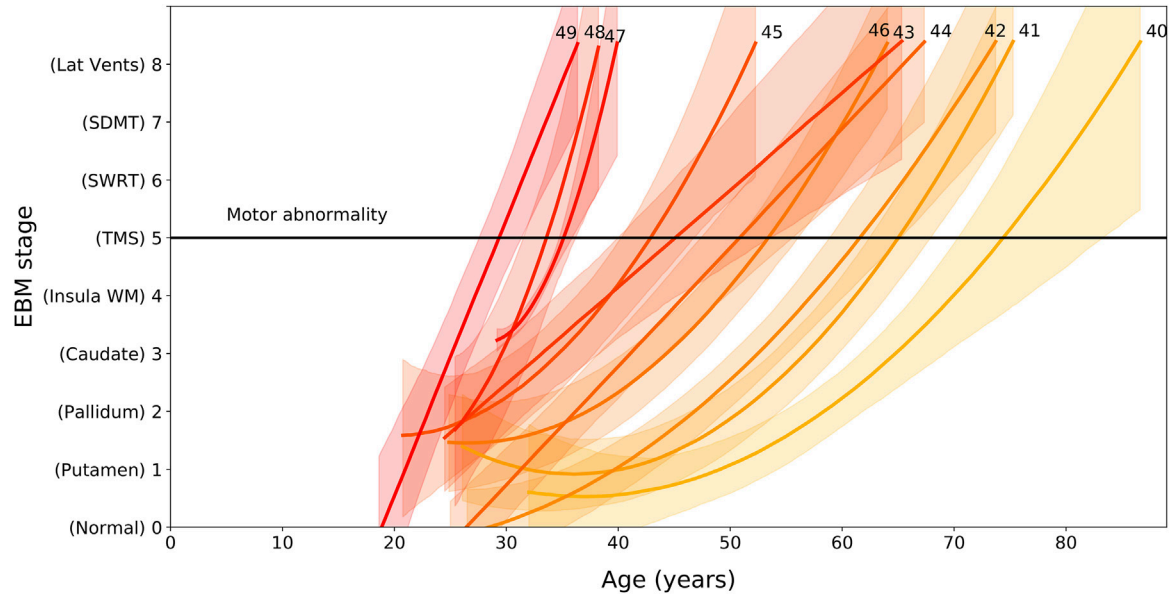


FIGURE 2 | EBM stage as a function of age and CAG repeat length, for PreHD participants with at least one follow-up across all years in the PREDICT-HD, TRACK-HD and IMAGE-HD cohorts. Polynomial mixed effects models are fit to each CAG group separately, which are coloured from low CAG repeat count in light yellow to high CAG repeat count in dark red, with the CAG repeat count denoted by integer values at the end of the curves. Stages are ordered along the vertical axis according to the ordering obtained by the EBM applied to the TRACK-HD cohort (**Figure 1**). The stage at which TMS becomes measurably abnormal is indicated by a black horizontal line (stage 5).

and select the model that provides the best fit as quantified by the size of the confidence intervals.

RESULTS

Event Sequences Are Consistent Across Studies

We find sequences of clinical and imaging events that are remarkably consistent across all three studies (**Figure 1** left column). For all three studies, the imaging biomarkers were placed before the clinical biomarkers with the exception of the lateral ventricles, which were positioned either last or second last for all cohorts. TMS was the fifth marker to become abnormal for all three cohorts, with SDMT and SWRT in variable positions after TMS. To quantify the similarity between event sequences, we calculated the Kendall's tau distance between each sequence separately, which returned values of 0.5 (TRACK-HD vs. PREDICT-HD, IMAGE-HD vs. PREDICT-HD) and 0.57 (TRACK-HD vs. IMAGE-HD), indicating positive correlations across all studies.

Event-Based Model Stage Reflects Clinical Stage

We find that EBM successfully stages individuals according to their clinical stage (HC, PreHD, or HD) in all three studies, when taking the maximum likelihood stage for each individual (**Figure 1** right column). As expected, the HC group is staged at or near zero, the PreHD group at intermediate stages, and the HD group across the later stages. The only exception is in the HD group in the PREDICT-HD cohort, where two of the three HD individuals are staged at zero; this is due to a combination of mismeasurement in the insula white matter and control-like clinical measurements for one individual, and mostly control-like volumetric and clinical measurements for the other individual.

Event-Based Model Stage Is Related to Age and Genetic Burden

We find that EBM stage and rate of progression depends on age and CAG length, with higher CAG lengths resulting in faster progression through the sequence (**Figure 2**). We can use the regression models shown **Figure 2** to calculate the average group-level age at each event as a function of CAG repeat length. We denote the onset of motor symptoms as equivalent to the event at which TMS becomes measurably abnormal (stage 5). Note that the dependency of motor onset on CAG is not smoothly monotonic (in particular CAG = 46); this is due to small sample sizes for these CAG lengths causing variability in the regression fits.

DISCUSSION

Here we applied a disease progression model, the EBM, to infer the patterns of change in brain and cognitive markers across

multiple cohorts in HD, and evaluated the consistency and genetic correlation of these changes. This is the first such cross-study model analysis in HD, and our findings suggest that the measurable changes in imaging and clinical volumes are largely independent of study protocols and cohort inclusion criteria. This has implications for large multi-centre clinical trials, which are necessary in HD due to its low prevalence, and suggests that the imaging and clinical biomarkers used in this analysis are suitable candidates for tracking disease progression.

Previously, we demonstrated that subcortical volumes become abnormal prior to other brain regions, which was supportive of the HD literature (Tabrizi et al., 2013; Byrne et al., 2018; Rodrigues et al., 2020). By applying the EBM to multiple cohorts we demonstrated that subcortical imaging biomarkers become abnormal prior to clinical markers. Across the three cohorts, the position of the caudate, pallidum, putamen and insula white matter varied in their position, but were consistently placed prior to clinical markers. The lateral ventricles were placed last (TRACK-HD, PREDICT-HD) or second to last (IMAGE-HD). The three non-imaging biomarkers are all ranked after the subcortical and white matter measures, with TMS first of these measures in all three cohorts. The differences in the relative positions of each imaging change across studies may be due to subtle between-sample variances related to cohort characteristics or imaging acquisitions, but by analyzing all data *via* the same imaging pipeline we can rule out the effects of different post-processing procedures. These results highlight the importance of using imaging biomarkers in clinical trials recruiting PreHD and early manifest HD participants, as clinical changes may not be sensitive enough to detect the pharmacological impacts of a therapy. Currently, the majority of clinical trials are focussed on manifest HD patients, but the end-goal of a number of therapeutic approaches is to treat PreHD individuals in order to delay or halt symptom onset. Trials for PreHD patients are unlikely to detect significant changes in clinical endpoints, and thus should also include imaging biomarkers as priority endpoints. The nature of these endpoints may vary dependent on the pharmaceutical mechanisms, but our results suggest that there are a variety of candidate regions available that change prior to clinical measures.

Importantly, we also demonstrate that the rate of progression through these stages is largely dependent on CAG repeat length, with wide variation seen in the age at which HD gene carriers with different CAG repeat lengths might be expected to pass through each stage. Our analysis of the link between CAG length, age and progression through the stages of our EBM suggest that those with shorter CAG repeat lengths undergo slower progression than those with longer CAG lengths. While this is supportive of previous work (Penney et al., 1997; Ruocco et al., 2008; Langbehn et al., 2011; Henley et al., 2012; Langbehn et al., 2019), **Figure 2** demonstrates how significantly this varies. Those with a CAG repeat length of 49 are expected to have abnormal sub-cortical and WM volumes by approximately 27 ± 2 years of age, while those with a CAG repeat length of 40 are estimated to be approximately 70 ± 5 years of age at the same stage. This large variability indicates that participants with larger CAG repeat lengths are expected to show faster progression during a clinical trial, and this should be considered during recruitment and treatment evaluation.

There are limitations to the analysis we present here. Firstly, we do not include biofluid biomarkers, such as neurofilament light, since these measures are only available for a limited selection of TRACK-HD data, and not at all for PREDICT-HD and IMAGE-HD. However, in previous work we demonstrate that these markers appear to be the first to show abnormalities in HD (Byrne et al., 2018; Rodrigues et al., 2020). In addition, we limited our investigation to a subset of available imaging biomarkers. This was done to aid interpretation, but different pharmacological mechanisms may require the consideration of other biomarkers not included here. Methodologically, we applied the basic cross-sectional EBM and hence were only able to recover the order of events, but not the time between them. Future work will use the recently developed temporal EBM (Wijeratne and Alexander, 2020) to properly leverage longitudinal data, allowing the time between events to be estimated. Finally, the basic EBM only considers a single sequence across the whole sample; it would be interesting to apply the subtyping version of the EBM (SuStaIn; Young et al., 2018) to investigate the possibility of multiple within-cohort subtypes.

By applying the EBM to multiple HD cohorts, we have confirmed that imaging biomarkers become abnormal prior to clinical and cognitive markers, and that there is large variation due to CAG repeat length in the age at which these markers become abnormal. By understanding both the sequence of changes in these markers and the correlation between the predicted individual-level stage and genetic burden, biomarkers can be more effectively selected for clinical trials in HD.

DATA AVAILABILITY STATEMENT

The datasets used in this study are available from the study authors co-ordinators (IMAGE-HD: NGK; PREDICT-HD: JSP; TRACK-HD; SJT) under a suitable data sharing agreement. Code used to run the Event-Based Model is available here: https://github.com/ucl-pond/kde_ebm.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by Monash University and Melbourne Health Human Research Ethics Committees. The patients/participants provided their written informed consent to participate in the TRACK-HD, PREDICT-HD, and IMAGE-HD studies.

AUTHOR CONTRIBUTIONS

PW and EJ contributed to conception and design of the study. PW performed the statistical analysis. EJ wrote the first draft of the manuscript. PW and EJ wrote sections of the manuscript. All

authors contributed to manuscript revision, read, and approved the submitted version.

FUNDING

PW was supported by funding from an MRC Skills Development Fellowship (MR/T027770/1). EJ, SG, RS, and ST were supported by funding from the Wellcome Trust (200181/Z/15/Z). The IMAGE-HD study was supported by the CHDI Foundation, Inc. (United States) (Grant Number A—3433) and the National Health and Medical Research Council (NHMRC) (AUS) (Grant Number: 606650). DA was supported by funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement number 666992, by the EPSRC under grant EP/M020533/1, and from the NIHR UCLH Biomedical Research Centre.

ACKNOWLEDGMENTS

We thank everyone involved in the TRACK-HD, PREDICT-HD, and IMAGE-HD studies.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fdata.2021.662200/full#supplementary-material

REFERENCES

Ashburner, J., and Ridgway, G. R. (2012). Symmetric Diffeomorphic Modeling of Longitudinal Structural MRI. *Front. Neurosci.* 6, 197. doi:10.3389/fnins.2012.00197

Bates, G. P., Dorsey, R., Gusella, J. F., Hayden, M. R., Kay, C., Leavitt, B. R., et al. (2015). Huntington Disease. *Nat. Rev. Dis. Primers* 1, 15005. doi:10.1038/nrdp.2015.5

Byrne, L. M., Rodrigues, F. B., Johnson, E. B., Wijeratne, P. A., De Vita, E., Alexander, D. C., et al. (2018). Evaluation of Mutant Huntingtin and Neurofilament Proteins as Potential Markers in Huntington's Disease. *Sci. Transl. Med.* 10, eaat7108. doi:10.1126/scitranslmed.aat7108

Cardoso, M. J., Modat, M., Wolz, R., Melbourne, A., Cash, D., Rueckert, D., et al. (2015). Geodesic Information Flows: Spatially-Variant Graphs and Their Application to Segmentation and Fusion. *IEEE Trans. Med. Imaging* 34, 1976–1988. doi:10.1109/TMI.2015.2418298

Firth, N. C., Primativo, S., Brotherhood, E., Young, A. L., Yong, K. X. X., Crutch, S. J., et al. (2020). Sequences of Cognitive Decline in Typical Alzheimer's Disease and Posterior Cortical Atrophy Estimated Using a Novel Event-Based Model of Disease Progression. *Alzheimers Demen.* 16 (7), 965–973. doi:10.1002/alz.12083

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., et al. (2012). An Event-Based Model for Disease Progression and its Application in Familial Alzheimer's Disease and Huntington's Disease. *NeuroImage* 60, 1880–1889. doi:10.1016/j.neuroimage.2012.01.062

Henley, S. M. D., Wild, E. J., Hobbs, N. Z., Scahill, R. I., Ridgway, G. R., Macmanus, D. G., et al. (2009). Relationship between CAG Repeat Length and Brain Volume in Premanifest and Early Huntington's Disease. *J. Neurol.* 256, 203–212. doi:10.1007/s00415-009-0052-x

Huntington Study Group(1996). Unified Huntington's Disease Rating Scale: Reliability and Consistency. Huntington Study Group. *Mov Disord.* 11, 136–142. doi:10.1002/mds.870110204

Langbehn, D., Brinkman, R., Falush, D., Paulsen, J., and Hayden, M.International Huntington's Disease Collaborative Group (2004). A New Model for Prediction of the Age of Onset and Penetrance for Huntington's Disease Based on CAG Length. *Clin. Genet.* 65, 267–277. doi:10.1111/j.1399-0004.2004.00241.x

Langbehn, D. R., Hayden, M. R., and Paulsen, J. S. (2010). CAG-repeat Length and the Age of Onset in Huntington Disease (HD): A Review and Validation Study of Statistical Approaches. *Am. J. Med. Genet.* 153B, 397–408. doi:10.1002/ajmg.b.30992

Langbehn, D. R., Stout, J. C., Gregory, S., Mills, J. A., Durr, A., Leavitt, B. R., et al. (2019). Association of CAG Repeats with Long-Term Progression in Huntington Disease. *JAMA Neurol.* 76 (11), 1375–1385. doi:10.1001/jamaneurol.2019.2368

MacLeod, C. M. (1991). Half a century of Research on the Stroop Effect: an Integrative Review. *Psychol. Bull.* 109, 163–203. doi:10.1037/0033-2909.109.2.163

Minkova, L., Gregory, S., Scahill, R. I., Abdulkadir, A., Kaller, C. P., Peter, J., et al. (2018). Cross-sectional and Longitudinal Voxel-Based Grey Matter Asymmetries in Huntington's Disease. *NeuroImage: Clin.* 17, 312–324. doi:10.1016/j.nicl.2017.10.023

Minkova, L., Habich, A., Peter, J., Kaller, C. P., Eickhoff, S. B., and Klöppel, S. (2017). Gray Matter Asymmetries in Aging and Neurodegeneration: A Review and Meta-Analysis. *Hum. Brain Mapp.* 38, 5890–5904. doi:10.1002/hbm.23772

Oxtoby, N. P., and Alexander, D. C. (2017). Imaging Plus X: Multimodal Models of Neurodegenerative Disease. *Curr. Opin. Neurol.* 30 (4), 371–379. doi:10.1097/WCO.0000000000000460

Paulsen, J. S., Langbehn, D. R., Stout, J. C., Aylward, E., Ross, C. A., Nance, M., et al. (2008). Detection of Huntington's Disease Decades before Diagnosis: the Predict-HD Study. *J. Neurol. Neurosurg. Psychiatry* 79, 874–880. doi:10.1136/jnnp.2007.128728

Penney, J. B., Vonsattel, J.-P., MacDonald, M. E., Gusella, J. F., and Myers, R. H. (1997). CAG Repeat Number Governs the Development Rate of Pathology in Huntington's Disease. *Ann. Neurol.* 41, 689–692. doi:10.1002/ana.410410521

Poudel, G. R., Stout, J. C., Domínguez D, J. F., Gray, M. A., Salmon, L., Churchyard, A., et al. (2015). Functional Changes during Working Memory in Huntington's Disease: 30-month Longitudinal Data from the IMAGE-HD Study. *Brain Struct. Funct.* 220, 501–512. doi:10.1007/s00429-013-0670-z

Rodrigues, F. B., Byrne, L. M., Tortelli, R., Johnson, E. B., Wijeratne, P. A., Arridge, M., et al. (2020). Mutant Huntingtin and Neurofilament Light Have Distinct Longitudinal Dynamics in Huntington's Disease. *Sci. Transl. Med.* 12 (574), eabc2888. doi:10.1126/scitranslmed.abc2888

Ruocco, H. H., Bonilha, L., Li, L. M., Lopes-Cendes, I., and Cendes, F. (2008). Longitudinal Analysis of Regional Grey Matter Loss in Huntington Disease: Effects of the Length of the Expanded Cag Repeat. *J. Neurol. Neurosurg. Psychiatry* 79, 130–135. doi:10.1136/jnnp.2007.116244

Smith, A. (1991). *Symbol Digit Modalities Test*. Los Angeles: Western Psychological Services.

Tabrizi, S. J., Langbehn, D. R., Leavitt, B. R., Roos, R. A., Durr, A., Craufurd, D., et al. (2009). Biological and Clinical Manifestations of Huntington's Disease in the Longitudinal TRACK-HD Study: Cross-Sectional Analysis of Baseline Data. *Lancet Neurol.* 8, 791–801. doi:10.1016/S1474-4422(09)70170-X

Tabrizi, S. J., Reilmann, R., Roos, R. A., Durr, A., Leavitt, B., Owen, G., et al. (2012). Potential Endpoints for Clinical Trials in Premanifest and Early Huntington's Disease in the TRACK-HD Study: Analysis of 24 Month Observational Data. *Lancet Neurol.* 11, 42–53. doi:10.1016/s1474-4422(11)70263-0

Tabrizi, S. J., Scahill, R. I., Owen, G., Durr, A., Leavitt, B. R., Roos, R. A., et al. (2013). Predictors of Phenotypic Progression and Disease Onset in Premanifest and Early-Stage Huntington's Disease in the TRACK-HD Study: Analysis of 36-month Observational Data. *Lancet Neurol.* 12 (7), 637–649. doi:10.1016/S1474-4422(13)70088-7

Wijeratne, P. A., and Alexander, D. C. (2020). *Learning transition times in event sequences: the Temporal Event-Based Hidden Markov Model of disease progression. Lecture Notes Comput. Sci.* 20210, 583–595. doi:10.1007/978-3-030-78191-0_45

Wijeratne, P. A., Johnson, E. B., Eshaghi, A., Aksman, L., Gregory, S., Johnson, H. J., et al. (2020). Robust Markers and Sample Sizes for Multicenter Trials of Huntington Disease. *Ann. Neurol.* 87 (5), 751–762. doi:10.1002/ana.25709

Wijeratne, P. A., Young, A. L., Oxtoby, N. P., Marinescu, R. V., Firth, N. C., Johnson, E. B., et al. (2018). An Image-Based Model of Brain Volume Biomarker Changes in Huntington's Disease. *Ann. Clin. Transl Neurol.* 5, 570–582. doi:10.1002/acn3.558

Young, A. L., Marinescu, R. V., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., et al. (2018). Uncovering the Heterogeneity and Temporal Complexity of Neurodegenerative Diseases with Subtype and Stage Inference. *Nat. Commun.* 9, 4273. doi:10.1038/s41467-018-05892-0

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A Data-Driven Model of Biomarker Changes in Sporadic Alzheimer's Disease. *Brain* 137, 2564–2577. doi:10.1093/brain/awu176

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 12 August 2021
doi: 10.3389/frai.2021.613261

Ordinal SuStaIn: Subtype and Stage Inference for Clinical Scores, Visual Ratings, and Other Ordinal Data

Alexandra L. Young[1,2,3], Jacob W. Vogel[4,5], Leon M. Aksman[6], Peter A. Wijeratne[2,3], Arman Eshaghi[3,7], Neil P. Oxtoby[2,3], Steven C. R. Williams[1] and Daniel C. Alexander[2,3] for the Alzheimer's Disease Neuroimaging Initiative*

[1]Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, United Kingdom, [2]Centre for Medical Image Computing, University College London, London, United Kingdom, [3]Department of Computer Science, University College London, London, United Kingdom, [4]Department of Psychiatry, University of Pennsylvania, Philadelphia, PA, Unites States, [5]Center for Biomedical Image Computing and Analytics, University of Pennsylvania, Philadelphia, PA, Unites States, [6]Stevens Neuroimaging and Informatics Institute, Keck School of Medicine, University of Southern California, Los Angeles, CA, Unites States, [7]Queen Square Multiple Sclerosis Centre, Department of Neuroinflammation, Faculty of Brain Sciences, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom

OPEN ACCESS

Edited by:
Raghvendra Mall, Qatar Computing Research Institute, Qatar

Reviewed by:
Ehsan Ullah, Qatar Computing Research Institute, Qatar
Murat Bilgel, National Institute on Aging (NIH), United States

***Correspondence:**
Alexandra L. Young
alexandra.young@kcl.ac.uk

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *06 April 2021*
Accepted: *20 July 2021*
Published: *12 August 2021*

Citation:
Young AL, Vogel JW, Aksman LM, Wijeratne PA, Eshaghi A, Oxtoby NP, Williams SCR and Alexander DC (2021) Ordinal SuStaIn: Subtype and Stage Inference for Clinical Scores, Visual Ratings, and Other Ordinal Data. Front. Artif. Intell. 4:613261. doi: 10.3389/frai.2021.613261

Subtype and Stage Inference (SuStaIn) is an unsupervised learning algorithm that uniquely enables the identification of subgroups of individuals with distinct pseudo-temporal disease progression patterns from cross-sectional datasets. SuStaIn has been used to identify data-driven subgroups and perform patient stratification in neurodegenerative diseases and in lung diseases from continuous biomarker measurements predominantly obtained from imaging. However, the SuStaIn algorithm is not currently applicable to discrete ordinal data, such as visual ratings of images, neuropathological ratings, and clinical and neuropsychological test scores, restricting the applicability of SuStaIn to a narrower range of settings. Here we propose 'Ordinal SuStaIn', an ordinal version of the SuStaIn algorithm that uses a scored events model of disease progression to enable the application of SuStaIn to ordinal data. We demonstrate the validity of Ordinal SuStaIn by benchmarking the performance of the algorithm on simulated data. We further demonstrate that Ordinal SuStaIn out-performs the existing continuous version of SuStaIn (Z-score SuStaIn) on discrete scored data, providing much more accurate subtype progression patterns, better subtyping and staging of individuals, and accurate uncertainty estimates. We then apply Ordinal SuStaIn to six different sub-scales of the Clinical Dementia Rating scale (CDR) using data from the Alzheimer's disease Neuroimaging Initiative (ADNI) study to identify individuals with distinct patterns of functional decline. Using data from 819 ADNI1 participants we identified three distinct CDR subtype progression patterns, which were independently verified using data from 790 ADNI2 participants. Our results provide insight into patterns of decline in daily activities in Alzheimer's disease and a mechanism for stratifying individuals into groups with difficulties in different domains. Ordinal SuStaIn is broadly applicable across different types of ratings data, including visual ratings from imaging, neuropathological ratings and clinical or behavioural ratings data.

Keywords: subtyping, staging, Alzheimer's disease, disease progression modelling, ordinal data

INTRODUCTION

Characterisation of disease progression patterns and heterogeneity among individuals can provide fundamental insights into the biology of a disease and is key to developing tools for patient stratification that can support precision medicine and healthcare. Disease progression models (Fonteijn et al., 2012; Jedynak et al., 2012; Donohue et al., 2014; Oxtoby et al., 2014; Young et al., 2014; Bilgel et al., 2016; Iturria-Medina et al., 2016; Schiratti, 2017; Koval et al., 2018; Li et al., 2018; Marinescu et al., 2019; Venkatraghavan et al., 2019; Firth et al., 2020) reconstruct the long-term temporal evolution of disease biomarkers from cross-sectional or short-term longitudinal data, enabling diagnosis, prognosis and stratification from biomarker measurements. In contrast to supervised machine learning techniques such as classification, which focus on a single disease stage, disease progression models infer fine-grained temporal patterns, providing the ability to generalise across disease stages and quantify disease trajectories in previously unseen detail. Disease progression models were primarily developed for use in Alzheimer's disease, where the decades-long disease process prevents the collection of long-term datasets that span the full disease time course, but they are increasingly being applied in other neurodegenerative diseases, such as Multiple Sclerosis (Eshaghi et al., 2018) and Huntington's disease (Wijeratne et al., 2018) and other long-term chronic conditions, such as respiratory diseases (Young et al., 2020b). However, the majority of disease progression modelling techniques rely on the assumption that all individuals follow a single common disease progression pattern, and so are unable to model disease subtypes which are prevalent in many diseases, and particularly in neurodegenerative diseases. Clustering identifies disease subgroups (Whitwell et al., 2009; Nettiksimmons et al., 2010, 2013, 2014; Noh et al., 2014; Racine et al., 2016; Zhang et al., 2016; Ferreira et al., 2020; Habes et al., 2020), providing new insights into disease heterogeneity, but lacks the ability to generalise across different disease stages, and so is unable to distinguish heterogeneity arising from differences in disease stage from heterogeneity due to the presence of disease subtypes.

The Subtype and Stage Inference (SuStaIn) algorithm (Young et al., 2018) allows disease progression modelling to be used in combination with clustering to identify subgroups of individuals with distinct disease trajectories. SuStaIn simultaneously clusters individuals into subgroups and characterises the trajectory that best defines each subgroup, thus capturing heterogeneity in both disease subtype and disease stage. The SuStaIn algorithm has been applied in a range of conditions including Alzheimer's disease (Young et al., 2018; Aksman et al., 2020; Garcia et al., 2020; Vogel et al., 2021), frontotemporal dementia (Young et al., 2018; Young et al., 2020a), Multiple Sclerosis (Eshaghi et al., 2020) and Chronic Obstructive Pulmonary disease (Young et al., 2020b). From a mathematical perspective any disease progression model can be used in combination with SuStaIn, but in practice some disease progression models may be unfeasibly computationally intensive. Two disease progression models have been used with SuStaIn to date: the event-based model (Fonteijn et al., 2012; Young et al., 2014; Firth et al., 2020) and the piecewise linear z-score model (Young et al., 2018). The event-based model describes disease progression as a series of events, where each event corresponds to a new biomarker becoming abnormal. The piecewise linear z-score model describes disease progression as a series of stages, with each stage corresponding to a biomarker linearly increasing to a new z-score relative to a control population. The advantage of each of these two models is that they are not too computationally intensive and work with purely cross-sectional data, enabling SuStaIn to perform stratification based on a single visit.

As is the case with most disease progression models, the disease progression models used in combination with SuStaIn to date are designed to take continuous biomarker measurements as input, for example those derived from blood or fluid samples or medical imaging. Whilst continuous measures offer fine-scaled resolution and so can provide high precision, discrete ordinal data, such as visual ratings of images, neuropathological ratings, and clinical and neuropsychological test scores can provide unique and complementary information. Clinical and cognitive test scores, for example, are widely collected in clinical settings and directly measure skills and symptoms that affect an individual's quality of life and reflect the severity of their disability. Meanwhile, neuropathological ratings offer direct measurement of disease pathologies, and thus can provide unique insights into the disease biology not possible with other techniques. Where imaging is used in a clinical setting, visual ratings of images are often already integrated into the clinical workflow, and thus can underpin diagnostic, prognostic and stratification tools that are more readily integrated into clinical practice. However, such measurements are not readily analysable by the majority of disease progression models, and neither of the disease progression models currently available for use with SuStaIn accommodate discrete ordinal data. The event-based model (Fonteijn et al., 2012; Young et al., 2014; Firth et al., 2020) doesn't model different severity levels, instead assuming each event is a transition from 'normal' to 'abnormal'. The piecewise linear z-score model (Young et al., 2018) doesn't allow for discrete data as it describes continuous biomarker trajectories with gaussian noise. There is a need for the development of disease progression modelling techniques that can be used on discrete ordinal data to enable a broader range of analyses to be carried out on these data types, in line with the techniques already available for continuous data.

Here we introduce the scored events model, allowing SuStaIn to be used with ordinal data. The scored events model describes disease progression as a series of events, where each event corresponds to a biomarker transitioning to a new score. We term the resulting algorithm 'Ordinal SuStaIn'. We verify the validity of Ordinal SuStaIn on simulated data, and that it outperforms the alternative option of using the existing piecewise linear z-score model ('Z-score SuStaIn') on ordinal data. We then demonstrate Ordinal SuStaIn by characterising heterogeneous trajectories of decline in subcategories of the Clinical Dementia Rating (CDR) scale.

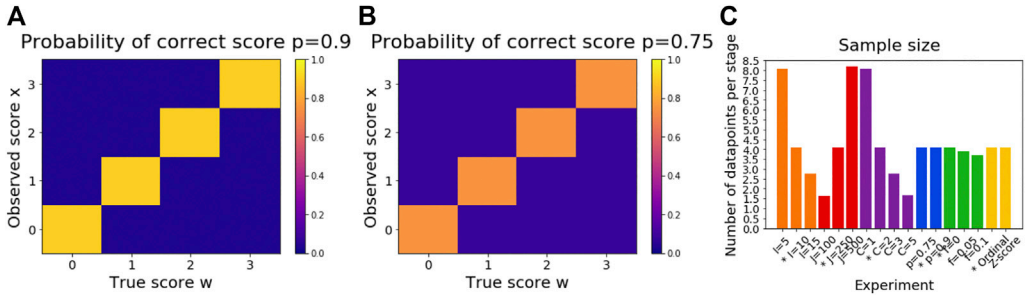


FIGURE 1 | Illustration of simulation settings. Subfigure **(A)** shows $P(x_{ij}|E_{iw})$ for the default proportion of correctly scored individuals $p = 0.9$. Subfigure **(B)** shows $P(x_{ij}|E_{iw})$ for the setting $p = 0.75$. $P(x_{ij}|E_{iw})$ can also be set to vary for each biomarker i and/or subject j. Subfigure **(C)** shows the expected number of datapoints for each stage of each subtype for each simulation setting.

MATERIALS AND METHODS

The Scored Events Model

We propose a scored events model to describe disease progression in Ordinal SuStaIn. The scored events model describes disease progression as a series of events, where each event corresponds to the transition of a biomarker to a new score. The occurrence of an event E_{iw} in biomarker i for score w is informed by the measurements x_{ij} of biomarker i in subject j, where each biomarker has its own set of scores $w_{ir} = w_{i1} \ldots w_{iW_i}$, and starts from a minimum score w_{i0}. The whole data set $X = \{x_{ij} | i = 1 \ldots I, j = 1 \ldots J\}$ is the set of measurements of each biomarker in each subject. The most likely ordering of the scored events is the sequence S that maximises the data likelihood

$$P\ (X|S) = \prod_{j=1}^{J}\left[\sum_{k=0}^{K} P(k) \prod_{i=1}^{I} P\left(x_{ij}|E_{iw}\right)\right],$$

where $w = s(i,k)$ is the score reached by biomarker i at stage k in the sequence S; at stage 0, w = w_{i0} for all biomarkers. The number of stages K is defined by the number of scored events included in the model, $K = 1 + \sum_{i=1}^{I} W_i$, i.e., the total number of scores included across all biomarkers. The form of the distribution $P(x_{ij}|E_{iw})$ is fully flexible and can be chosen by the user. The scored events model simply takes as input the probability each datapoint has each score: for each measurement x_{ij} of biomarker i in subject j the user specifies the probability $P(x_{ij}|E_{iw})$ that the 'true' score of measurement x_{ij} is E_{iw} for each score w as a matrix with dimensions $J \times W_i$ for each biomarker i. Here we use a categorical distribution (see **Figure 1** for a visualisation) where

$$P\left(x_{ij}|E_{iw}\right) = \begin{cases} p & \text{if } x_{ij} = w \\ \dfrac{1-p}{W_i} & \text{if } x_{ij} \neq w \end{cases}$$

thus p indicates the proportion of correctly scored individuals for each biomarker, and all other scores are assumed to be equally probable.

Ordinal SuStaIn

The SuStaIn algorithm (Young et al., 2018) assumes a dataset consists of c clusters of individuals (subtypes) that undergo a common disease progression pattern, S_c. Each individual is a sample of an unknown subtype c at an unknown stage k along the disease progression pattern for that subtype. SuStaIn simultaneously optimises subtype membership and subtype progression patterns (which describe the stages of the disease). SuStaIn fits an increasing number of clusters up to a user-defined maximum, using Markov Chain Monte Carlo (MCMC) sampling to obtain samples of the progression pattern for each subtype, providing an estimate of the posterior distribution of each subtype progression pattern. Information criterion can be used to choose the optimal number of clusters by evaluating the number of clusters that best balances accuracy and complexity, such as the Cross-Validation Information Criterion used in (Young et al., 2018). Our proposed Ordinal SuStaIn algorithm uses the scored events model detailed above to describe the evolution of biomarkers at different stages. To this end, Ordinal SuStaIn uses the same implementation of the SuStaIn algorithm as in (Young et al., 2018), but replaces the data likelihood $P(X|S_c)$ for each subtype c with that of the scored events model described above.

Simulated Data

We generated a series of simulated datasets to test the performance of Ordinal SuStaIn. To generate each dataset we randomly chose C subtype progression patterns, each described by a sequence S in which a set of scored events occur. We fixed the expected proportion π_c of individuals belonging to each subtype c to be

$$\pi_c = \frac{C - c + 2}{\sum_{c=1}^{C}[C - c + 2]},$$

or equivalently,

$$\pi_{C-c+1} = \frac{c + 1}{\sum_{c=1}^{C}[c + 1]},$$

such that the proportion of individuals in each subtype decreased from the most prevalent subtype $c = 1$ to the least prevalent

subtype $c = C$. We then randomly assigned $j = 1 \ldots J$ individuals to $c = 1 \ldots C$ subtypes and $k = 0 \ldots K$ stages, using a weighted random sampling of subtype membership c_j based on the proportion π_c of individuals belonging to each subtype, and a uniform random sampling of stage k_j. The set of expected biomarker scores for each individual $\overline{W_j} = \{w_i \forall\, i \; where \; w_i = s(i, k_j) \; if \; k_j > 0 \; and \; w_i = w_{i0} \; if \; k_j = 0\}$ was then evaluated and each individual's biomarker data $\overline{X_j}$ was then sampled according to the categorical distribution $P(x_{ij} = w_{ij}) = p$ and $P(x_{ij} \neq w_{ij}) = \frac{1-p}{W_i}$, such that $P(x_{ij}|E_{iw})$ follows the categorical distribution described above, as illustrated in **Figures 1A,B**. In our experiments we varied the number of biomarkers I, number of subjects J, number of subtypes C, proportion of correctly scored individuals p and a proportion of misdiagnosed individuals f who followed random subtype progression patterns not included in the simulated set of sequences S. By default we fixed the simulation settings to $I = 10$, $J = 250$, $C = 2$, $p = 0.9$, and $f = 0$, varying each setting in turn to test settings of $I = [5, 10, 15]$, $J = [100, 250, 500]$, $C = [1, 2, 3, 5]$, $p = [0.75, 0.9]$, and $f = [0, 0.05, 0.1]$. We fixed the number of scored events to $W_i = 3$ for all biomarkers i. Each experiment was performed three times for different randomly chosen subtype progression patterns and simulated datasets. The expected number of datapoints for each stage of each subtype varies across the different simulation settings, as illustrated in **Figure 1C**.

Comparison With Z-Score SuStaIn

We performed one further simulation in which we used the default settings to generate simulated data but used Z-score SuStaIn rather than Ordinal SuStaIn to estimate the subtype progression patterns and subtypes and stages of individuals. Z-score SuStaIn uses a piecewise linear z-score model of disease progression, which describes disease progression as a series of events, where each event corresponds to a biomarker reaching a new z-score relative to a control population. The data in the control population is assumed to be normally distributed and the data is z-scored using this control population such that the control population has a mean of 0 and standard deviation of 1. In the piecewise linear z-score model, the biomarkers start at 0 (at stage 0), accumulating linearly between z-score events (each of which corresponds to a new stage) and accumulate to a final maximum z-score (reached at the last stage). The z-score events and the maximum z-score are specified by the user. To apply Z-score SuStaIn we z-scored the data using a control population consisting of individuals assigned to stage 0 in each experiment. The z-score events in Z-score SuStaIn were set to be the same as those in Ordinal SuStaIn by z-score transforming the score corresponding to each scored event. The maximum z-score was set to be the same as the maximum score of the scored event model by z-score transforming the maximum scores.

Performance Evaluation: Progression Pattern Estimation

We estimated the most probable progression pattern $\overline{S_c}$ from the MCMC samples of the progression pattern by ordering the scored events according to their mean position in the sequence across samples. We measured the accuracy of the subtype progression patterns by calculating the average Kendall rank correlation τ (Kendall, 1945) between the most probable subtype progression patterns $\overline{S_c}$ estimated by SuStaIn and the ground truth subtype progression patterns $\widehat{S_c}$ in each simulation. This is computed as

$$\tau = \frac{P - Q}{\sqrt{(P + Q + T)(P + Q + U)}},$$

where P is the number of concordant pairs, Q is the number of discordant pairs, T is the number of ties in $\overline{S_c}$, and U is the number of ties in $\widehat{S_c}$. Correspondence between the ground truth and simulated subtypes was achieved by matching each simulated subtype progression pattern $\overline{S_c}$ with the most similar ground truth subtype progression pattern $\widehat{S_c}$. In nearly all experiments this was equivalent to matching the ground truth and simulated subtype progression patterns based on the proportion of individuals belonging to each subtype. The exception was for experiments with $C = 5$ subtypes in which the fraction would sometimes be swapped between subtypes of similar sizes, and so matching the subtype progression patterns based on their correspondence with the ground truth ensured that correspondence was achieved between subtypes of similar sizes. We estimated the confidence in the position assigned to each scored event by evaluating the proportion of MCMC samples in which each scored event appeared in the same position as in the most probable progression pattern. We evaluated the accuracy of the confidence estimate by determining whether the ground truth position of each scored event fell within the 95% confidence estimates output by SuStaIn. To do this we tested whether the ground truth position of each scored event was within two standard deviations of the estimated mean position of each scored event across MCMC samples.

Performance Evaluation: Subtyping and Staging

We computed the probability each individual belonged to each subtype and stage by computing the probability they belonged to each subtype (summed over stage) and stage (summed over subtype) for each MCMC sample and then averaging over MCMC samples, thus taking into account the uncertainty in the progression pattern of each subtype. We then assigned each individual to their most probable subtype and most probable stage. We estimated the confidence of the subtype and stage assignments by evaluating the probability of the subtype and stage that each individual had been assigned to. We evaluated the accuracy of the confidence estimates by determining whether the ground truth subtype and stage of each individual fell within the 95% confidence estimates output by SuStaIn. To do this we tested whether the ground truth subtype of each individual was assigned an average probability of at least 0.05, and whether the ground truth stage of each individual had a cumulative probability of more than 0.025 and less than 0.975.

Performance Evaluation: Number of Subtypes

When comparing the estimated subtype progression patterns and subtype and stage assignments with the ground truth, we fixed the number of subtypes to be the same as the ground truth number of

subtypes to enable a direct comparison. To give an indication of the accuracy of the number of subtypes estimated by SuStaIn we fitted up to $C + 1$ subtypes in each experiment. We then evaluated whether the 95% confidence intervals of the overall model likelihood (obtained from the MCMC samples of the model likelihood) for the ground truth number of subtypes C overlapped with the 95% confidence intervals of the overall model likelihood for one less $(C - 1)$ subtype and one more $(C + 1)$ subtype than the ground truth number of subtypes. We considered SuStaIn to underestimate the number of subtypes if the 95% confidence intervals of the C subtypes model likelihood overlapped the confidence intervals for $C - 1$ subtypes, or if the average model likelihood was greater for $C - 1$ subtypes. We considered SuStaIn to overestimate the number of subtypes if the average model likelihood was greater for $C + 1$ subtypes than C subtypes, and the 95% confidence intervals of the model likelihood for $C + 1$ subtypes didn't overlap the confidence intervals for C subtypes.

Alzheimer's Disease Neuroimaging Initiative Data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organisations, as a $60 million, 5 years public-private partnership. For up-to-date information, see http://www.adni-info.org. Written consent was obtained from all participants, and the study was approved by the Institutional Review Board at each participating institution.

CDR sub-scores (Hughes et al., 1982; Morris, 1993) from 819 participants in ADNI1 and 790 participants in ADNI2 were collated to obtain two independent datasets measuring sub-scores of memory, orientation, judgement, community, home and personal care. Each CDR sub-score can be assigned a score of 0 (no impairment), 0.5 (questionable impairment), 1 (mild impairment), 2 (moderate impairment) or 3 (severe impairment). Of the 819 ADNI1 participants, 229 were cognitively normal, 397 had mild cognitive impairment and 193 had a dementia diagnosis. Of the 790 ADNI2 participants, 293 were cognitively normal, 349 had mild cognitive impairment and 148 had a dementia diagnosis. We further collated follow-up CDR sub-scores at 6, 12, 18, 24 and 36 months follow-up visits to test the longitudinal consistency of the subtypes and stages assigned by Ordinal SuStaIn.

We ran Ordinal SuStaIn separately on baseline data from each of the ADNI1 and ADNI2 studies to obtain two independent estimates of CDR subtype progression patterns. We set the proportion p in $P(x_{ij}|E_{iw})$ with an accurate score to 0.75 for each sub-score, based on the inter-rater reliability of CDR scores in the literature (Schafer et al., 2004). None of the ADNI participants had a score of three on any CDR sub-scale and so this score was excluded from the scored events model. We selected the optimal number of subtypes by performing three-fold cross-validation in each dataset and evaluating the Cross-Validation Information Criterion (Gelman et al., 2014; Young et al., 2018).

Individuals were assigned to subtypes and stages at baseline and at follow-up visits using Ordinal SuStaIn, with the subtyping and staging being performed independently in each dataset (i.e., using the subtype progression patterns estimated from the baseline data in each dataset separately). Subtypes were considered to be longitudinally consistent between a pair of visits if both visits were labelled as the same subtype. Stages were considered to be longitudinally consistent between a pair of visits if the stage either remained the same or increased at the later of the two visits.

RESULTS

Simulated Data: Progression Pattern

Figure 2A shows the accuracy of SuStaIn for estimating subtype progression patterns under different simulation settings. In general, Ordinal SuStaIn gave a good accuracy across all settings, with a Kendall rank correlation between the estimated subtype progressions and the ground truth of >0.63 for all settings. When comparing Ordinal SuStaIn and Z-score SuStaIn under the default settings, the Kendall rank correlation using Z-score SuStaIn was only 0.33, compared to 0.95 for Ordinal SuStaIn. The confidence estimates of the position of each scored event provided by Ordinal SuStaIn (**Figure 2B**) gave a good indication of the true accuracy of the estimated progression patterns measured against the ground truth (**Figure 2B** reflects the trend in **Figure 2A**). Likewise, **Figure 2C** shows that the ground truth position of each scored event was generally within the 95% confidence intervals estimated by Ordinal SuStaIn for at least 95% of scored events (minimum of 94%, maximum of 100%). The confidence intervals obtained using Z-score SuStaIn were much less accurate with only 69% of the ground truth positions of the scored events being within the 95% confidence intervals estimated by Z-score SuStaIn.

The Kendall rank correlation between the estimated progression patterns and the ground truth varied substantially with different simulation settings. The Kendall rank correlation decreased substantially when the number of biomarkers was set to I = 15 compared with I = 5 and I = 10, when the number of subjects was set to J = 100 rather than J = 250 and J = 500, when the number of clusters was set to C = 3 or C = 5 rather than C = 1 or C = 2, and when the proportion of correctly scored individuals was set to $p = 0.75$ compared to $p = 0.9$. As shown in **Figure 3A**, increasing the number of biomarkers, decreasing the number of subjects and increasing the number of clusters all reduce the number of datapoints per subtype and stage combination, with this decrease in sample size correlating with the decrease in the accuracy of the progression pattern. **Figure 3A** also shows that decreasing the proportion p of individuals that are scored correctly from $p = 0.9$ to $p = 0.75$, which makes the data noisier, further decreases the accuracy of the estimated progression patterns in addition to the effect of sample size.

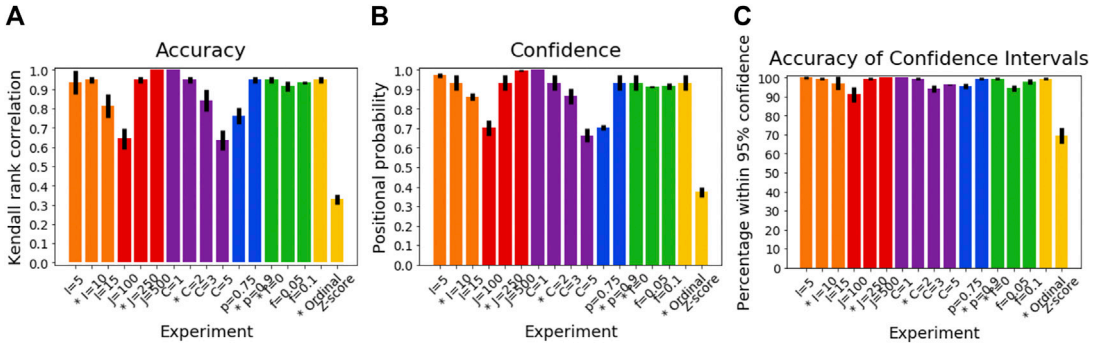


FIGURE 2 | Performance of SuStaIn for recovering progression patterns. **(A)** Accuracy of Ordinal SuStaIn for recovering the ground truth subtype progression patterns, **(B)** the confidence SuStaIn assigned to the estimated subtype progression patterns, and **(C)** the accuracy of the confidence intervals SuStaIn assigned to the estimated subtype progression patterns. The x-axis shows the experiments in which we varied the simulated number of biomarkers I (orange), number of subjects J (red), number of subtypes C (purple), proportion p with an accurate score (i.e., the categorical probability each test score is accurate; blue), the proportion f of misdiagnosis (i.e., the proportion of individuals that follow randomly chosen alternative progression patterns; green), and the choice of algorithm (either the proposed Ordinal SuStaIn algorithm or the existing Z-score SuStaIn algorithm). The default value for each simulation setting is indicated with an asterisk on the x-axis.

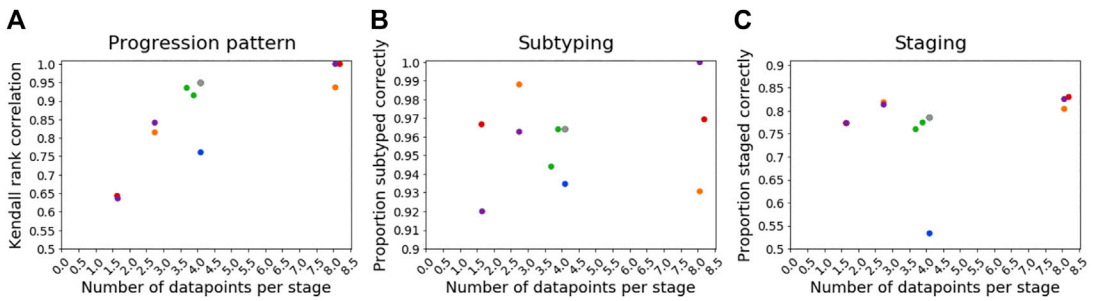


FIGURE 3 | Relationship between sample size and accuracy**.** Each subfigure shows a scatter plot comparing the expected number of datapoints per stage for each simulation and the accuracy of Ordinal SuStaIn for **(A)** estimating subtype progression patterns, **(B)** subtyping individuals, and **(C)** staging individuals. Each simulation setting is plotted using the same colours used in **Figures 2**, **4**, **5**, except the default setting, which is shown in grey. The simulation using Z-score SuStaIn was excluded from these figures.

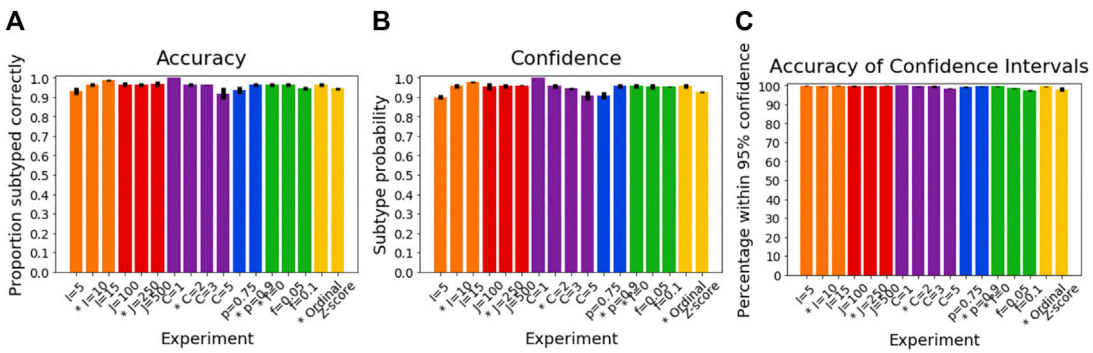


FIGURE 4 | Performance of SuStaIn for subtyping individuals. **(A)** Accuracy of Ordinal SuStaIn for recovering the ground truth subtypes of individuals, **(B)** the confidence SuStaIn assigned to the estimated subtypes, and **(C)** the accuracy of the confidence intervals SuStaIn assigned to the estimated subtypes. As in **Figure 2**, the x-axis shows the experiments in which we varied the simulated number of biomarkers I (orange), number of subjects J (red), number of subtypes C (purple), proportion p with an accurate score (i.e., the categorical probability each test score is accurate; blue), the proportion f of misdiagnosis (i.e., the proportion of individuals that follow randomly chosen alternative progression patterns; green), and the choice of algorithm (either the proposed Ordinal SuStaIn algorithm or the existing Z-score SuStaIn algorithm). The default value for each simulation setting is indicated with an asterisk on the x-axis.

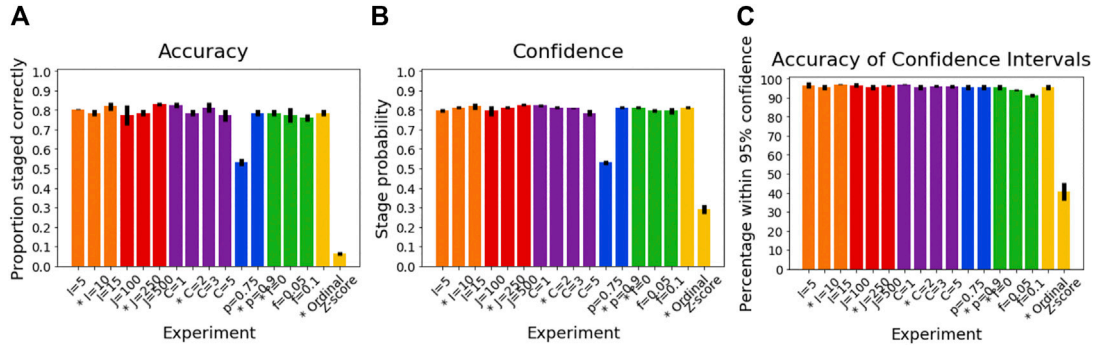


FIGURE 5 | Performance of SuStaIn for staging individuals. **(A)** Accuracy of Ordinal SuStaIn for recovering the ground truth stages of individuals, **(B)** the confidence SuStaIn assigned to the estimated stages, and **(C)** the accuracy of the confidence intervals SuStaIn assigned to the estimated stages. As in **Figures 2**, **3**, the x-axis shows the experiments in which we varied the simulated number of biomarkers I (orange), number of subjects J (red), number of subtypes C (purple), proportion p with an accurate score (i.e., the categorical probability each test score is accurate; blue), the proportion f of misdiagnosis (i.e., the proportion of individuals that follow randomly chosen alternative progression patterns; green), and the choice of algorithm (either the proposed Ordinal SuStaIn algorithm or the existing Z-score SuStaIn algorithm). The default value for each simulation setting is indicated with an asterisk on the x-axis.

Simulated Data: Subtyping

Figure 4A shows how the accuracy of SuStaIn for subtyping individuals varies with different simulation settings. SuStaIn was able to subtype individuals with high accuracy, with more than 92% of individuals being subtyped correctly for all simulation settings. **Figure 4B** shows that the confidence was a good reflection of the subtyping accuracy, following the same trend as **Figure 4A**. **Figure 4C** shows that all simulation settings gave 95% confidence intervals that were correct in at least 95% of subjects (minimum of 97%, maximum 100%). **Figure 3B** shows that the accuracy of the subtyping was not particularly related to the sample size. However, the sample size does remain reasonably large for each subtype across all simulation settings: the last subtype in the C = 5 experiment was the smallest, but still had an expected sample size of 25 subjects.

Simulated Data: Staging

Figure 5A shows the accuracy of the SuStaIn stages of individuals for different simulation settings. The SuStaIn stages were around 80% accurate for most simulation settings. There were two notable exceptions. The first was when the proportion of correctly scored individuals was set to $p = 0.75$, introducing more noise in the data and reducing the staging accuracy to 53%. The second was when Z-score SuStaIn was used rather than Ordinal SuStaIn, which staged only 6% of individuals correctly. **Figure 5B** shows that Z-score SuStaIn also has a lower confidence in the stages assigned to each individual, but that the stages are not within the 95% confidence interval estimated by Z-score SuStaIn, with only 40% of individual's stages falling within the 95% confidence interval. For all other settings the confidence assigned by SuStaIn was a good reflection of the accuracy of the stages (**Figure 5B** follows the same trend as **Figure 5A**), and the confidence intervals were a good reflection of the confidence in each individuals stage assignment (**Figure 5C**), with at least the expected 95% of individuals ground truth stages falling within the 95% confidence intervals estimated by SuStaIn (minimum of 91% and maximum of 97%). **Figure 3C** shows that the staging accuracy increases slightly with sample size, but that the effect of noisy data (reducing the proportion of correctly scored individuals from $p = 0.9$ to $p = 0.75$) is much greater. **Figure 6** shows the relationship between the ground truth stage and the stage assigned by Z-score SuStaIn. Z-score SuStaIn systematically underestimates the stage of each individual, as well as being less accurate than Ordinal SuStaIn.

Simulated Data: Number of Subtypes

The number of subtypes was estimated accurately for all simulation settings, except when a proportion of misdiagnosed individuals f were included, or when Z-score SuStaIn was used instead of Ordinal SuStaIn. For f = 0.05, SuStaIn over-estimated the number of subtypes in two out of three experiments and for f = 0.10, SuStaIn over-estimated the number of subtypes in all three experiments. Z-score SuStaIn over-estimated the number of subtypes in two out of three experiments.

Application to Clinical Dementia Rating Sub-scores

Figure 7 shows the subtype progression patterns estimated from applying Ordinal SuStaIn to CDR sub-scores in ADNI1 and ADNI2 separately. Three subtypes with distinct progression patterns were identified independently in each dataset, which we describe as 1) 'typical'—the most numerous group, with memory problems at early SuStaIn stages, followed by difficulties with orientation and judgement and problem solving, and then difficulties with home life and community affairs, 2) 'orientation-spared'—remaining relatively well-oriented until later SuStaIn stages, and 3)

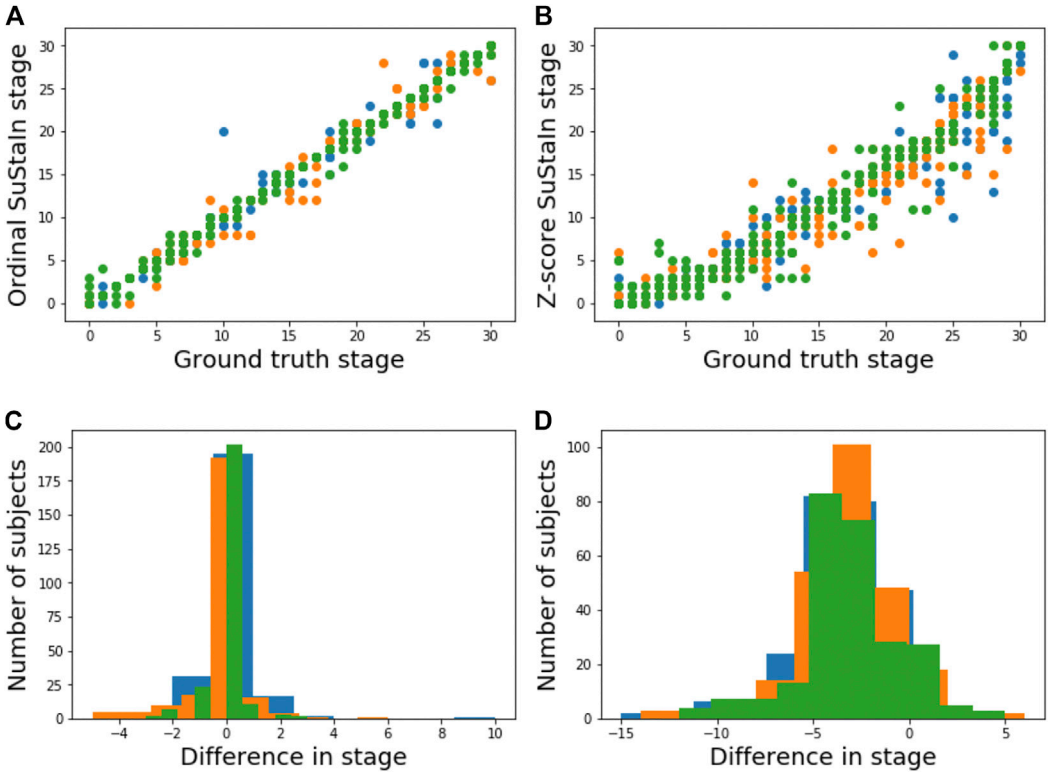


FIGURE 6 | Comparison of staging performance using Ordinal SuStaIn and Z-score SuStaIn**.** The top row shows scatter plots comparing the ground truth stage in simulation and the estimated SuStaIn stage obtained from **(A)** Ordinal SuStaIn and **(B)** Z-score SuStaIn across three simulations (shown in different colours) performed using the default settings. The bottom row shows histograms of the difference between the ground truth stage and the stage estimated by **(C)** Ordinal SuStaIn and **(D)** Z-score SuStaIn across the three simulations.

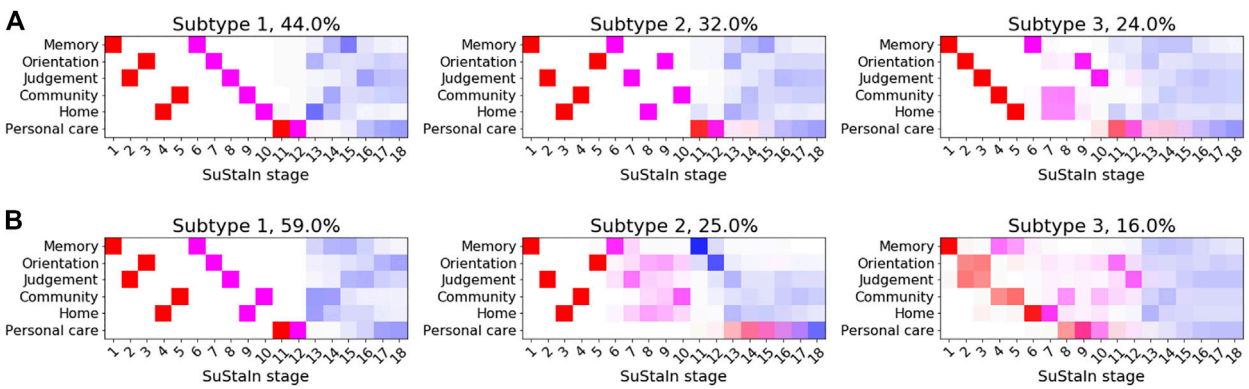


FIGURE 7 | Clinical subtypes of Alzheimer's disease based on CDR sub-scores. Subtypes of CDR ratings subtypes identified by applying Ordinal SuStaIn to **(A)** ADNI1 and **(B)** ADNI2. Each entry in the diagram represents the proportion of MCMC samples in which a particular scored event appears at a particular position along the progression pattern, with CDR = 0.5 shown in red, CDR = 1 in magenta and CDR = 2 in blue.

'outliers'—not following the 'typical' or 'orientation-spared' CDR sub-score progression pattern. The progression patterns were consistent between the two datasets, supporting the existence of three Alzheimer's subgroups with distinct clinical progression.

Subtyping and Staging Using Clinical Dementia Rating Sub-scores

Figures 8A,B show the distribution of the stages assigned to individuals by Ordinal SuStaIn at the baseline visit in ADNI1 and ADNI2. As expected, cognitively normal individuals had the

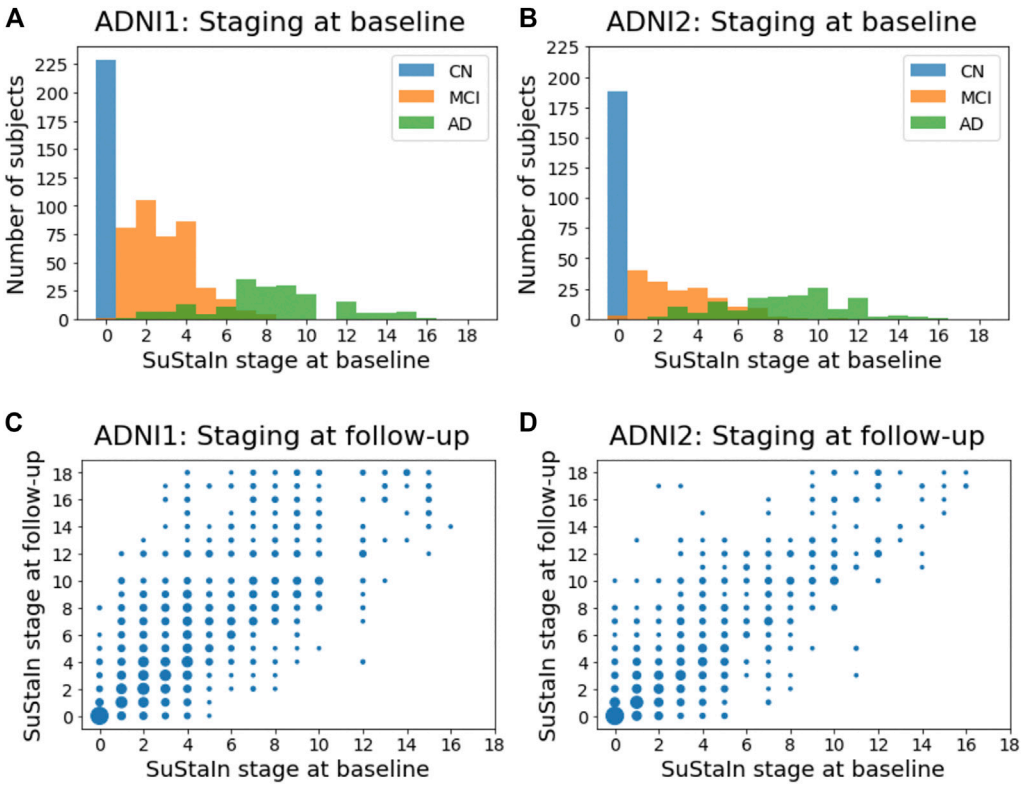


FIGURE 8 | Staging individuals using CDR sub-scores. The top row shows histograms of the SuStaIn stages of individuals in **(A)** ADNI1 and **(B)** ADNI2. The bottom row shows scatter plots comparing the SuStaIn stages of individuals at baseline and follow-up in **(C)** ADNI1 and **(D)** ADNI2. The size of each point represents the number of individuals. CN = cognitively normal; MCI = mild cognirive impairment; AD = Alzheimer's dsease.

lowest stages, followed by individuals with mild cognitive impairment, whilst individuals with Alzheimer's disease had the highest stages. There was a clear separation between cognitively normal and Alzheimer's disease subjects, with all cognitively normal subjects being assigned to stage 0 and all Alzheimer's disease subjects being assigned to stage 2 or above. **Figures 8C,D** compare the stages assigned to individuals by Ordinal SuStaIn at baseline and at follow-up. The follow-up visits were generally longitudinally consistent, i.e., at follow-up individuals either remained at the same stage or advanced in stage compared to baseline. In ADNI 1, 2113 of 2456 follow-up visits (86%) were longitudinally consistent, and in ADNI2, 1606 of 1885 follow-up visits (85%) were longitudinally consistent.

The subtypes assigned to individuals by Ordinal SuStaIn generally remained consistent at follow-up visits. Assigning individuals to subtypes using CDR scores is difficult as several of the stages are predicted to give the same CDR values across more than one subtype. For example, at stage 5 of all subtypes, CDR values are predicted to be 0 for the personal care rating and 0.5 for all the other sub-scales. Likewise, at stage 6 of all subtypes, CDR values are predicted to be 0 for the personal care rating, one for the memory score, and 0.5 for all the other sub-scales. Naively comparing each pair of visits that had CDR scores available at both visits (excluding individuals assigned to SuStaIn stage 0 at either visit and therefore unable to be subtyped), we found that the same subtype was assigned at both visits in 3,017 of 5,129 pairs of visits (59%) from ADNI1, and 2035 of 2,728 pairs of visits (75%) from ADNI2. Performing the same analysis but instead considering only individuals confidently assigned to subtypes (probability greater than or equal to 0.75), and thus removing individuals who were at stages where the subtypes are indistinguishable, we found that the same subtype was assigned at both visits in 143 of 190 pairs of visits (75%) from ADNI1, and in 157 of 169 pairs of visits (93%) from ADNI2.

DISCUSSION

In this study we developed Ordinal SuStaIn, an extension of the SuStaIn algorithm to allow SuStaIn to be used with discrete scored data. We demonstrated strong performance of Ordinal SuStaIn on simulated data and much better performance than using Z-score SuStaIn, which is designed for continuous data only. We applied Ordinal SuStaIn to CDR scores to identify three CDR subtypes that were

longitudinally consistent and replicable across independent data from ADNI1 and ADNI2.

The simulation results highlight the scenarios in which Ordinal SuStaIn performs best. In particular, the progression patterns are more accurately estimated when the average number of data points is more than three per stage. However, the confidence estimates still provided accurate information about the range of possible progression patterns and subtypes and stages of individuals, regardless of the simulation setting. The accuracy of the progression patterns also does not hugely impact on the subtyping and staging accuracy. In general, noise in the data has the largest effect of all settings, adversely affecting the ability to estimate the progression patterns and the stages of individuals. We also found that the number of subtypes is likely to be overestimated when a proportion of misdiagnosed individuals are included in the dataset. Misdiagnosed individuals are typically grouped into an outlier cluster with no distinct progression pattern.

We therefore propose the following guidelines for using Ordinal SuStaIn:

- Report the uncertainty in the progression patterns and the subtypes and stages of individuals by showing the positional variance diagrams or other visual representations of the uncertainty.
- In cases where there is low confidence take uncertainty into account in any subsequent analysis and reporting of results by clearly presenting the caveat that there is low confidence in a particular progression pattern.
- Small clusters with high uncertainty (proportion of individuals belonging to the cluster less than 10% and high uncertainty in the progression patterns illustrated by the positional variance diagrams) in the progression pattern should be reported as possibly being groups of outliers rather than subtypes.
- Where possible choose datasets and scored events to have an average of more than three data points per stage.
- Where possible choose biomarkers with a good signal to noise ratio.

Ordinal SuStaIn requires the user to input the probability $P(x_{ij}|E_{iw})$ that the 'true' score of measurement x_{ij} is E_{iw}. This allows complete flexibility in the probability distributions of the scores, which can vary by biomarker, score, and even by individual if desired. This allows the user to model, for example, some scores being difficult to distinguish from one another, whilst others are easily distinguished, or individualised confidence ratings for each score. $P(x_{ij}|E_{iw})$ would ideally be estimated by comparing assigned scores for each biomarker with a ground truth, in which the scorer is blinded to the ground truth score. In the absence of a ground truth, $P(x_{ij}|E_{iw})$ can be approximated by looking at test-retest reliability.

Z-score SuStaIn performed poorly at estimating progression patterns and stages of individuals for discrete data. Z-score SuStaIn uses a piecewise linear z-score model, which assumes that each biomarker transitions linearly between scores. This alters the expected value of each biomarker at each stage, with the majority of stages modelling biomarker values that don't exist in the data, leading to inaccuracy in the estimation of the subtype progression patterns and the stages of individuals. Z-score SuStaIn further assumes the errors on the data are normally distributed, which means that there are predicted to be more individuals with lower and higher scores than exist in the data. This causes a systematic overestimation of the stages of individuals at early stages and an underestimation of the stages of individuals at late stages. In this case the overall trend is to underestimate the stages of individuals as there are more stages representing scored events that have a positively skewed distribution than a negatively skewed distribution. Z-score SuStaIn also tends to overestimate the number of subtypes in the data to account for poor modelling of the subtype progression patterns.

Ordinal SuStaIn identified three clinical Alzheimer's subgroups with distinct patterns of decline in CDR sub-scores. The subgroups were independently identified in ADNI1 and ADNI2 and the subtypes and stages were longitudinally consistent at follow-up visits taken over a 3 year time frame. These subgroups may simply illustrate different cognitive trajectories experienced by individuals, there may be different underlying biological disease processes (Mukherjee et al., 2018), or there may be a proportion of individuals with other neurodegenerative diseases or atypical variants (Scheltens et al., 2017). Further work will be required to validate these subtypes in a wider range of clinical settings, and to test whether the subtypes correspond to distinct biological subgroups.

There are now three forms of SuStaIn that can be used in different settings: the new Ordinal SuStaIn algorithm proposed here, Z-score SuStaIn and Event-based SuStaIn. Ordinal SuStaIn uses a scored events model to describe discrete scored data, Z-score SuStaIn uses a piecewise linear z-score model to describe continuous data with normally distributed noise, and Event-based SuStaIn uses an event-based model to describe discrete or continuous biomarkers transitioning from normal to abnormal. Future work will explore whether it is possible to develop an integrated version of SuStaIn that can allow different types of data to be modelled simultaneously. Extensions to model subtypes conditioned on different variables would also be a valuable addition, for example modelling how genetics, demographics, lifestyle factors, multi-morbidity, and electronic health records are related to subtype assignment or how subtype assignment alters the probability of different outcomes, such as developing a particular condition or long-term health outcomes. Another important avenue for future work is incorporating longitudinal data to estimate the time between different stages.

All forms of the SuStaIn algorithm rely on several assumptions to infer temporal subtype progression patterns from cross-sectional data. One assumption is that biomarker trajectories increase monotonically with disease progression, enabling identifiability of the progression patterns. This monotonicity assumption is made at the population level rather than at an individual level, which enables SuStaIn to allow for reversion in disease stage; individuals who revert will be assigned a lower stage at follow-up than at baseline. In

future work there may be possibilities to relax this assumption by allowing a subset of biomarkers to be non-monotonic or incorporating longitudinal data to establish the time directionality. Another design choice in the SuStaIn algorithm is that the number of stages is fixed based on the number of biomarkers and scores. This simplifies the discrete optimisation procedure underlying SuStaIn by reducing the number of dimensions of the search space but can lead to redundant model complexity. Future versions will test whether it is possible to optimise the number of stages to enable more compact subtype progression patterns. However, under the current version of the SuStaIn algorithm, stages of a subtype progression pattern that are under-represented by samples can be identified by looking at the uncertainty in the positional variance diagrams. In addition, the model complexity can be reduced pre-emptively by limiting the number of features for small datasets, for example by using the rule of thumb described earlier of ensuring at least three subjects per stage. Another assumption that leads to redundancy in the subtype progression patterns is that each subtype progression pattern is unique; in fact, some subtypes may merge or split at some points in the progression. Future versions of the SuStaIn algorithm will explore whether merging and splitting of subtype progression patterns can be incorporated.

We proposed Ordinal SuStaIn, a variant of the SuStaIn algorithm for use with discrete scored data. We demonstrated that Ordinal SuStaIn out-performs available versions of SuStaIn in this setting and provides good performance in simulation. Ordinal SuStaIn is applicable to any discrete scored data. Here we applied Ordinal SuStaIn to CDR scores to reveal three distinct CDR subtypes in Alzheimer's disease, however Ordinal SuStaIn is readily applicable to visual ratings data, such as from neuropathology or imaging, other clinical, neuropsychological or behavioural scores, and across a wide range of conditions, including other neurodegenerative diseases and respiratory diseases.

DATA AVAILABILITY STATEMENT

The Alzheimer's disease data used in this study are publicly available from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). Source code for the Ordinal SuStaIn algorithm is available at https://github.com/ucl-pond/pySuStaIn together with a jupyter notebook enabling the simulations to be reproduced.

ETHICS STATEMENT

For data from the Alzheimer's Disease Neuroimaging Initiative, informed written consent was obtained by the study organisers from all participants or their designated caregiver(s) and the study was approved by the institutional ethical review board for each site.

AUTHOR CONTRIBUTIONS

AY developed the scored events model, wrote the programming code for Ordinal SuStaIn, analysed the data and wrote the manuscript. JV tested the performance of Ordinal SuStaIn whilst it was under development. LA and PW implemeted and optimised the python version of the SuStaIn algorithm. AY and DA designed the SuStaIn algorithm. All authors contributed to designing the experiments and reviewing and editing the manuscript.

FUNDING

AY is supported by a Skills Development Fellowship from the Medical Research Council (MR/T027800/1). JV is supported by the National Institute of Health (T32MH019112). PW is supported by a MRC Skills Development Fellowship (MR/T027770/1). AE was supported by an award from the International Progressive MS Alliance, award reference number PA-1603-08175. NO is a UKRI Future Leaders Fellow (MR/S03546X/1). NO and DA were supported by the NIHR UCLH Biomedical Research Centre and SW was supported by the NIHR Biomedical Research Centre at South London and Maudsley NHS Foundation Trust and King's College London. This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement No 666992. Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.;Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (http://www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

REFERENCES

Aksman, L. M., Oxtoby, N. P., Scelsi, M. A., Wijeratne, P. A., Young, A. L., Alves, I. L., et al. (2020). Tau-first subtype of Alzheimer's disease consistently identified across in vivo and post mortem studies. *bioRxiv*. doi:10.1101/2020.12.18.418004

Bilgel, M., Prince, J. L., Wong, D. F., Resnick, S. M., and Jedynak, B. M. (2016). A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *Neuroimage* 134, 658–670. doi:10.1016/j.neuroimage.2016.04.001

Donohue, M. C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R. G., Raman, R., Gamst, A. C., et al. (2014). Estimating long-term multivariate progression from short-term data. *Alzheimer's Demen.* 10, S400–S410. doi:10.1016/j.jalz.2013.10.003

Eshaghi, A., Marinescu, R. V., Young, A. L., Firth, N. C., Prados, F., Jorge Cardoso, M., et al. (2018). Progression of regional grey matter atrophy in multiple sclerosis. *Brain* 141, 1665–1677. doi:10.1093/brain/awy088

Eshaghi, A., Young, A., Wijertane, P., Prados, F., Arnold, D., Narayanan, S., et al. (2020). Identifying multiple sclerosis subtypes using machine learning and MRI data. *Nat. Commun.* 12, 2078.

Ferreira, D., Nordberg, A., and Westman, E. (2020). Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology* 94, 436–448. doi:10.1212/WNL.0000000000009058

Firth, N. C., Primativo, S., Brotherhood, E., Young, A. L., Yong, K. X. X., Crutch, S. J., et al. (2020). Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimer's Demen.* 16, 965–973. doi:10.1002/alz.12083

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., et al. (2012). An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* 60, 1880–1889. doi:10.1016/j.neuroimage.2012.01.062

Garcia, M. E., Young, A. L., Schott, J. M., and Alexander, D. C. (2020). *Multimodal modelling of the heterogeneity of Alzheimer's Disease*. Alzheimer's Association International Conference.

Gelman, A., Hwang, J., and Vehtari, A. (2014). Understanding predictive information criteria for Bayesian models. *Stat. Comput.* 24, 997–1016. doi:10.1007/s11222-013-9416-2

Habes, M., Grothe, M. J., Tunc, B., Mcmillan, C., Wolk, D. A., and Davatzikos, C. (2020). Disentangling Heterogeneity in Alzheimer's Disease and Related Dementias Using Data-Driven Methods. *Biol. Psychiatry* 88, 70–82. doi:10.1016/j.biopsych.2020.01.016

Hughes, C. P., Berg, L., Danziger, W., Coben, L. A., and Martin, R. L. (1982). A New Clinical Scale for the Staging of Dementia. *Br. J. Psychiatry* 140, 566–572. doi:10.1192/bjp.140.6.566

Iturria-Medina, Y., Sotero, R. C., Toussaint, P. J., Mateos-Perez, J. M., and Evans, A. C. (2016). Early role of vascular dysregulation on late-onset Alzheimer's disease based on multifactorial data-driven analysis. *Nat. Commun.* 7, 11934. doi:10.1038/ncomms11934

Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., et al. (2012). A computational neurodegenerative disease progression score: Method and results with the Alzheimer's disease neuroimaging initiative cohort. *Neuroimage* 63, 1478–1486. doi:10.1016/j.neuroimage.2012.07.059

Kendall, M. G. (1945). The Treatment of Ties in Ranking Problems. *Biometrika* 33, 239–251. doi:10.1093/biomet/33.3.239

Koval, I., Schiratti, J-B., Routier, A., Bacci, M., Colliot, O., Allassonnière, S., et al. (2018). Spatiotemporal Propagation of the Cortical Atrophy. *Popul. Individual Patterns* 9, 1–13. doi:10.3389/fneur.2018.00235

Li, D., Iddi, S., Thompson, W. K., Rafii, M. S., Aisen, P. S., Donohue, M. C., et al. (2018). Bayesian latent time joint mixed-effects model of progression in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 10, 657–668. doi:10.1016/j.dadm.2018.07.008

Marinescu, R. V., Eshaghi, A., Lorenzi, M., Young, A. L., Oxtoby, N. P., Garbarino, S., et al. (2019). DIVE: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. *Neuroimage* 192, 166–177. doi:10.1016/j.neuroimage.2019.02.053

Morris, J. C. (1993). The Clinical Dementia Rating (CDR) Current version and scoring rules. *Neurology* 43 (11), 2412–2414. doi:10.1212/wnl.43.11.2412-a

Mukherjee, S., Mez, J., Trittschuh, E., Saykin, A. J., Gibbons, L. E., Fardo, D. W., et al. (2018). Genetic data and cognitively-defined late-onset Alzheimer's disease subgroups. *Mol. Psychiatry*, 367615.

Nettiksimmons, J., Beckett, L., Schwarz, C., Carmichael, O., Fletcher, E., and Decarli, C. (2013). Subgroup of ADNI normal controls characterized by atrophy and cognitive decline associated with vascular damage. *Psychol. Aging* 28, 191–201. doi:10.1037/a0031063

Nettiksimmons, J., DeCarli, C., Landau, S., and Beckett, L. (2014). Biological heterogeneity in ADNI amnestic mild cognitive impairment. *Alzheimer's Demen.* 10, 511–521. doi:10.1016/j.jalz.2013.09.003

Nettiksimmons, J., Harvey, D., Brewer, J., Carmichael, O., DeCarli, C., Jack, C. R., et al. (2010). Subtypes based on cerebrospinal fluid and magnetic resonance imaging markers in normal elderly predict cognitive decline. *Neurobiol. Aging* 31, 1419–1428. doi:10.1016/j.neurobiolaging.2010.04.025

Noh, Y., Jeon, S., Lee, J. M., Seo, S. W., Kim, G. H., Cho, H., et al. (2014). Anatomical heterogeneity of Alzheimer disease: based on cortical thickness on MRIs. *Neurology* 83, 1936–1944. doi:10.1212/wnl.0000000000001003

Oxtoby, N. P., Young, A. L., Young, A. L., Fox, N. C., Daga, P., Cash, D. M., et al. (2014). "Learning imaging biomarker trajectories from noisy Alzheimer's disease data using a Bayesian multilevel model," in *Bayesian and Graphical Models for Biomedical Imaging* (Berlin, Germany: Springer International Publishing), 85–94. doi:10.1007/978-3-319-12289-2_8

Racine, A. M., Koscik, R. L., Berman, S. E., Nicholas, C. R., Clark, L. R., Okonkwo, O. C., et al. (2016). Biomarker clusters are differentially associated with longitudinal cognitive decline in late midlife. *Brain* 139, 2261–2274. doi:10.1093/brain/aww142

Scheltens, N. M. E., Tijms, B. M., Koene, T., Barkhof, F., Teunissen, C. E., Wolfsgruber, S., et al. (2017). Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts. *Alzheimer's Dement* 13(11), 1226–1236. doi:10.1016/j.jalz.2017.03.002

Schafer, K. A., Tractenberg, R., Sano, M., Mackell, J., Thomas, R., Gamst, A., et al. (2004). Reliability of monitoring the clinical dementia rating in multicenter clinical trials. *Alzheimer Dis. Assoc. Disord.*. 18: 219–222.

Schiratti, J. (2017). A Bayesian Mixed-Effects Model to Learn Trajectories of Changes from Repeated Manifold-Valued Observations. *J. Mach. Learn. Res.*. 18: 1–33.

Venkatraghavan, V., Bron, E. E., Niessen, W. J., and Klein, S. (2019). Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage* 186, 518–532. doi:10.1016/j.neuroimage.2018.11.024

Vogel, J. W., Young, A. L., Oxtoby, N. P., Smith, R., Ossenkoppele, R., Strandberg, O. T., et al. (2021). Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat. Med.* 27, 871–881. doi:10.1038/s41591-021-01309-6

Whitwell, J. L., Przybelski, S. A., Weigand, S. D., Ivnik, R. J., Vemuri, P., Gunter, J. L., et al. (2009). Distinct anatomical subtypes of the behavioural variant of frontotemporal dementia: A cluster analysis study. *Brain* 132, 2932–2946. doi:10.1093/brain/awp232

Wijeratne, P. A., Young, A. L., Oxtoby, N. P., Marinescu, R. V., Firth, N. C., Johnson, E. B., et al. (2018). An image-based model of brain volume biomarker changes in Huntington's disease. *Ann. Clin. Transl Neurol.* 5, 570–582. doi:10.1002/acn3.558

Young, A. L., Bocchetta, M., Cash, D. M., Convery, R. S., Moore, K. M., Neason, M. R., et al. (2020a). *Characterizing the Clinical Features and Atrophy Patterns of MAPT-Related Frontotemporal Dementia With Disease Progression Modeling. Neurology*. doi:10.1212/WNL.0000000000012410

Young, A. L., Bragman, F. J. S., Rangelov, B., Han, M. K., Galbán, C. J., Lynch, D. A., et al. (2020b). Disease Progression Modeling in Chronic Obstructive Pulmonary Disease. *Am. J. Respir. Crit. Care Med.* 201, 294–302. doi:10.1164/rccm.201908-1600OC

Young, A. L., Marinescu, R-V. V., Oxtoby, N. P., Bocchetta, M., Yong, K., Firth, N., et al. (2018). Uncovering the heterogeneity and temporal complexity of

neurodegenerative diseases with Subtype and Stage Inference. *Nat. Commun.* 9, 4273. doi:10.1038/s41467-018-05892-0

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain* 137, 2564–2577. doi:10.1093/brain/awu176

Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., and Yeo, B. T. T. (2016). Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc. Natl. Acad. Sci. USA* 113, E6535–E6544. doi:10.1073/pnas.1611073113

Conflict of Interest: AE has received speaker's honoraria from Biogen and At The Limits educational programme. He has received travel support from the National Multiple Sclerosis Society and honorarium from the Journal of Neurology, Neurosurgery and Psychiatry for Editorial Commentaries. He has received research grants from Biogen, and Roche. He serves on the editorial board of Neurology. AE and DA hold an equity stake in Queen Square Analytics.

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Big Data

METHODS
published: 20 August 2021
doi: 10.3389/fdata.2021.676168

Disease Modelling of Cognitive Outcomes and Biomarkers in the European Prevention of Alzheimer's Dementia Longitudinal Cohort

James Howlett[1†], *Steven M. Hill*[1†‡], *Craig W. Ritchie*[2] *and Brian D. M. Tom*[1*]

[1]*MRC Biostatistics Unit, University of Cambridge, Cambridge, United Kingdom,* [2]*Centre for Clinical Brain Sciences, University of Edinburgh, Edinburgh, United Kingdom*

OPEN ACCESS

Edited by:
Neil P. Oxtoby, University College London, United Kingdom

Reviewed by:
Vikram Venkatraghavan, Erasmus Medical Center, Netherlands
Cameron Shand, University College London, United Kingdom

***Correspondence:**
Brian D. M. Tom
brian.tom@mrc-bsu.cam.ac.uk

[†]*These authors have contributed equally to this work and share first authorship*

[‡]***Present Address:***
Steven M. Hill, Cancer Research UK Manchester Institute Cancer Biomarker Centre, University of Manchester, Manchester, United Kingdom

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Big Data

Received: *04 March 2021*
Accepted: *30 July 2021*
Published: *20 August 2021*

Citation:
Howlett J, Hill SM, Ritchie CW and Tom BDM (2021) Disease Modelling of Cognitive Outcomes and Biomarkers in the European Prevention of Alzheimer's Dementia Longitudinal Cohort. Front. Big Data 4:676168. doi: 10.3389/fdata.2021.676168

A key challenge for the secondary prevention of Alzheimer's dementia is the need to identify individuals early on in the disease process through sensitive cognitive tests and biomarkers. The European Prevention of Alzheimer's Dementia (EPAD) consortium recruited participants into a longitudinal cohort study with the aim of building a readiness cohort for a proof-of-concept clinical trial and also to generate a rich longitudinal data-set for disease modelling. Data have been collected on a wide range of measurements including cognitive outcomes, neuroimaging, cerebrospinal fluid biomarkers, genetics and other clinical and environmental risk factors, and are available for 1,828 eligible participants at baseline, 1,567 at 6 months, 1,188 at one-year follow-up, 383 at 2 years, and 89 participants at three-year follow-up visit. We novelly apply state-of-the-art longitudinal modelling and risk stratification approaches to these data in order to characterise disease progression and biological heterogeneity within the cohort. Specifically, we use longitudinal class-specific mixed effects models to characterise the different clinical disease trajectories and a semi-supervised Bayesian clustering approach to explore whether participants can be stratified into homogeneous subgroups that have different patterns of cognitive functioning evolution, while also having subgroup-specific profiles in terms of baseline biomarkers and longitudinal rate of change in biomarkers.

Keywords: Alzheimer's disease, biomarkers, cognitive functioning, disease modelling, European prevention of Alzheimer's dementia, latent class mixed models, precision medicine, Bayesian profile regression

1 INTRODUCTION

Alzheimer's disease (AD), the leading cause of dementia globally (Livingston et al., 2017), is characterised by synaptic dysfunction and neurodegeneration (e.g., neuronal loss), triggered by sequential accumulation of amyloid plaques and neurofibrillary tangles (aggregates of hyperphosphorylated tau proteins) (Braak and Braak, 1991). The exact ordering of the pathological cascade of events, leading to clinical symptoms of cognitive deterioration and dementia, has been actively researched over the last decade. Jack and colleagues (Jack et al., 2010; Jack et al., 2013) hypothesised that there is an underlying disease process and that the temporal ordering of changes in key biomarkers and their dynamics characterise the full spectrum of the disease throughout the different successive stages of pre-clinical, prodromal and dementia.

On the whole there have been few treatment successes (and none of these are disease-modifying) despite substantive investment in pharmacological compounds for Alzheimer's disease in symptomatic populations and early promise shown in pre-clinical studies (Gauthier et al., 2016; Winblad et al., 2016; Anderson et al., 2017). There may be a number of possible explanations for the many failures including inadequate drug dosages, incorrect treatment targets and inappropriate trial populations where the disease process is too far along to be amenable to treatment (Raket, 2020; Shi et al., 2020; Yiannopoulou and Papageorgiou, 2020). There is a consensus that the genesis of AD pathology occurs decades before the onset of dementia symptoms (Braak and Braak, 1997; Hardy and Selkoe, 2002; Jack et al., 2010; Bateman et al., 2012; Braak and Del Tredici, 2012; Jack et al., 2013). This thus presents an opportunity for early disease course modification before dementia onset and even prior to clinical symptoms. As such there is great interest—from both academia and industry—in accurately identifying groups of individuals with higher likelihood of progressing to AD dementia for natural history studies, early phase treatment trials and for participation in secondary prevention trials where, for example, they may have evidence of AD pathology through relevant biomarker abnormalities but no clinical evidence of symptoms of dementia (Ritchie et al., 2016; Watts, 2018).

Current proposals for defining an individual's probability for developing AD dementia or for modelling cognitive deterioration based on biomarkers and/or clinical symptoms have been focused on the stage of AD close to dementia onset. Various disease progression and sub-type approaches have been proposed and developed. These include survival and multi-state models for investigating transitions between disease states (Hubbard and Zhou, 2011; Vos et al., 2013; van den Hout, 2016; Wei and Kryscio, 2016; Robitaille et al., 2018; Zhang et al., 2019); mixed effects models (linear, generalized, non-linear) that incorporate subject-specific random effects and can be extended to handle latent time shifts, random change points, latent factors, processes and classes, hidden states, and multiple outcomes (Hall et al., 2000; Jedynak et al., 2012; Liu et al., 2013; Proust-Lima et al., 2013; Donohue et al., 2014; Samtani et al., 2014; Lai et al., 2016; Zhang et al., 2016; Geifman et al., 2018; Li et al., 2018; Wang et al., 2018; Lorenzi et al., 2019; Proust-Lima et al., 2019; Villeneuve et al., 2019; Younes et al., 2019; Bachman et al., 2020; Kulason et al., 2020; Raket, 2020; Segalas et al., 2020; Williams et al., 2020) and can be combined with models for event-history data (Marioni et al., 2014; Blanche et al., 2015; Proust-Lima et al., 2016; Rouanet et al., 2016; Li et al., 2017; Iddi et al., 2019; Li and Luo, 2019; Wu et al., 2020); event-based models which attempt to model the pathological cascade of events occurring as the disease develops and progresses through disease stages (Fonteijn et al., 2012; Young et al., 2014; Chen et al., 2016; Goyal et al., 2018; Oxtoby et al., 2018); and various clustering approaches for discovering risk stratification/disease progression groups and endotypes. For example, those based on hierarchical, partitioning and model-based clustering algorithms/methods (Dong et al., 2016; Racine et al., 2016; Dong et al., 2017; ten Kate et al., 2018; Young et al., 2018). Moreover, various machine learning and other statistical approaches have been proposed for both disease progression, prediction and subgroup identification in Alzheimer's disease (Fiot et al., 2014; Schmidt-Richberg et al., 2016; Cheng et al., 2017; Bhagwat et al., 2018; Khanna et al., 2018; de Jong et al., 2019; Martí-Juan et al., 2019; Brand et al., 2020; Golriz Khatami et al., 2020; Lei et al., 2020; Martí-Juan et al., 2020; Lin et al., 2021; Zhang et al., 2021).

However, in the earlier stages of disease, the development of disease models is far more challenging due to the relatively slow progression of the disease and clinical measures being insufficiently sensitive to detect such subtle changes. In order to develop disease models in the early stages when individuals do not have symptoms, or express only subjective complaints of cognitive decline or have only mild cognitive symptoms, it is necessary to undertake longitudinal follow-up of these individuals measuring reliable biomarkers of pathological changes alongside clinical outcomes. Ideally individuals would be followed-up over an extended period of time to ensure sufficient proportions make transitions through the various disease stages to dementia. Ultimately, these disease models would better inform patient selection into trials, improve understanding of AD progression in individuals and allow a more tailored approach to clinical management and targeting of disease modifying treatments to individuals (i.e., precision medicine) based on a range of biomarker modalities (e.g., neuroimaging, cerebrospinal fluid (CSF), blood), cognitive and clinical measures and risk factors.

Against this backdrop, the European Prevention of Alzheimer's Dementia (EPAD) consortium (Ritchie et al., 2016) was initiated as a large public-private partnership, and funded by the Innovative Medicines Initiative (IMI) Joint Undertaking. A total of 39 European organisations or "partners" were involved in the EPAD consortium. EPAD was developed as an interdisciplinary research initiative with an aim of improving the understanding of the early stages of Alzheimer's disease and delivering new preventative treatments.

The EPAD Longitudinal Cohort Study (LCS) was a prospective, multi-centre, pan-European study set up with the dual objectives of developing accurate longitudinal models over the entire course of Alzheimer's disease (AD) prior to the onset of dementia and creating a trial-ready cohort for potential recruitment into the EPAD Proof-of-Concept (PoC) Trial (Solomon et al., 2018). It was designed as a long-term observational study with recruitment from different types of existing parent cohorts (PCs) across Europe (e.g., population-based, memory clinics) and then, later on, more directly from clinical settings. It aimed to provide both a well-phenotyped population covering the full continuum of risk of subsequent AD dementia development and enough participants with particular profiles potentially eligible for an adaptive designed trial. This aim was achieved through monitoring of the evolving characteristics of the EPAD cohort and use of a flexible and dynamic approach to selection into the LCS that allowed over- and under-sampling by particular characteristics already available in the PCs. The other component of the EPAD programme, the EPAD PoC Trial, was designed to provide an environment for testing multiple interventions for the secondary prevention of AD dementia.

Using the data collected in the LCS on cognitive and clinical outcomes, biomarkers and risk factors, we aim to develop state-of-the-art models for disease progression and stratification which can be used 1) to inform selective recruitment and adaptation in clinical trials, 2) for longitudinal prediction and stratification, 3) for subgroup identification based on both baseline and longitudinal biomarker profiles and, ultimately, 4) to help improve treatment and clinical management decisions. We adopt a two-stage approach, where we first identify subpopulations/classes with different underlying, potentially AD-related cognitive/functional trajectory patterns (i.e., latent clinical phenotypes) over time after controlling for known exogenous risk factors (constitutional and genetic). These latent phenotypes are then jointly modelled with endogenous neuroimaging and CSF biomarkers to identify homogeneous subgroups/clusters based on biomarker profiles (i.e., neuropathological endotypes) that are linked to these trajectory patterns.

2 METHODS

2.1 Data

We performed all analyses on the V. IMI data release from the EPAD cohort (http://ep-ad.org/open-access-data/access/). Briefly, a total of 2,096 participants were screened and entered the cohort. Any participants who failed screening, had a baseline global clinical dementia rating (CDR) ≥1, or had a diagnosis of Alzheimer's dementia at baseline were excluded, leaving 1,828 eligible participants. Participants were aged at least 50 years old, with either a CDR global score of 0 (n = 1,313) or 0.5 (n = 498) (The CDR global scores for seventeen participants were missing.) Recruitment occurred across 31 centres from 10 different European countries. Follow-up visits were designed to occur at 6 months, 1 year and yearly thereafter. Unfortunately, the LCS closed at the end of the IMI-funding period and therefore the maximum number of visits was five. Of the 1,828 participants with a baseline visit, 1,567 attended the 6-months visit, 1,188 attended the 1-year visit and 396 and 89 attended the 2-years and 3-years visits respectively. Two hundred and fifty four participants only had a baseline visit, 389 had two visits (including five who had a baseline and 1-year visit but not 6-months), 791 had three visits (including 2 who had baseline, 1-year and 2-years visits but not a 6-months visit; the remaining attended the first three visits), 307 had four visits (including 2 who had baseline, 6-months, 2-years and 3-years visits but not a 1-year visit; the remaining had all visits up to 2 years) and 87 had five visits. We restrict our study to the 1,574 participants who had more than one visit.

The variables used in the models can be considered to belong to four domains: 1) outcomes, 2) baseline risk factors, 3) baseline biomarkers, and 4) longitudinal biomarkers.

Outcomes

The outcomes used were transformations of CDR sum of boxes (CDRSB) and Mini-Mental State Examination (MMSE) scores. To deal with floor and ceiling effects of CDRSB, a logistic transformation was applied to CDRSB as defined in **Eq. 1**:

$$tCDRSB = -\log\left(\frac{(CDRSB + 0.1)}{(18 - CDRSB + 0.1)}\right) \quad (1)$$

A normalising transformation was applied to MMSE values, converting MMSE from a 0–30 scale to nMMSE on a 0–100 scale to deal with curvilinearity (Philipps et al., 2014). CDRSB was scheduled to be collected at all visits but MMSE was not designed to be collected at the 6-months visit.

Baseline Risk Factors

Baseline risk factors included age, sex, education, family history of AD (first-degree relatives), and APOEϵ4 carrier status. Age is treated as a continuous variable. Sex, family history, and APOE are binary. Education was recorded in the LCS as years of formal education. However, as the values have different interpretations for different countries, years of education was converted to a three-category highest educational attainment level variable labelled 1, 2, and 3 on a country-specific basis (European Commission/EACEA/Eurydice, 2018). Level 1 is defined as up to secondary education, level 2 as beyond secondary education up to undergraduate ordinary degree, and level 3 as postgraduate studies.

Baseline Biomarkers

Baseline biomarkers included:

- the ratio of phosphorylated tau (pTau) to amyloid-beta 42 (Aβ), derived from CSF samples using the fully automated Roche Elecsys System in a single laboratory;
- volumetric imaging variables of the total of the left and right hippocampi and of the total of the four ventricles adjusting for head size by dividing by the pseudo total intracranial factor (HV and VV), processed by IXICO using the learning embeddings for atlas propagation (LEAP) method (Wolz et al., 2010);
- neurological radiological reads variables obtained through central assessment of magnetic resonance (MR) images by IXICO raters following a standardised, compliant and efficient workflow (Ritchie et al., 2020; ten Kate et al., 2018):
  - average of left and right medial temporal lobe atrophy (MTA);
  - Fazekas scale deep (FSD) and Fazekas scale periventricular (FSPV); and
  - five regional age-related white matter change (ARWMC) variables.

For EPAD participants, values of pTau/Aβ > 0.024 are here defined as CSF "AD positive" based on the biomarker cut-offs derived by Roche for EPAD using the methodology in (Hansson et al., 2018; Schindler et al., 2018), and reflect either decreased concentrations of Aβ (a marker of amyloidosis) or increased levels of pTau (a marker of neurofibrillary tangles). All radiological reads biomarkers were converted to binary variables <1 and ≥1, except for Fazekas scale deep which was dichotomised instead at 2. A score of 0 for all radiological reads

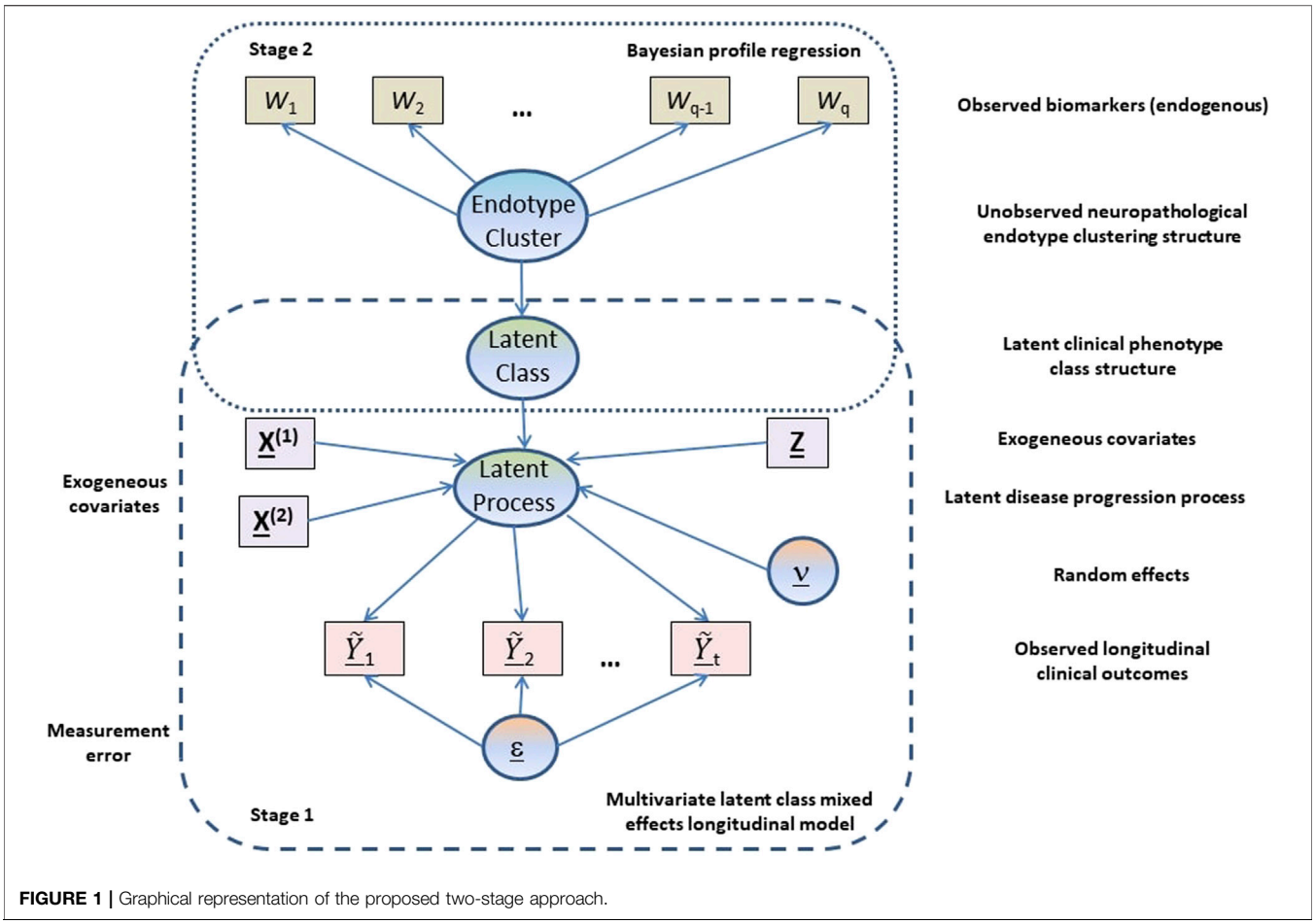


FIGURE 1 | Graphical representation of the proposed two-stage approach.

variables indicates no pathology and scores ≥1 (and ≥2 for Fazekas scale deep) indicate some pathology. A score of 0.5 in the average of left and right MTA is assumed to provide inconclusive evidence of pathology. A combined ARWMC variable was created that counted the number of age-related regions with evidence of white matter lesion cerebrovascular pathology, and a count ≥3 indicated that the majority of regions had signs of pathology.

Longitudinal Biomarkers

Longitudinal biomarkers considered were derived from the MR volumetric imaging variables of total hippocampal volume and total ventricular volume adjusting for head size. The processing of the longitudinal volumetric variables was also performed by IXICO using LEAP. The rates of change in the adjusted total hippocampal and ventricles volumes were calculated by dividing the difference between the last observed and baseline volumes by the time in study (in years) between the taking of the last and baseline volumes. These rate of change (i.e., annualised change) variables were used in our analyses to describe the longitudinal changes in biomarkers.

2.2 Statistical Methods

Our analysis is based on a two-stage approach (see **Figure 1**) where in the first stage a multivariate latent class linear mixed effects modelling approach is adopted to model the longitudinal cognitive and clinical outcomes adjusting for constitutional and genetic risk factors purported to be important in AD disease progression or related to selection into the EPAD LCS. From the multivariate latent class linear mixed effects model, latent clinical phenotypes corresponding to the latent classes are extracted to characterise the various mean trajectory profiles which individuals may follow over time. These latent phenotypes result from a hard assignment of individuals to specific latent classes based on their posterior probabilities of class membership. In the second stage, a probabilistic outcome-guided clustering approach based on Dirichlet process mixture modelling called Bayesian profile regression is applied to the latent phenotypes alongside the CSF and neuroimaging biomarkers. This aims to identify homogeneous clusters of participants with particular neuropathological endotypes characterised by biomarker profiles linked to clinical disease progression. Note that the latent phenotypes and endotypes are not meant to represent a grouping orthogonal to disease severity or stage, but reflect and characterise potential underlying processes and features that give rise to or are associated with disease severity or stage.

The specific statistical formulation of this two-stage modelling approach for disease progression, trajectory stratification and subgroup identification are outlined in the next two subsections. Missing response data are assumed to be missing at random

(MAR) for both stages, which allows valid inference using likelihood approaches.

2.2.1 Multivariate Latent Class Linear Mixed Effects Model

We used a multivariate latent class linear mixed effects model (MLCMM) to identify G mean profiles of trajectories corresponding to G latent classes or sub-populations of individuals (Lai et al., 2016; Proust-Lima et al., 2017). This model assumes that a latent process $\Lambda_i(t)$ generates the K longitudinal outcomes at time t, and this latent process is characterised by the mean trajectory profile corresponding to the latent class membership of individual i. Y_{ijk} is a measure of outcome k ($k = 1, \ldots, K$) for subject i ($i = 1, \ldots, N$) at measurement occasion j ($j = 1, \ldots, n_{ik}$), with associated time of outcome measurement from start of study t_{ijk}. Y_{ijk} is related to $\Lambda_i(t_{ijk})$ via an outcome-specific link function. Here for the purposes of this paper, we assume a linear transformation link function (others are possible) for the outcomes with outcome-specific parameters. That is, $\tilde{Y}_{ijk} = \frac{Y_{ijk} - \eta_{1k}}{\eta_{2k}}$, $k = 1, \ldots, K$. These transformations allow the transformed outcomes to be interpreted as noisy measurements of the underlying latent process with outcome-specific measurement errors.

The general formulation of the linear mixed effects part of our model given membership to latent class g is

$$\tilde{Y}_{ijk}|_{c_i=g} = \Lambda_i(t_{ijk})|_{c_i=g} + \epsilon_{ijk} \tag{2}$$

with

$$\Lambda_i(t_{ijk})|_{c_i=g} = X_{ijk}^{(1)T}\beta + X_{ijk}^{(2)T}\gamma_g + Z_{ijk}^T v_{ig}, \tag{3}$$

where c_i is the latent class variable, $X_{ijk}^{(1)}$ are the covariates associated with the class-independent fixed effects β and $X_{ijk}^{(2)}$ are the covariates associated with the class-specific fixed effects γ_g. Z_{ijk} are the covariates associated with the class-specific random effects v_{ig}, which are from a zero-mean multivariate normal with variance-covariance matrix $\omega_g B$, where B is left unspecified and ω_g is a positive proportionality factor (with $\omega_G = 1$ to ensure identifiability). The measurement errors $\{\epsilon_{ijk}\}$ are assumed to be independent Gaussian random variables with mean 0 and outcome-specific variances σ_k^2 ($k = 1, \ldots, K$).

The latent variable c_i equals g when subject i belongs to latent class g. To complete the specification of our multivariate latent class mixed model, the probability of individual i belonging to class g is described by the multinomial logistic submodel without covariates given by **Eq. 4**:

$$\pi_{ig} = P(c_i = g) = \frac{e^{\xi_{0g}}}{\sum_{l=1}^{G} e^{\xi_{0l}}}, \tag{4}$$

where ξ_{0g} is the intercept parameter for class g. Extension of this latent class membership submodel to include covariates is straightforward. The full MLCMM is fitted using maximum likelihood estimation within R (R Core Team, 2017) using the multlcmm function in the lcmm package (Proust-Lima et al., 2017).

In our application, we included the logistic transformed CDRSB, tCDRSB, and normalised MMSE, nMMSE, as outcomes ($K = 2$) in our MLCMM formulation. For both these outcomes and the latent process, a higher value indicates less cognitive/functional impairment (i.e., better cognitive functioning). We used time in study as the time scale and allowed class-specific fixed intercepts and slopes (time in study effects). As maximum follow-up in the EPAD study population was 3 years and 4 months and the majority of subjects had two or three visits, we considered only linear trends in an individual's underlying disease process. The baseline risk factors described in Section 2.1 were introduced into **Eq. 3** with associated class-independent fixed effects. We included only class-specific random intercepts into the latent process model, which are introduced to induce correlation across the longitudinal observations of an outcome for an individual and to better align participants in terms of where they fall on the disease time scale. The variance of the random intercept for the reference class is not estimated by the model and is set to be 1. The best choice of the number of latent classes was made using the Bayesian Information Criterion (BIC) and the relative entropy.

All observations with either a recorded CDRSB or MMSE were considered for inclusion in the model provided that individuals had 2 or more visits. These corresponded to 4,795 visits on 1,574 participants. Of which, there were 3,228 visits with both CDRSB and MMSE present, 1,558 visits with only CDRSB present, and nine visits with only MMSE present. Of the 1,574 individuals, 86 had five observation-visits, 305, 789, 384 and 10 had 4, 3, 2 and 1 observation-visits with either CDRSB or MMSE or both present respectively. However, 31 individuals had missing APOEϵ4 carrier status information and were excluded. This thus resulted in 1,543 individuals be included in the MLCMM analysis.

2.2.2 Bayesian Profile Regression

Bayesian profile regression (Molitor et al., 2010) is a non-parametric outcome-guided clustering approach that links an outcome variable to covariates via cluster membership. Here, it was applied to identify $G^\star$ clusters of participants, with each cluster characterised by particular clinical disease progression phenotypes (latent classes from the MLCMM analysis) and a particular CSF/neuroimaging biomarker profile. These clusters can be interepreted as corresponding to different neuropathological endotypes.

Bayesian profile regression uses a Dirichlet process mixture model (DPMM), which can be regarded as the limit of a finite mixture model as the number of components goes to infinity. That is, for observed data D_i for subject i, we have the following DPMM likelihood:

$$p(D_i|\pi^\star, \Theta) = \sum_{h=1}^{\infty} p(c_i^\star = h|\pi^\star)\mathrm{p}(\mathrm{D_i}|\mathrm{c_i^\star} = \mathrm{h}, \Theta) \tag{5}$$

$$= \sum_{h=1}^{\infty} \pi_h^\star f(D_i|\Theta_h), \tag{6}$$

where $c_i^\star \in \mathbb{Z}^+$ denotes latent cluster membership, $\pi^\star = (\pi_1^\star, \pi_2^\star, \ldots)^{\mathrm{T}}$ are mixture component (cluster) weights and $\Theta^T = (\Theta_1^T, \Theta_2^T, \ldots)$ are component-specific parameters for the mixture component densities, indexed by $h \in \mathbb{Z}^+$.

In addition to covariates W_i for subject i, Bayesian profile regression models an outcome Y_i^* that also informs the clustering and is assumed to be conditionally independent of the covariates given cluster assignment c_i^*. Furthermore, covariates can be a mix of discrete and continuous, $W_i^T = (W_i^{(\mathrm{d})T}, W_i^{(\mathrm{c})T})$, with discrete

covariates $W_i^{(\mathrm{d})}$ and continuous covariates $W_i^{(\mathrm{c})}$ also assumed to be conditionally independent given c_i^*. We therefore have observed data $D_i^T = (Y_i^*, W_i^{(\mathrm{d})T}, W_i^{(\mathrm{c})T})$ for subject i and **Eq. 6** becomes

$$p\left(Y_i^*, W_i^{(\mathrm{d})}, W_i^{(\mathrm{c})} | \pi^\star, \Theta\right) = \sum_{h=1}^{\infty} \pi_h^\star f\left(Y_i^* | \Theta_h^{(o)}\right) f\left(W_i^{(\mathrm{d})} | \Theta_h^{(\mathrm{d})}\right) f\left(W_i^{(\mathrm{c})} | \Theta_h^{(\mathrm{c})}\right), \tag{7}$$

where $\Theta_h^T = (\Theta_h^{(o)T}, \Theta_h^{(\mathrm{d})T}, \Theta_h^{(\mathrm{c})T})$ are the component-specific parameters for the outcome, discrete covariate and continuous covariate densities respectively.

The stick-breaking construction of the DPMM (Sethuraman, 1994) is used within Bayesian profile regression which gives the following formulation for the prior on the mixture weights: $\pi_1^\star = V_1$ and $\pi_h^\star = V_h \prod_{l<h}(1 - V_l)$ for $h \geq 2$ with $V_h \overset{iid}{\sim} \mathrm{Beta}(1, \alpha)$. The concentration hyperparameter α, which itself has a gamma prior distribution, affects the mixture weight distribution and implicitly informs the number of non-empty clusters. One of the key desirable properties of a DPMM approach to clustering is the removal of the need to pre-specify the number of clusters. Prior distributions are also placed on the component-specific parameters Θ_h and Markov chain Monte Carlo (MCMC) is used to fit the resulting profile regression model (see (Liverani et al., 2015) for details of the prior distributions and for computational aspects of the MCMC).

In our application, the outcome variable for each subject is the latent class predicted from the MLCMM analysis, i.e. $Y_i^* = \hat{c}_i$. This is treated as a categorical variable with cluster-dependent parameters: $Y_i^* | \Theta_h^{(o)} \sim \mathrm{Cat}(\theta_{h,1}^{(o)}, \theta_{h,2}^{(o)}, \ldots, \theta_{h,\hat{G}}^{(o)})$, where $\hat{G}$ is the estimated number of latent classes in the MLCMM. The covariates used in the model are the baseline and longitudinal biomarkers described in Section 2.1. In particular, we included five binary baseline covariates (pTau/Aβ, MTA, FSD, FSPV, ARWMC combined) for each subject, each independently taking a Bernoulli distribution given cluster assigmment: $W_{i,q}^{(\mathrm{d})} | \Theta_{h,q}^{(\mathrm{d})} \sim \mathrm{Bern}(\theta_{h,q}^{(\mathrm{d})})$ for $q = 1, \ldots, 5$. Additionally, four continuous covariates (standardised) were included—adjusted total hippocampal and ventricles volumes at baseline, HV and VV, and their corresponding longitudinal rate of changes, HV rate and VV rate—jointly taking a multivariate Gaussian distribution given cluster assignment: $W_i^{(\mathrm{c})} | \Theta_h^{(\mathrm{c})} \sim \mathcal{N}_4(\mu_h, \Sigma_h)$. This allows for the correlation between the continuous covariates to be taken into account.

Since the clustering assignments and number of clusters vary across the MCMC iterations, it is useful to obtain a "representative" clustering that summarises the MCMC output. Following (Molitor et al., 2010; Liverani et al., 2015), we find a "representative" clustering based on the $N \times N$ posterior similarity matrix S, where element (i, j) of S is the proportion of MCMC iterations where subjects i and j are assigned to the same cluster. The partitioning around medoids (PAM) clustering algorithm (Kaufman and Rousseeuw, 1990) is applied to the posterior dissimilarity matrix $1 - S$ to find a clustering of the subjects that is consistent with S, with the optimal number of clusters selected using the silhouette width method (Rousseeuw, 1987).

An advantage of the DPMM clustering framework is that it takes uncertainty in the clustering (including the number of clusters) into account. This allows the uncertainty associated with the "representative" clustering to be investigated. If we let $C_h^{(\mathrm{rep})}$ denote the subset of subjects allocated to cluster h in the "representative" clustering, then at MCMC iteration r we can calculate the average value of mixture component parameters for subjects in $C_h^{(\mathrm{rep})}$. For example, for the Bernoulli distribution parameter for binary covariate q we calculate

$$\bar{\theta}_{h,q}^{(\mathrm{d})}(r) = \frac{1}{n_h} \sum_{i \in C_h^{(\mathrm{rep})}} \theta_{c_i^*(r),q}^{(\mathrm{d})}(r) \tag{8}$$

where n_h is the number of subjects in $C_h^{(\mathrm{rep})}$ and $\theta_{c_i^*(r),q}^{(\mathrm{d})}(r)$ is the sampled Bernoulii parameter for the cluster $c_i^*(r)$ that subject i is allocated to at MCMC iteration r. The distribution of $\bar{\theta}_{h,q}^{(\mathrm{d})}(r)$ across the MCMC iterations (i.e., the posterior distribution) gives an insight into the uncertainty of cluster h in the "representative" clustering; narrower credible intervals indicates a more consistent clustering. These distributions can be computed for all of the "representative" clusters and for all of the mixture component parameters associated with the outcome variable and covariates.

Bayesian profile regression is implemented in the R package PReMiuM (Liverani et al., 2015) and this was used to fit the model and perform the post-processing analysis (PReMiuM package version 3.2.3; R version 3.6.3; default settings for hyperparameters used; run for 350,000 MCMC iterations with first 100,000 discarded as burn-in). Convergence of the MCMC procedure was investigated by checking agreement between the "representative" clusterings from six independent chains (quantified using the adjusted Rand index) and by inspection of posterior parameters (see (Liverani et al., 2015) for more details of convergence diagnostics). Consensus clustering of the consensus dissimilarity matrix, obtained through averaging of the dissimilarity matrices from the six independent chains and applying PAM to this matrix, resulted in the final representative clustering structure. The adjusted Rand indices assessing agreement between the final representative consensus clustering with the representative clusterings from the six independent chains are calculated and reported. Moreover, the lower triangular part of the individual posterior dissimilarity matrices from the six independent chains are compared to the lower triangular part of the consensus posterior dissimilarity matrix using Pearson's correlation. Risk and covariate profiles are derived through pooling of MCMC iterations across the six chains and using the final representative consensus clustering. Additionally, Bayesian profile regression without the latent classes as outcome was performed to obtain a baseline/reference clustering structure based purely on the biomarkers. All 1,543 subjects included in the MLCMM analysis were included in the Bayesian profile regression analysis.

2.2.3 Validation

The final results of our multivariate latent class mixed model and Bayesian profile regression analysis on the full data-set were assessed for class and cluster validity through stability assessment under repeated sub-setting. We repeatedly (i.e., ten times) split the full data-set into two subsets, by first stratifying the full data-set by number of visits and then randomly allocating (with equal probability) within each strata a participant to belong to either

the first or second subset. Our proposed two-stage approach is then applied in turn to each subset to estimate a latent class structure followed by clustering structure as described in the previous two subsections. For assessing class validity for each split, we begin by using the multivariate latent class model trained on one subset to predict the class membership of participants in the other subset and vice versa. We next cross-tabulate the out-of-sample predictions of class memberships (based on the model trained on the subset that does not include the participant for whom a prediction is being made) with the in-sample class membership assignments (obtained from the model trained on the subset which includes the participant for whom a prediction is being made) to assess out-of-sample performances in the models trained on the two subsets and stability of class structure across the two subsets. Cohen's kappa statistic and the adjusted Rand index are used to measure the out-of-sample performances and across subset stability. Finally, the validity/stability of the class structure obtained from the full data-set is evaluated by a comparison of the in-sample class assignments based on the final multivariate latent class mixed model on the full data-set to the in-sample class assignments obtained from the multivariate latent class models for the subsets, using again Cohen's kappa and the adjusted Rand index.

For assessing clustering validity, we first apply Bayesian profile regression to the biomarker and latent class assignment data for each subset in turn and obtain the consensus results over six chains as described earlier. Next, the consensus dissimilarity matrices for the subsets are compared to their corresponding block diagonals of the consensus dissimilarity matrix from our Bayesian profile regression on the full data-set using Pearson's correlation. To assess cluster stability, the corresponding PAM consensus representative clustering structures from each subset are compared to the final representative clustering from the full data-set using the adjusted Rand index. Moreover, we make predictions for the held-out subsets that allow us to compare 1) their predicted dissimilarity matrices with the corresponding off-diagonal blocks of the final consensus dissimilarity matrix from the full data-set using Pearson's correlation, and 2) their predicted clustering structures with the PAM consensus representative clustering obtained using a model trained on the held-out subset (clustering predictions are obtained by using the predicted dissimilarity matrices to assign participants in the held-out subset to the PAM consensus representative cluster from the training subset that they are closest to).

External validation was not possible as we do not have access to data from studies on similar populations with the corresponding extensive baseline and longitudinal biomarker and phenotypic information to EPAD.

3 RESULTS

3.1 Baseline Characteristics of the European Prevention of Alzheimer's Dementia Longitudinal Cohort Study Population

Table 1 describes the group of 1,574 participants with two or more visits in the EPAD longitudinal cohort. The mean age of these participants was 65.4 years with a standard deviation of 7.4 years. Around 56% were female and 63% had their highest educational attainment beyond secondary education—an indication of a highly educated cohort of participants recruited; reflecting the eligibility criterion on minimum years of formal education. The cohort was enriched for participants with a family history of AD (first degree relatives) and APOEϵ4 carriers, without diminished decision-making capacity. For the group, this enrichment corresponded to 65.5 and 37.5% with a known family history of AD and a known carrier for APOEϵ4 respectively. 78% of this group (n = 1,226) had a global CDR of 0, while the remaining 22% had a score of 0.5 (n = 346); two participants had unknown baseline CDR global. Around 82% of those with a family history of AD had a CDR global score of 0. Whereas 70% of those without a family history of AD had a CDR score of 0. Thus there was a clear association between CDR global score and family history of AD favouring the recruitment of participants with a family history who do not have any baseline cognitive impairment and for those without a family history enriching for early symptomatics ($p < 0.0001$; χ^2-test). No evidence for an association between CDR global score and APOEϵ4 carrier status was found ($p = 0.10$), with 80% of non-carriers and 76% of carriers having CDR global equal 0.

Table 1 also summarises the distributions of the EPAD cognitive and clinical outcomes and CSF and neuroimaging biomarkers at baseline. Ten percent of participants had an MMSE score below 27 and 12.5% had a CDRSB score of 1 or above; suggesting that the majority of participants had high levels of cognitive functioning at baseline. However, varying degrees of disease pathology at baseline were indicated on considering a range of biomarkers. AD positivity was estimated around 20% using the ratio of phosphorylated tau to amyloid-beta 42 in CSF. Convincing evidence for the widening of the choroid fissure to different degrees (average of left and right MTA ≥1) was found in about a quarter of the participants, whilst varying percentages of white matter lesion cerebrovascular pathology were seen ranging from 6 to 68% based on age-related regional white matter changes or based on an overall impression of the brain using the Fazekas scales (approximately 16 and 39%). Nearly a quarter of the participants (23.5%) had indications of cerebrovascular pathology in three or more of the five age-related white matter regions. The mean adjusted total hippocampal and ventricles volumes at baseline (with standard deviation) were 5,793mm^3 (703mm^3) and 32,991mm^3 (17,669mm^3).

3.2 Disease Progression and Latent Phenotypes—Results From MLCMM

Our MLCMM was able to identify four distinct mean trajectories. **Figure 2** shows these four mean trajectory profiles on the latent process scale and on the original scales for CDRSB and MMSE. Latent clinical phenotype classes 0 to 3 had, respectively, 1,050 (68.0%), 97 (6.3%), 106 (6.9%), and 290 (18.8%) individuals hard assigned to them based on a posterior classification of participants' class membership through the selection of the participant's class with the highest posterior class-membership probability. Latent phenotype class 0, which had the majority of participants, is

TABLE 1 | Baseline characteristics of the 1,574 participants with more than one visit.

	Variable		Mean (SD)	Frequency (%)	No. Unknown
Risk Factors	Age, years		65.4 (7.4)		0
	Sex	Female		888 (56.4)	0
		Male		686 (43.6)	
	Education	Level 1		587 (37.3)	0
		Level 2		393 (25.0)	
		Level 3		594 (37.7)	
	Family history of AD	No		543 (34.5)	0
		Yes		1,031 (65.5)	
	APOEϵ4 carrier	No		965 (62.5)	31
		Yes		578 (37.5)	
Outcomes	CDRSB	0		1,162 (73.9)	2
		0.5		214 (13.6)	
		≥1		196 (12.5)	
	MMSE	29–30		999 (63.5)	1
		27–28		417 (26.5)	
		≤26		157 (10.0)	
	Transformed CDRSB, tCDRSB		4.60 (1.04)		2
	Normalised MMSE, nMMSE		83.6 (14.6)		1
Biomarkers	pTau/Aβ	≤0.024		1,240 (80.5)	33
		>0.024		301 (19.5)	
	MTA average	0		800 (51.2)	13
		0.5		375 (24.0)	
		≥1		386 (24.7)	
	Fazekas scale deep	<2		1,317 (84.4)	13
		≥2		244 (15.6)	
	Fazekas scale periventricular	<1		947 (60.7)	13
		≥1		614 (39.3)	
	ARWMC basal ganglia	<1		1,379 (88.3)	13
		≥1		182 (11.7)	
	ARWMC frontal	<1		506 (32.4)	13
		≥1		1,055 (67.6)	
	ARWMC infratentorial	<1		1,465 (93.9)	13
		≥1		96 (6.1)	
	ARWMC parieto-occipital	<1		786 (50.4)	13
		≥1		775 (49.6)	
	ARWMC temporal	<1		1,268 (81.2)	13
		≥1		293 (18.8)	
	ARWMC combined	<3		1,194 (76.5)	13
		≥3		367 (23.5)	
	Total hippocampal volume (adj), mm^3		5,793 (703)		62
	Total ventricular volume (adj), mm^3		32,991 (17,669)		168

characterised by individuals having the highest levels of cognitive functioning with no signs of impairment at baseline and no decline throughout the course of the study. Class 1 contained individuals who showed some signs of cognitive/functional impairment at baseline but appeared to improve over time. Class 2 was characterised by individuals who appeared cognitively and functionally unimpaired at baseline (although cognitive functioning levels were not as high as those in class 0) but then declined on follow-up. Whereas class 3 contained individuals who showed the most evident signs of early cognitive/functional impairment at baseline and continued to show impairment on follow-up.

Table 2 reports the results from our four-class MLCMM. A higher baseline age and a lower level of education are associated with higher levels of cognitive/functional impairment; consistent with findings from the neurodegenerative and AD literature. Due to how individuals were recruited into the study (through use of a flexible and dynamic approach to selection), biased effects of family history of AD and APOEϵ4 carrier status were expected and therefore the corresponding estimates of these effects were not interpreted as they were notably affected by the selection mechanism. For example, both were found not to be statistically significantly associated with cognitive/functional impairment and the effect of family history of AD was in the opposite direction to that reported in the literature.

The measurement error variances for tCDRSB and nMMSE are 0.531 and 3.527 respectively indicating that tCDRSB has a stronger relationship to the underlying latent disease process. The estimated class-specific proportionality factors, $\hat{\omega}_g$ $(g = 0, 1, 2)$,

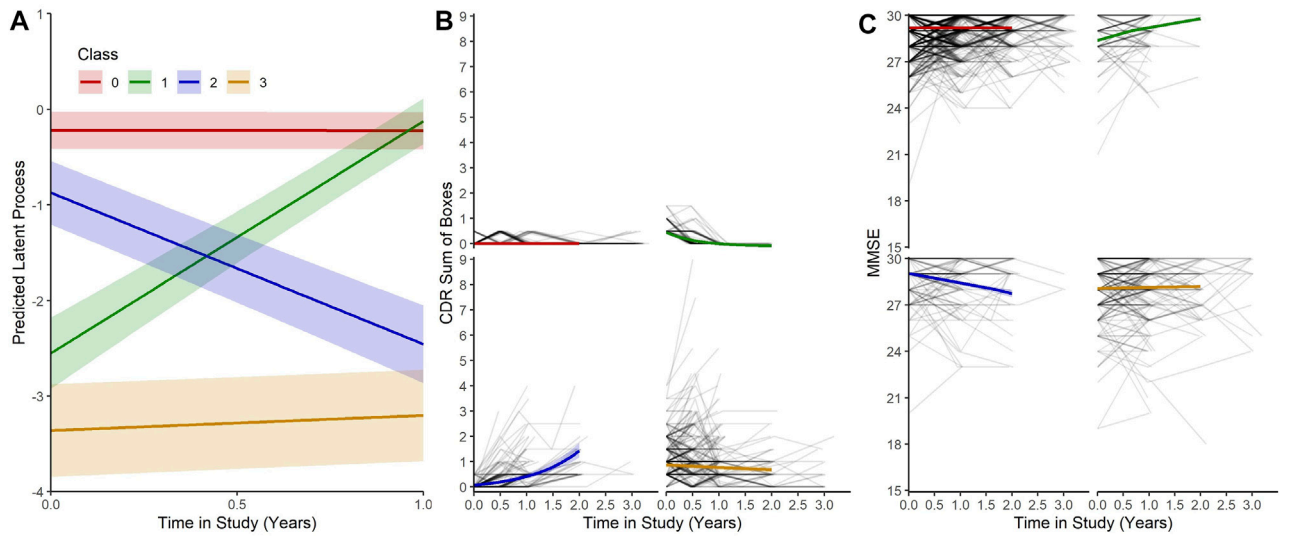


FIGURE 2 | (A) Predicted trajectories with 95% confidence bands for each class on the latent process scale given mean values for each of the covariates. **(B)** and **(C)** Predicted trajectories with 95% confidence bands for each class on the CDRSB and MMSE scale given mean values for the each of the covariates with observed outcomes for each participant.

TABLE 2 | Results of the 4-class MLCMM on the 1,543 participants.

		Coefficient (SE)	*p*-value
Class membership model	Intercept class 0	1.30 (0.07)	<0.0001
	Intercept class 1	−1.04 (0.12)	<0.0001
	Intercept class 2	−0.87 (0.13)	<0.0001
Fixed effects model	Intercept class 0	0 (not estimated)	—
	Intercept class 1	−2.33 (0.15)	<0.0001
	Intercept class 2	−0.65 (0.14)	<0.0001
	Intercept class 3	−3.14 (0.22)	<0.0001
	Time in study class 0	−0.0040 (0.014)	0.773
	Time in study class 1	2.43 (0.16)	<0.0001
	Time in study class 2	−1.58 (0.11)	<0.0001
	Time in study class 3	0.16 (0.04)	0.0001
	Age	−0.0033 (0.0014)	0.022
	Sex male	−0.015 (0.019)	0.443
	Education level 2	0.022 (0.024)	0.367
	Education level 3	0.043 (0.022)	0.049
	Family history of AD	0.016 (0.021)	0.453
	APOEϵ4	−0.021 (0.019)	0.290
Link function parameters	tCDRSB η_1	5.31 (0.07)	<0.0001
	tCDRSB η_2	0.73 (0.04)	<0.0001
	nMMSE η_1	87.96 (0.48)	<0.0001
	nMMSE η_2	3.80 (0.30)	<0.0001

which here correspond to the variances of the class-specific random intercepts are 0.0004 for class 0, 0.397 for class 1, and 0.957 for class 2. (The variance for the random intercept corresponding to class 3 was set to 1 for identifiability.) The log-likelihood of this model was −16,842.95, the BIC 33,869.44, and the (relative) entropy 0.947. By comparison, the equivalent three-class model had a log-likelihood of −17,097.26, a higher BIC of 34,348.70, and a lower entropy of 0.932. Thus our four-class model was preferred. It showed excellent ability to discriminate between latent trajectory classes.

Moreover assessment of validity of our four-class model through class stability under repeated sub-setting gave mean Cohen's kappa and adjusted Rand index values (with standard deviations) of 0.987 (0.008) and 0.989 (0.006), respectively, across the twenty subset comparisons and 0.993 (0.004) and 0.995 (0.003) for the ten comparisons against the full data-set. These results indicate near perfect agreement with evidence for stability across subsets and validity of the class structure derived based on the full data-set. Across the ten splits, the number of discordant classifications seen when the in-sample latent class membership predictions for subsets are compared to the class memberships predicted by our four-class model on the full data-set ranged from 3 to 13 out of the 1,543 participants (0.19–0.84%). For the twenty subsets across the ten splits, four-class multivariate latent class mixed models were always found to provide a better fit (based on BIC) than the alternative three-class multivariate latent class mixed models, and these four-class models had similar class structure as our four-class model on the full data-set.

We further characterised these four latent phenotype classes by baseline and change variables and (marginally) compared these variables across classes using analysis of variance (ANOVA) tests for the continuous variables and χ^2 tests for binary and categorical variables. The results are shown in **Table 3**. We observe increasing trends in mean age and mean baseline ventricles volume across the latent classes from 0 to 3 and a decreasing trend in mean baseline hippocampal volume. Class 3 differed from the other three classes in having the highest proportions of males, lowest educational level attainers, those with AD positivity at baseline and with evidence on baseline MTA of widening of choroid fissure in varying degrees from widen to end stage atrophy. There was evidence found for differences amongst the

TABLE 3 | Characterisation of the baseline and change variables by latent phenotype classes.

		Mean (SD)				
Variable		**Class 0**	**Class 1**	**Class 2**	**Class 3**	**ANOVA *p*-value**
Age, years		63.9 (7.0)	65.6 (6.6)	68.1 (7.0)	69.5 (7.0)	<0.0001
Total hippocampal volume (adj), mm^3		5,911 (644)	5,814 (725)	5,609 (715)	5,429 (768)	<0.0001
Total ventricular volume (adj), mm^3		30,715 (16,348)	35,404 (19,823)	37,962 (18,838)	38,396 (19,405)	<0.0001
Annual (adj) hippocampal volume change, mm^3/yr		−9.4 (83.5)	−30.4 (61.9)	−40.2 (85.3)	−55.3 (99.1)	<0.0001
Annual (adj) ventricular volume change, mm^3/yr		988 (910)	1,430 (1,354)	1,958 (1,651)	1,688 (1,586)	<0.0001
		Frequency (%)				
Variable		**Class 0**	**Class 1**	**Class 2**	**Class 3**	**χ^2 *p*-value**
Sex	Female	614 (58.5)	54 (55.7)	65 (61.3)	137 (47.2)	0.005
	Male	436 (41.5)	43 (44.3)	41 (38.7)	153 (52.8)	—
Education	Level 1	368 (35.0)	35 (36.1)	34 (32.1)	138 (47.6)	0.009
	Level 2	267 (25.4)	24 (24.7)	30 (28.3)	63 (21.7)	—
	Level 3	415 (39.5)	38 (39.2)	42 (39.6)	89 (30.7)	—
Family history of AD	No	319 (30.4)	36 (37.1)	40 (37.7)	136 (46.9)	<0.0001
	Yes	731 (69.6)	61 (62.9)	66 (62.3)	154 (53.1)	—
APOEϵ4 carrier	No	656 (62.5)	68 (70.1)	73 (68.9)	168 (57.9)	0.078
	Yes	394 (37.5)	29 (29.9)	33 (31.1)	122 (42.1)	—
pTau/Aβ	≤0.024	906 (87.5)	74 (77.9)	71 (71.0)	167 (59.4)	<0.0001
	>0.024	130 (12.5)	21 (22.1)	29 (29.0)	114 (40.6)	—
MTA average	<1	856 (82.1)	63 (67.0)	75 (71.4)	158 (54.7)	<0.0001
	≥1	186 (17.9)	31 (33.0)	30 (28.6)	131 (45.3)	—
Fazekas scale deep (FSD)	<2	893 (85.7)	83 (88.3)	76 (72.4)	236 (81.7)	0.002
	≥2	149 (14.3)	11 (11.7)	29 (27.6)	53 (18.3)	—
Fazekas scale periventricular (FSPV)	<1	660 (63.3)	61 (64.9)	53 (50.5)	155 (53.6)	0.002
	≥1	382 (36.7)	33 (35.1)	52 (49.5)	134 (46.4)	—
ARWMC basal ganglia	<1	929 (89.2)	83 (88.3)	90 (85.7)	246 (85.1)	0.248
	≥1	113 (10.8)	11 (11.7)	15 (14.3)	43 (14.9)	—
ARWMC frontal	<1	346 (33.2)	36 (38.3)	27 (25.7)	86 (29.8)	0.182
	≥1	696 (66.8)	58 (61.7)	78 (74.3)	203 (70.2)	—
ARWMC infratentorial	<1	988 (94.8)	87 (92.6)	96 (91.4)	266 (92.0)	0.195
	≥1	54 (5.2)	7 (7.4)	9 (8.6)	23 (8.0)	—
ARWMC parieto-occipital	<1	549 (52.7)	48 (51.1)	45 (42.9)	127 (43.9)	0.025
	≥1	493 (47.3)	46 (48.9)	60 (57.1)	162 (56.1)	—
ARWMC temporal	<1	863 (82.8)	77 (81.9)	78 (74.3)	223 (77.2)	0.043
	≥1	179 (17.2)	17 (18.1)	27 (25.7)	66 (22.8)	—
ARWMC combined	<3	815 (78.2)	71 (75.5)	72 (68.6)	211 (73.0)	0.061
	≥3	227 (21.8)	23 (24.5)	33 (31.4)	78 (27.0)	—

classes in the presence of white matter hyperintensities in the entire brain as measured by the Fazekas scales, with latent class 2 having the highest proportion of participants with abnormal pathology. No evidence of any further differences between classes 3 and 2 (or between classes 0 and 1) was found with regard to age-related regional white matter changes. However evidence of differences between the lower two classes (0 and 1) compared to the upper two classes (2 and 3) was found for a number of these neuroimaging variables associated with white matter lesions (**Table 3**).

On examining possible associations of longitudinal changes in volumetric imaging measures and the latent phenotype classes, we observe an increasing annualised hippocampal shrinkage with increasing class, without a similar trend being seen between annualised ventricular enlargement and latent phenotype class. Notably, latent phenotype class 2 had the largest annualised increase in ventricles volume (**Table 3**).

3.3 Neuropathological Endotypes—Results From Profile Regression

For the profile regression analysis, which links CSF and neuroimaging biomarkers to the latent clinical trajectory phenotype, we ran six independent MCMC chains, and obtained a posterior similarity matrix and associated PAM "representative" clustering from the output of each chain (see Section 2.2.2 for details). Agreement between these six chains was high with mean pairwise Pearson's correlation of 0.95 (standard deviation 0.03) between the dissimilarity matrices and mean pairwise adjusted Rand index of 0.90 (standard deviation 0.05) between the representative clusterings. This, together with inspection of posterior parameters for each chain, suggests that there is no strong evidence against convergence of the MCMC and there is a good level of robustness of the clustering structure. Three of the "representative" clusterings have six clusters, while the other three had seven.

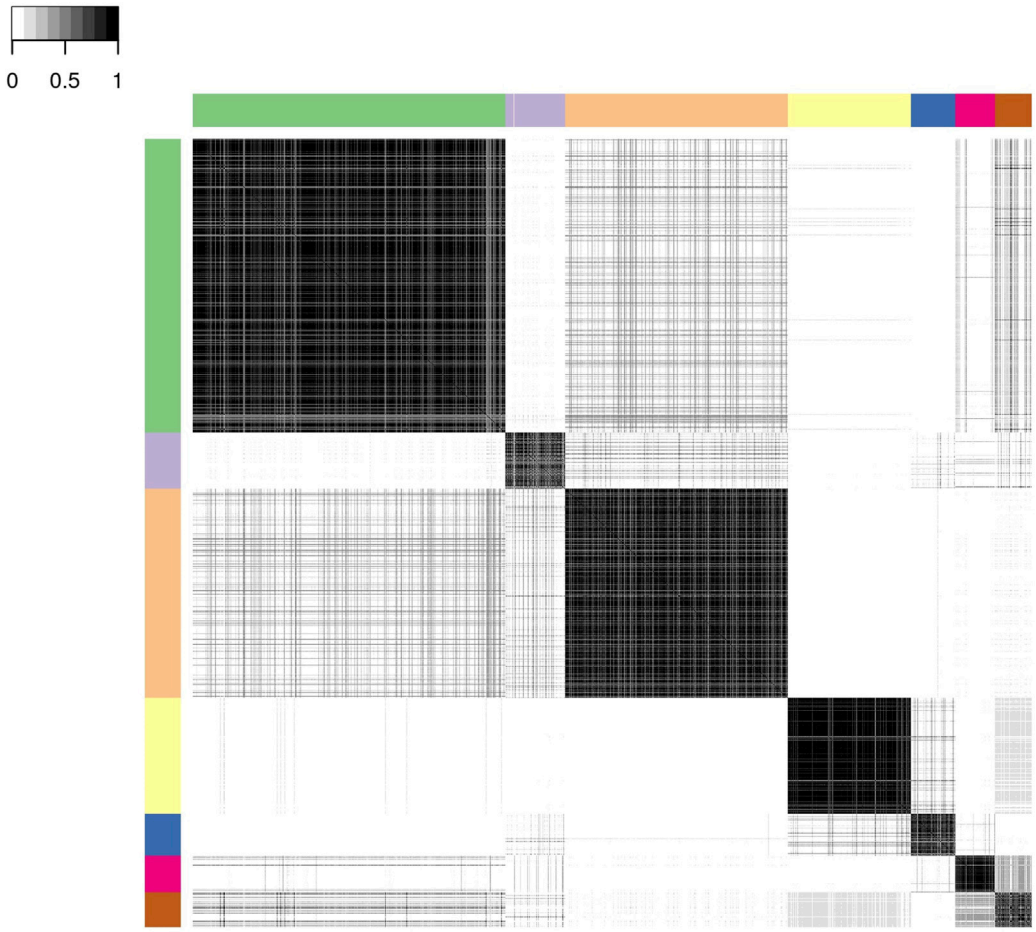


FIGURE 3 | Posterior similarity matrix for the consensus across the six MCMC chains from the Bayesian profile regression analysis on the 1,543 EPAD participants. Each entry (i, j) of this 1,543 × 1,543-matrix represents the proportion of times participants i and j are assigned to the same cluster over the 250,000 × 6 MCMC iterations. Color bars indicate the seven final PAM consensus representative clusters of participants identified. See **Figure 4** for more information regarding these clusters.

We present below the results of applying consensus clustering to aggregate the output from the six MCMC chains. The mean Pearson's correlation between the consensus dissimilarity matrix and the six independent chains' dissimilarity matrices was 0.98 (standard deviation 0.01). Similarly, the mean adjusted Rand index between the consensus representative clustering and the six representative clusterings from each chain was 0.93 (standard deviation 0.03). The final consensus representative clustering has seven clusters.

Figure 3 shows the consensus posterior similarity matrix that summarises the output from across the six MCMC chains and the seven representative clusters that were identified from this matrix. **Figure 4** and **Table 4** describe these seven clusters and their distinct biomarker profiles (i.e., neuropathological endotypes). Cluster 1, which is the largest cluster (comprising of 575 out of 1,543 participants), estimated the posterior mean probability of belonging to latent phenotype class 0 to be 92% (in agreement with the empirical estimate of 94%). It was characterised by participants with lower than expected/average probabilities of having abnormal pathology on the various biomarkers and above average healthy indicators of baseline and longitudinal volumetric measures for hippocampus and ventricles. We label this cluster as a "healthy brain" neuropathological endotype. It had on average the youngest participants, with a mean age (SD) of 61.4 (6.2) years.

Cluster 2, which is a mixture of participants from both latent phenotype classes 0 and 1 (85 and 15% respectively), had somewhat lower than average AD positivity risk (but within the margin of uncertainty of the overall mean) and had stable hippocampal volume over time, but otherwise had higher than expected risk of abnormal pathology on the other biomarkers, including medial temporal lobe atrophy (MTA) indicating hippocampal involvement, and 1.59 standard deviations (SDs) higher baseline ventricles volume and 0.32 SD faster annual rate of increase in ventricles volume above their average, which is being tolerated so far. This cluster appears to be a non-AD driven cluster with "kindling" cerebrovascular disease. We label it as an "at-risk-of-vascular dementia" neuropathological endotype. The

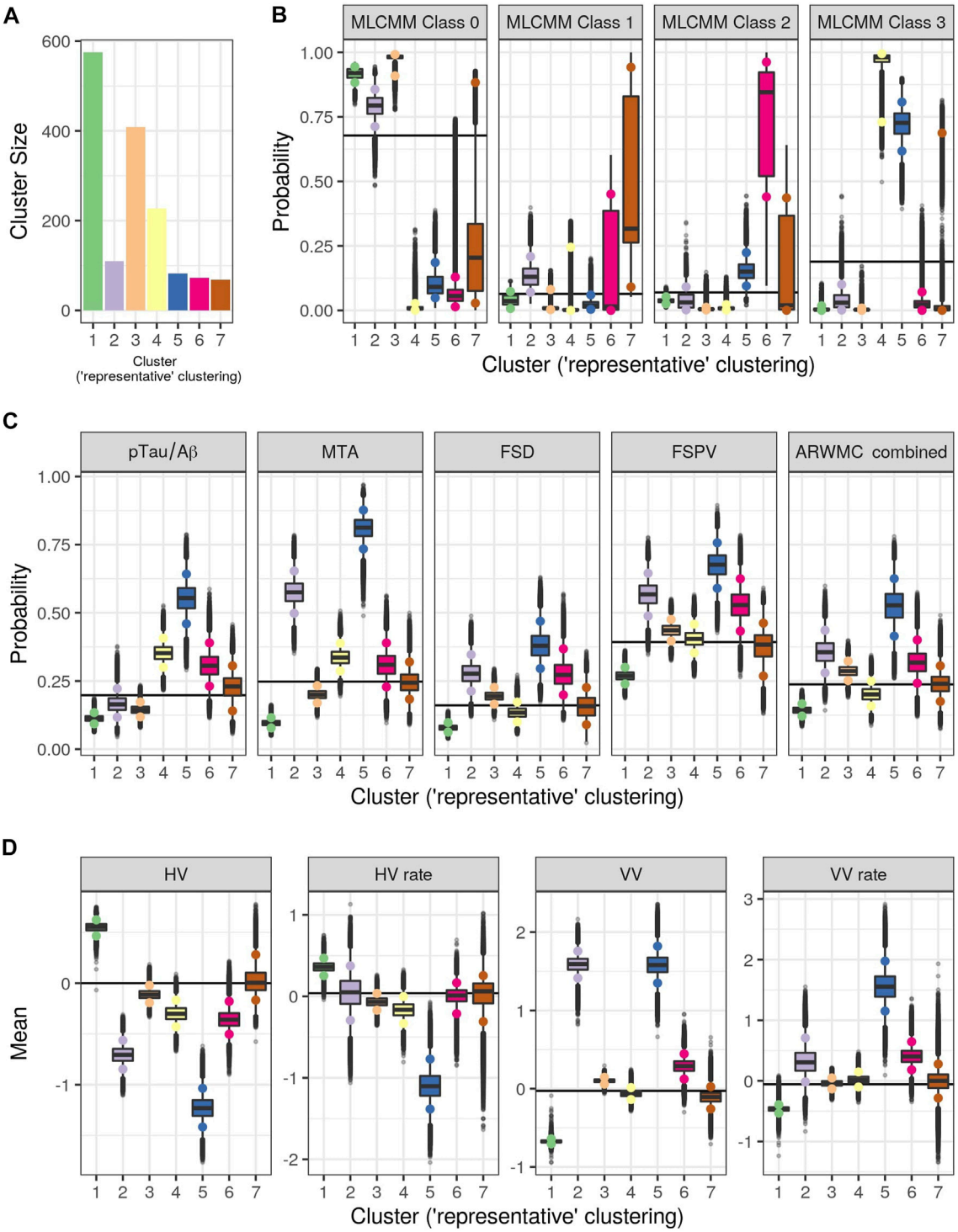


FIGURE 4 | Results from Bayesian Profile Regression analysis. **(A)** Cluster sizes for the final PAM representative consensus clusters. **(B–D)** Posterior distributions for mean mixture component parameter values for each of the "representative" clusters (see Section 2.2.2). **(B)** Outcome variable (parameters are the probability of belonging to each MLCMM latent class). **(C)** Binary covariates (parameters are the probability of the covariate having value of one). **(D)** Continuous covariates (parameters are the mean covariate value). For **(A–D)**, colors indicate clusters (see also **Figure 3**). For **(B–D)**, black horizontal lines indicate the mean parameter values across all subjects and the coloured circles indicate the upper and lower limit of the 90% credible interval.

TABLE 4 | Results from the Bayesian profile regression analysis on the 1,543 participants.

		Posterior means											
		Probability of abnormal pathology					SD distance from overall mean				Class membership probability			
Clusters	N (%)	pTau/ Aβ	MTA	FSD	FSPV	ARWMC combined	Mean HV	Mean HV rate	Mean VV	Mean VV rate	Class 0	Class 1	Class 2	Class 3
1	575 (37.3)	0.113	0.096	0.079	0.269	0.143	0.549	0.362	−0.674	−0.463	0.917	0.040	0.038	0.005
2	110 (7.1)	0.166	0.575	0.278	0.567	0.357	−0.706	0.047	1.590	0.321	0.791	0.133	0.038	0.039
3	409 (26.5)	0.145	0.200	0.195	0.436	0.287	−0.111	−0.066	0.101	−0.042	0.976	0.015	0.006	0.003
4	227 (14.7)	0.353	0.337	0.135	0.405	0.202	−0.300	−0.166	−0.063	0.020	0.010	0.040	0.009	0.941
5	82 (5.3)	0.553	0.810	0.380	0.675	0.524	−1.229	−1.093	1.583	1.558	0.101	0.024	0.154	0.721
6	72 (4.7)	0.308	0.309	0.276	0.529	0.319	−0.354	0.000	0.287	0.407	0.064	0.177	0.731	0.028
7	68 (4.4)	0.228	0.247	0.157	0.376	0.240	0.025	0.026	−0.109	−0.004	0.248	0.464	0.173	0.115
Overall empirical mean		0.194	0.247	0.158	0.393	0.236	5,793	-23.8	32,997	1,274	0.680	0.063	0.069	0.188
SD							705	88.7	17,687	1,264				

mean age (SD) of participants was 69.1 (7.0) years and percentage APOEe4 positive was 28%, the lowest amongst the clusters.

Cluster 3, which is the second largest in size (409 participants; all from latent phenotype class 0), is characterised by lower than expected risk of both AD positivity and MTA abnormal pathology and no clinically meaningful pathological indications on volumetric neuroimaging; but with evidence of white matter lesion pathology. We describe this cluster as a "healthy ageing" endotype especially as participants are on average older than those in cluster 1, with a mean age (SD) of 66.1 (6.5) years, and they appear to be able to compensate for some cerebrovascular disease. Moreover, none of the differences from overall average for any of the biomarkers were particularly large.

Cluster 4 has 15% of the participants—with average age (SD) of 68.1 (6.8) years–all of whom belong to latent phenotype class 3 (questionably cognitively impaired class). It is characterised by clinically meaningful increased risk of AD positivity and MTA abnormal pathology, early pathological indications on hippocampal volume markers and slightly increased proportion of APOEϵ4 carriers relative to the overall average (0.4 versus 0.375) and may represent a subgroup of "AD high risk" participants.

Cluster 5 represents the 5% of the cohort who have the highest risk, worst baseline levels and fastest rate of worsening on markers. It comprises of a mixture of participants from latent phenotype classes 0 (6%), 2 (17%) and 3 (77%). We consider this to be an "AD-related cluster". Moreover, it has the highest mean age of 74.2 years (SD 5.6 years) amongst the seven clusters and, notably, the highest proportion of APOEϵ4 carriers (0.46) despite the EPAD selection mechanism.

Finally clusters 6 and 7, which are the most uncertain ones (i.e., empirical class membership proportions of 100% in class 2 for cluster 6 and class 1 for cluster 7 do not match with the corresponding mean posterior probabilities for these classes of 73 and 46% respectively), correspond to clusters where there are, respectively, evidence of increased abnormal pathology on all markers (except hippocampal atrophy and MTA) and no particular overall evidence of increased abnormal pathology beyond expected on any particular biomarker. Cluster 6 may be another AD-related cluster, but one, possibly, in an earlier stage of progression (cf cluster 5) as they are on average 5.6 years younger, with a mean age (SD) of 68.6 (6.3) years. Cluster 7 appears to have individuals with both unclear biomarker profiles and unclear cognitive trajectories, and therefore we describe it as an "ambiguous" cluster. The mean age (SD) here is 66.0 (6.5) years.

We assessed clustering validity through stability under repeated sub-setting (10 splits, totalling 20 subsets of the data). Out of the twenty consensus representative clustering structures obtained from applying Bayesian profile regression to the twenty subsets, 8 and 10 of these clustering structures consisted of four and five clusters respectively, while the other two comprised three and six clusters. Agreement between the consensus clusterings and the clusterings from the corresponding six independent MCMC chains across the twenty subsets were again high with mean adjusted Rand index of 0.93 (standard deviation of 0.10). The reduced number of clusters relative to the seven clusters found using the full data-set is likely due to the 50% reduction in sample size for the subsets. A comparison of the consensus representative clustering obtained using the subsets of data with the consensus representative clustering obtained using the full data-set (restricted to those individuals in each subset for the comparisons) resulted in a mean adjusted Rand index of 0.69 (standard deviation of 0.09). Furthermore, comparing the 20 consensus posterior dissimilarity matrices obtained from the subsets against those obtained using the corresponding

submatrices of the full data-set resulted in a mean Pearson's correlation of 0.860 (standard deviation of 0.049). These comparisons indicate good agreement between the results obtained on the subsets and those obtained using the full data-set, giving evidence for stability of our results.

Additionally, the held-out prediction analyses (training on one subset and predicting for the other in each split) resulted in a mean Pearson's correlation of 0.674 (standard deviation of 0.045) between the predicted posterior dissimilarity matrices and the corresponding submatrices obtained using the full data-set. Comparing each of the estimated consensus representative clustering obtained from one subset of the split with the predicted clustering for this subset (predictions obtained using a model trained on the other subset of the data split only) resulted in a mean adjusted Rand index of 0.464 (standard deviation of 0.076) over the 20 comparisons. This performance on challenging held-out prediction tasks gives further support for the validity and stability of our clustering results.

The consensus representative clustering structure obtained using Bayesian profile regression without the MLCMM class outcome (i.e., only using biomarker covariates) had an adjusted Rand index of 0.48 with the clustering that did include the outcome, indicating that the outcome is playing an influential role in the clustering analysis and is facilitating interpretation of the clusters in terms of linking them to latent clinical phenotypes.

4 DISCUSSION

In this paper, we demonstrate the usefulness of our two-stage approach in, firstly, characterising the evolution of correlated cognitive and clinical outcomes for LCS participants via an underlying latent process in which its trajectory depends on one of four latent clinical phenotypes, and then in providing biological insight through the identification of subgroups based on distinct biomarker profiles (i.e., neuropathological endotypes) linked to the latent phenotypes. Our approach recognises that the longitudinal cognitive and clinical outcomes are the downstream clinical manifestations/consequences of earlier endogenous biological changes occurring within the brain whether they be due to normal brain ageing or pathological due to a specific underlying disease process. It however does not attempt to assess the exact ordering of the pathological cascade of events.

Our intention here was not to provide a comprehensive clinical and biological investigation of the EPAD LCS data but to demonstrate the utility of our two-stage strategy in uncovering meaningful clinical and biological structure within this heterogeneous population. Therefore we chose to use a reduced set of coarser, but still relevant, ATN (amyloid-beta deposition (A), pathologic tau (T), and neurodegeneration (N)) and cerebrovascular biomarkers to demonstrate our two-stage approach. If interest lies in a more thorough investigation, then our approach can be extended to incorporate a larger set of biomarkers, providing more granular information (e.g., both left and right MTAs and hippocampal volumes and all five ARWMC regions could be considered instead of the average, total or majority as was done in this paper; with additional markers such as the Koedam score, which measures parietal atrophy, included), and additional correlated cognitive or clinical outcomes (e.g., specific cognitive domains). However, with more biomarkers being considered, this could result in increased uncertainty and instability in clustering structure obtained through use of Bayesian profile regression. Therefore we would recommend the incorporation of a variable selection component into the Bayesian profile regression analysis in order to identify the actual drivers of the clustering structure. Related issues may arise regarding both the number and relevance of latent classes arrived at when additional outcomes are added to the multivariate latent class mixed effects analysis, especially when weakly informative or conflicting outcomes are included.

The latent process arising from the multivariate latent class mixed modelling (MLCMM) approach appeared to be more highly correlated with the observed transformed CDR sum of boxes score than to the normalised Mini-Mental State Examination score, possibly reflecting the former being more sensitive to underlying changes than the latter early on. Nevertheless both CDRSB and MMSE produced concurring patterns with each other across the four latent phenotype classes (see **Figure 2**). These four trajectories correspond to a normal cognitive functioning class throughout, a reversion class, a declining class and a (questionable) cognitively impaired class. They are consistent with what has been reported previously in the literature, although the reversion class probably reflects measurement error. Interestingly, with our Bayesian profile regression analysis, we were able to find endotypes covering the full spectrum from "healthy brain" to "AD-related" within the EPAD cohort; reflecting one of the aims of the EPAD LCS to provide a well-phenotyped population covering the full continuum of risk of subsequent AD dementia development.

We note that the diminishing numbers at each visit reflect both the staggered opening of the 31 recruitment centres across Europe and the LCS concluding at the end of the IMI funding period. Attempts to further fund the cohort as a whole across Europe were not successful, in large part due to the COVID-19 pandemic. Attempts are ongoing to follow-up these participants in a series of studies across Europe to provide longer term clinical and biological outcomes.

The second objective of the EPAD LCS was to create a trial-ready cohort for potential recruitment into the EPAD PoC Trial. Unfortunately, this trial was not realised. However, our approach can still be used to demonstrate trial-readiness with respect to both minimising screen failures and identifying participants with particular biomarker profiles eligible for recruitment. For example, participants identified/pre-screened as belonging to the "healthy brain" or "healthy ageing" clusters would not be considered for inclusion into trials thereby reducing screen-failure rates currently seen in AD-related trials due to the low prevalence of AD pathology in individuals without dementia, especially among cognitively unimpaired. Whereas, for example, individuals in clusters 4, 5 or 6 may be specifically targeted for phase II trials in which volumetric neuroimaging biomarkers are used as "surrogate" endpoints. While secondary prevention trials in pre-clinical populations with no baseline cognitive impairment may be more inclined to focus recruitment on participants from cluster 6 (or class 2) when the primary endpoint is a cognitive one.

The novelty of our approach is not only in characterising the longitudinal cognitive and clinical outcomes into latent

phenotype trajectories and in identifying neuropathological endotypes, but going beyond identifying substructures to also being able to do future longitudinal clinical prediction in individuals. Briefly, we would combine the posterior predictive probabilities of class membership obtained from both the Bayesian profile regression and the MLCMM, based on the observed relevant biomarker, cognitive and risk factor data, to update the individual's mixture component probabilities in the MLCMM. We would then use these as weights to average over the linear mixed effects submodels corresponding to the four classes in order to predict future transformed CDRSB and normalised MMSE. Currently, the uncertainty attached to the latent trajectory classes is not taken account of in the Bayesian profile regression analysis in our two-stage approach, although this can be rectified by using Markov melding (Goudie et al., 2019). However, we expect this to have little impact on our findings.

In conclusion, we have introduced a two-stage approach for the modelling of longitudinal cognitive and clinical outcomes, biomarkers (baseline and longitudinal) and risk factors to analyse the data from the EPAD Longitudinal Cohort Study and shown its clinical and biological utility in the areas of trajectory stratification, subgroup identification and prediction. In the long term we envisage this approach to be applicable more widely to precision medicine and secondary prevention in Alzheimer's dementia research and practice.

DATA AVAILABILITY STATEMENT

A publicly available dataset was analyzed in this study. This dataset can be found at: http://ep-ad.org/erap/, doi: 10.34688/epadlcs_v.imi_20.10.30.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Independent Ethics Committee or other relevant ethical review board for written approval as required by local laws and regulations. A copy of approval is required by the University of Edinburgh as Sponsor before the study commences at each site. The study is designed and conducted in accordance with the guidelines for Good Clinical Practice (GCP), and with the ethical principles as proclaimed in the Declaration of Helsinki. All participants are required to provide written informed consent prior to participation in any research activities laid out in the EPAD LCS protocol. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

JH, SH, and BT developed the disease progression and probabilistic clustering models, analyzed the data, and drafted the manuscript. CR and BT conceptualised the approach for Alzheimer's disease. CR is the chief investigator of the EPAD Longitudinal Cohort Study and provided clinical input, reviewed and gave critical feedback on the manuscript. The authors read and approved the final manuscript.

FUNDING

This work was funded through the EU/EFPIA Innovative Medicines Initiative Joint Undertaking EPAD grant agreement 115736. BT is supported through the United Kingdom Medical Research Council programme grant No. (MC_UU_00002/2) and supported by the NIHR the Cambridge Biomedical Research Centre.

ACKNOWLEDGMENTS

We thank all the EPAD LCS participants for their enthusiastic involvement in this study.

REFERENCES

Anderson, R. M., Hadjichrysanthou, C., Evans, S., and Wong, M. M. (2017). Why Do So many Clinical Trials of Therapies for Alzheimer's Disease Fail?. *The Lancet* 390, 2327–2329. doi:10.1016/s0140-6736(17)32399-1

Bachman, A. H., and Ardekani, B. A.; for the Alzheimer's Disease Neuroimaging Initiative (2020). Change point Analyses in Prodromal Alzheimer's Disease. *Biomarkers in Neuropsychiatry* 3, 100028. doi:10.1016/j.bionps.2020.100028

Bateman, R. J., Xiong, C., Benzinger, T. L. S., Fagan, A. M., Goate, A., Fox, N. C., et al. (2012). Clinical and Biomarker Changes in Dominantly Inherited Alzheimer's Disease. *N. Engl. J. Med.* 367, 795–804. doi:10.1056/nejmoa1202753

Bhagwat, N., Viviano, J. D., Voineskos, A. N., and Chakravarty, M. M.; Alzheimer's Disease Neuroimaging Initiative (2018). Modeling and Prediction of Clinical Symptom Trajectories in Alzheimer's Disease Using Longitudinal Data. *Plos Comput. Biol.* 14, e1006376. doi:10.1371/journal.pcbi.1006376

Blanche, P., Proust-Lima, C., Loubère, L., Berr, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2015). Quantifying and Comparing Dynamic Predictive Accuracy of Joint Models for Longitudinal Marker and Time-To-Event in Presence of Censoring and Competing Risks. *Biom* 71, 102–113. doi:10.1111/biom.12232

Braak, H., and Braak, E. (1997). Frequency of Stages of Alzheimer-Related Lesions in Different Age Categories. *Neurobiol. Aging* 18, 351–357. doi:10.1016/s0197-4580(97)00056-0

Braak, H., and Braak, E. (1991). Neuropathological Stageing of Alzheimer-Related Changes. *Acta Neuropathol.* 82, 239–259. doi:10.1007/bf00308809

Braak, H., and Del Tredici, K. (2012). Alzheimer's Disease: Pathogenesis and Prevention. *Alzheimer's Demen.* 8, 227–233. doi:10.1016/j.jalz.2012.01.011

Brand, L., Nichols, K., Wang, H., Shen, L., and Huang, H. (2020). Joint Multi-Modal Longitudinal Regression and Classification for Alzheimer's Disease Prediction. *IEEE Trans. Med. Imaging* 39, 1845–1855. doi:10.1109/tmi.2019.2958943

Chen, G., Shu, H., Chen, G., Ward, B. D., Antuono, P. G., Zhang, Z., et al. (2016). Staging Alzheimer's Disease Risk by Sequencing Brain Function and Structure, Cerebrospinal Fluid, and Cognition Biomarkers. *Jad* 54, 983–993. doi:10.3233/jad-160537

Cheng, B., Liu, M., Liu, M., Shen, D., Li, Z., and Zhang, D. (2017). Multi-Domain Transfer Learning for Early Diagnosis of Alzheimer's Disease. *Neuroinform* 15, 115–132. doi:10.1007/s12021-016-9318-5

de Jong, J., Emon, M. A., Wu, P., Karki, R., Sood, M., Godard, P., et al. (2019). Deep Learning for Clustering of Multivariate Clinical Patient Trajectories with Missing Values. *GigaScience* 8, giz134. doi:10.1093/gigascience/giz134

Dong, A., Toledo, J. B., Honnorat, N., Doshi, J., Varol, E., Sotiras, A., et al. (2017). Heterogeneity of Neuroanatomical Patterns in Prodromal Alzheimer's Disease: Links to Cognition, Progression and Biomarkers. *Brain* 140, 735–747. doi:10.1093/brain/aww319

Dong, A., Honnorat, N., Gaonkar, B., and Davatzikos, C. (2016). CHIMERA: Clustering of Heterogeneous Disease Effects via Distribution Matching of Imaging Patterns. *IEEE Trans. Med. Imaging* 35, 612–621. doi:10.1109/tmi.2015.2487423

Donohue, M. C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R. G., Raman, R., Gamst, A. C., et al. (2014). Estimating Long-Term Multivariate Progression from Short-Term Data. *Alzheimer's Demen.* 10, S400–S410. doi:10.1016/j.jalz.2013.10.003

European Commission/EACEA/Eurydice (2018). *The Structure of the European Education Systems 2018/19: Schematic Diagrams. Eurydice Facts and Figures*. Luxembourg: Publications Office of the European Union Luxembourg.

Fiot, J.-B., Raguet, H., Risser, L., Cohen, L. D., Fripp, J., Vialard, F.-X., et al. (2014). Longitudinal Deformation Models, Spatial Regularizations and Learning Strategies to Quantify Alzheimer's Disease Progression. *NeuroImage: Clin.* 4, 718–729. doi:10.1016/j.nicl.2014.02.002

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., et al. (2012). An Event-Based Model for Disease Progression and its Application in Familial Alzheimer's Disease and Huntington's Disease. *NeuroImage* 60, 1880–1889. doi:10.1016/j.neuroimage.2012.01.062

Gauthier, S., Albert, M., Fox, N., Goedert, M., Kivipelto, M., Mestre-Ferrandiz, J., et al. (2016). Why Has Therapy Development for Dementia Failed in the Last Two Decades?. *Alzheimer's Demen.* 12, 60–64. doi:10.1016/j.jalz.2015.12.003

Geifman, N., Kennedy, R. E., Schneider, L. S., Buchan, I., and Brinton, R. D. (2018). Data-driven Identification of Endophenotypes of Alzheimer's Disease Progression: Implications for Clinical Trials and Therapeutic Interventions. *Alzheimers Res. Ther.* 10, 1–7. doi:10.1186/s13195-017-0332-0

Golriz Khatami, S., Robinson, C., Birkenbihl, C., Domingo-Fernández, D., Hoyt, C. T., and Hofmann-Apitius, M. (2020). Challenges of Integrative Disease Modeling in Alzheimer's Disease. *Front. Mol. Biosci.* 6, 158. doi:10.3389/fmolb.2019.00158

Goudie, R. J. B., Presanis, A. M., Lunn, D., De Angelis, D., and Wernisch, L. (2019). Joining and Splitting Models with Markov Melding. *Bayesian Anal.* 14, 81–109. doi:10.1214/18-BA1104

Goyal, D., Tjandra, D., Migrino, R. Q., Giordani, B., Syed, Z., Wiens, J., et al. (2018). Characterizing Heterogeneity in the Progression of Alzheimer's Disease Using Longitudinal Clinical and Neuroimaging Biomarkers. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 10, 629–637. doi:10.1016/j.dadm.2018.06.007

Hall, C. B., Lipton, R. B., Sliwinski, M., and Stewart, W. F. (2000). A Change point Model for Estimating the Onset of Cognitive Decline in Preclinical Alzheimer's Disease. *Statist. Med.* 19, 1555–1566. doi:10.1002/(sici)1097-0258(20000615/30)19:11/12<1555::aid-sim445>3.0.co;2-3

Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J. Q., Bittner, T., et al. (2018). CSF Biomarkers of Alzheimer's Disease concord with Amyloid-β PET and Predict Clinical Progression: A Study of Fully Automated Immunoassays in BioFINDER and ADNI Cohorts. *Alzheimer's Demen.* 14, 1470–1481. doi:10.1016/j.jalz.2018.01.010

Hardy, J., and Selkoe, D. J. (2002). The Amyloid Hypothesis of Alzheimer's Disease: Progress and Problems on the Road to Therapeutics. *Science* 297, 353–356. doi:10.1126/science.1072994

Hubbard, R. A., and Zhou, X. H. (2011). A Comparison of Non-homogeneous Markov Regression Models with Application to Alzheimer's Disease Progression. *J. Appl. Stat.* 38, 2313–2326. doi:10.1080/02664763.2010.547567

Iddi, S., Li, D., Aisen, P. S., Rafii, M. S., Thompson, W. K., and Donohue, M. C. (2019). Predicting the Course of Alzheimer's Progression. *Brain Inform.* 6, 1–18. doi:10.1186/s40708-019-0099-0

Jack, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., et al. (2013). Tracking Pathophysiological Processes in Alzheimer's Disease: an Updated Hypothetical Model of Dynamic Biomarkers. *Lancet Neurol.* 12, 207–216. doi:10.1016/s1474-4422(12)70291-0

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical Model of Dynamic Biomarkers of the Alzheimer's Pathological cascade. *Lancet Neurol.* 9, 119–128. doi:10.1016/s1474-4422(09)70299-6

Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., et al. (2012). A Computational Neurodegenerative Disease Progression Score: Method and Results with the Alzheimer's Disease Neuroimaging Initiative Cohort. *NeuroImage* 63, 1478–1486. doi:10.1016/j.neuroimage.2012.07.059

Kaufman, L., and Rousseeuw, P. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis*. John Wiley & Sons.

Khanna, S., Domingo-Fernández, D., Iyappan, A., Emon, M. A., Hofmann-Apitius, M., and Fröhlich, H. (2018). Using Multi-Scale Genetic, Neuroimaging and Clinical Data for Predicting Alzheimer's Disease and Reconstruction of Relevant Biological Mechanisms. *Sci. Rep.* 8, 1–13. doi:10.1038/s41598-018-29433-3

Kulason, S., Xu, E., Tward, D. J., Bakker, A., Albert, M., Younes, L., et al. (2020). Entorhinal and Transentorhinal Atrophy in Preclinical Alzheimer's Disease. *Front. Neurosci.* 14, 804. doi:10.3389/fnins.2020.00804

Lai, D., Xu, H., Koller, D., Foroud, T., and Gao, S. (2016). A Multivariate Finite Mixture Latent Trajectory Model with Application to Dementia Studies. *J. Appl. Stat.* 43, 2503–2523. doi:10.1080/02664763.2016.1141181

Lei, B., Yang, M., Yang, P., Zhou, F., Hou, W., Zou, W., et al. (2020). Deep and Joint Learning of Longitudinal Data for Alzheimer's Disease Prediction. *Pattern Recognition* 102, 107247. doi:10.1016/j.patcog.2020.107247

Li, D., Iddi, S., Thompson, W. K., Rafii, M. S., Aisen, P. S., Donohue, M. C., et al. (2018). Bayesian Latent Time Joint Mixed-effects Model of Progression in the Alzheimer's Disease Neuroimaging Initiative. *Alzheimer's Demen. Diagn. Assess. Dis. Monit.* 10, 657–668. doi:10.1016/j.dadm.2018.07.008

Li, K., Chan, W., Doody, R. S., Quinn, J., and Luo, S. (2017). Prediction of Conversion to Alzheimer's Disease with Longitudinal Measures and Time-To-Event Data. *Jad* 58, 361–371. doi:10.3233/jad-161201

Li, K., and Luo, S. (2019). Dynamic Predictions in Bayesian Functional Joint Models for Longitudinal and Time-To-Event Data: An Application to Alzheimer's Disease. *Stat. Methods Med. Res.* 28, 327–342. doi:10.1177/0962280217722177

Lin, J., Li, K., and Luo, S. (2021). Functional Survival Forests for Multivariate Longitudinal Outcomes: Dynamic Prediction of Alzheimer's Disease Progression. *Stat. Methods Med. Res.* 30, 99–111. doi:10.1177/0962280220941532

Liu, W., Zhang, B., Zhang, Z., and Zhou, X.-H. (2013). Joint Modeling of Transitional Patterns of Alzheimer's Disease. *PloS ONE* 8, e75487. doi:10.1371/journal.pone.0075487

Liverani, S., Hastie, D. I., Azizi, L., Papathomas, M., and Richardson, S. (2015). PReMiuM: An R Package for Profile Regression Mixture Models Using Dirichlet Processes. *J. Stat. Softw.* 64, 1–30. doi:10.18637/jss.v064.i07

Livingston, G., Sommerlad, A., Orgeta, V., Costafreda, S. G., Huntley, J., Ames, D., et al. (2017). Dementia Prevention, Intervention, and Care. *The Lancet* 390, 2673–2734. doi:10.1016/s0140-6736(17)31363-6

Lorenzi, M., Filippone, M., Frisoni, G. B., Alexander, D. C., and Ourselin, S.; Alzheimer's Disease Neuroimaging Initiative (2019). Probabilistic Disease Progression Modeling to Characterize Diagnostic Uncertainty: Application to Staging and Prediction in Alzheimer's Disease. *NeuroImage* 190, 56–68. doi:10.1016/j.neuroimage.2017.08.059

Marioni, R. E., Proust-Lima, C., Amieva, H., Brayne, C., Matthews, F. E., Dartigues, J.-F., et al. (2014). Cognitive Lifestyle Jointly Predicts Longitudinal Cognitive Decline and Mortality Risk. *Eur. J. Epidemiol.* 29, 211–219. doi:10.1007/s10654-014-9881-8

Martí-Juan, G., Sanroma, G., and Piella, G.; Alzheimer's Disease Neuroimaging Initiative and Alzheimer's Disease Metabolomics Consortium (2019). Revealing Heterogeneity of Brain Imaging Phenotypes in Alzheimer's Disease Based on Unsupervised Clustering of Blood Marker Profiles. *PloS ONE* 14, e0211121. doi:10.1371/journal.pone.0211121

Martí-Juan, G., Sanroma-Guell, G., and Piella, G. (2020). A Survey on Machine and Statistical Learning for Longitudinal Analysis of Neuroimaging Data in Alzheimer's Disease. *Comp. Methods Programs Biomed.* 189, 105348. doi:10.1016/j.cmpb.2020.105348

Molitor, J., Papathomas, M., Jerrett, M., and Richardson, S. (2010). Bayesian Profile Regression with an Application to the National Survey of Children's Health. *Biostatistics* 11, 484–498. doi:10.1093/biostatistics/kxq013

Oxtoby, N. P., Young, A. L., Cash, D. M., Benzinger, T. L. S., Fagan, A. M., Morris, J. C., et al. (2018). Data-driven Models of Dominantly-Inherited Alzheimer's Disease Progression. *Brain* 141, 1529–1544. doi:10.1093/brain/awy050

Philipps, V., Amieva, H., Andrieu, S., Dufouil, C., Berr, C., Dartigues, J.-F., et al. (2014). Normalized Mini-Mental State Examination for Assessing Cognitive Change in Population-Based Brain Aging Studies. *Neuroepidemiology* 43, 15–25. doi:10.1159/000365637

Proust-Lima, C., Amieva, H., and Jacqmin-Gadda, H. (2013). Analysis of Multivariate Mixed Longitudinal Data: a Flexible Latent Process Approach. *Br. J. Math. Stat. Psychol.* 66, 470–487. doi:10.1111/bmsp.12000

Proust-Lima, C., Philipps, V., and Dartigues, J. F. (2019). A Joint Model for Multiple Dynamic Processes and Clinical Endpoints: Application to Alzheimer's Disease. *Stat. Med.* 38, 4702–4717. doi:10.1002/sim.8328

Proust-Lima, C., Philipps, V., and Liquet, B. (2017). Estimation of Extended Mixed Models Using Latent Classes and Latent Processes: The R Package lcmm. *J. Stat. Softw.* 78, 1–56. doi:10.18637/jss.v078.i02

Proust-Lima, C., Dartigues, J.-F., and Jacqmin-Gadda, H. (2016). Joint Modeling of Repeated Multivariate Cognitive Measures and Competing Risks of Dementia and Death: a Latent Process and Latent Class Approach. *Statist. Med.* 35, 382–398. doi:10.1002/sim.6731

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing.

Racine, A. M., Koscik, R. L., Berman, S. E., Nicholas, C. R., Clark, L. R., Okonkwo, O. C., et al. (2016). Biomarker Clusters Are Differentially Associated with Longitudinal Cognitive Decline in Late Midlife. *Brain* 139, 2261–2274. doi:10.1093/brain/aww142

Raket, L. L. (2020). Statistical Disease Progression Modeling in Alzheimer Disease. *Front. Big Data* 3, 24. doi:10.3389/fdata.2020.00024

Ritchie, C. W., Muniz-Terrera, G., Kivipelto, M., Solomon, A., Tom, B., and Molinuevo, J. L. (2020). The European Prevention of Alzheimer's Dementia (EPAD) Longitudinal Cohort Study: Baseline Data Release V500.0. *J. Prev. Alzheimers Dis.* 7, 8–13. doi:10.14283/jpad.2019.46

Ritchie, C. W., Molinuevo, J. L., Truyen, L., Satlin, A., Van der Geyten, S., and Lovestone, S. (2016). Development of Interventions for the Secondary Prevention of Alzheimer's Dementia: the European Prevention of Alzheimer's Dementia (EPAD) Project. *The Lancet Psychiatry* 3, 179–186. doi:10.1016/s2215-0366(15)00454-x

Robitaille, A., van den Hout, A., Machado, R. J. M., Bennett, D. A., Čukić, I., Deary, I. J., et al. (2018). Transitions across Cognitive States and Death Among Older Adults in Relation to Education: A Multistate Survival Model Using Data from Six Longitudinal Studies. *Alzheimer's Demen.* 14, 462–472. doi:10.1016/j.jalz.2017.10.003

Rouanet, A., Joly, P., Dartigues, J. F., Proust-Lima, C., and Jacqmin-Gadda, H. (2016). Joint Latent Class Model for Longitudinal Data and Interval-censored Semi-competing Events: Application to Dementia. *Biom* 72, 1123–1135. doi:10.1111/biom.12530

Rousseeuw, P. J. (1987). Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *J. Comput. Appl. Math.* 20, 53–65. doi:10.1016/0377-0427(87)90125-7

Samtani, M., Raghavan, N., Novak, G., Nandy, P., and Narayan, V. A. (2014). Disease Progression Model for Clinical Dementia Rating–Sum of Boxes in Mild Cognitive Impairment and Alzheimer's Subjects from the Alzheimer's Disease Neuroimaging Initiative. *Ndt* 10, 929. doi:10.2147/ndt.s62323

Schindler, S. E., Gray, J. D., Gordon, B. A., Xiong, C., Batrla-Utermann, R., Quan, M., et al. (2018). Cerebrospinal Fluid Biomarkers Measured by Elecsys Assays Compared to Amyloid Imaging. *Alzheimer's Demen.* 14, 1460–1469. doi:10.1016/j.jalz.2018.01.013

Schmidt-Richberg, A., Ledig, C., Guerrero, R., Molina-Abril, H., Frangi, A., Rueckert, D., et al. (2016). Learning Biomarker Models for Progression Estimation of Alzheimer's Disease. *PloS ONE* 11, e0153040. doi:10.1371/journal.pone.0153040

Segalas, C., Helmer, C., and Jacqmin-Gadda, H. (2020). A Curvilinear Bivariate Random Changepoint Model to Assess Temporal Order of Markers. *Stat. Methods Med. Res.* 29, 2481–2492. doi:10.1177/0962280219898719

Sethuraman, J. (1994). A Constructive Definition of Dirichlet Priors. *Stat. Sinica* 4, 639–650.

Shi, J., Sabbagh, M. N., and Vellas, B. (2020). Alzheimer's Disease beyond Amyloid: Strategies for Future Therapeutic Interventions. *BMJ* 371, m3684. doi:10.1136/bmj.m3684

Solomon, A., Kivipelto, M., Molinuevo, J. L., Tom, B., and Ritchie, C. W. (2018). European Prevention of Alzheimer's Dementia Longitudinal Cohort Study (EPAD LCS): Study Protocol. *BMJ Open* 8, e021017. doi:10.1136/bmjopen-2017-021017

ten Kate, M., Ingala, S., Schwarz, A. J., Fox, N. C., Chételat, G., van Berckel, B. N. M., et al. (2018). Secondary Prevention of Alzheimer's Dementia: Neuroimaging Contributions. *Alzheimers Res. Ther.* 10, 1–21. doi:10.1186/s13195-018-0438-z

ten Kate, M., Dicks, E., Visser, P. J., van der Flier, W. M., Teunissen, C. E., Barkhof, F., et al. (2018). Atrophy Subtypes in Prodromal Alzheimer's Disease Are Associated with Cognitive Decline. *Brain* 141, 3443–3456. doi:10.1093/brain/awy264

van den Hout, A. (2016). *Multi-state Survival Models for Interval-Censored Data*. New York: CRC Press.

Villeneuve, S. C., Houot, M., Cacciamani, F., Verrijp, M., Dubois, B., Sikkes, S., et al. (2019). Latent Class Analysis Identifies Functional Decline with Amsterdam IADL in Preclinical Alzheimer's Disease. *Alzheimer's Demen. Translational Res. Clin. Interventions* 5, 553–562. doi:10.1016/j.trci.2019.08.009

Vos, S. J., Xiong, C., Visser, P. J., Jasielec, M. S., Hassenstab, J., Grant, E. A., et al. (2013). Preclinical Alzheimer's Disease and its Outcome: a Longitudinal Cohort Study. *Lancet Neurol.* 12, 957–965. doi:10.1016/s1474-4422(13)70194-7

Wang, G., Xiong, C., McDade, E. M., Hassenstab, J., Aschenbrenner, A. J., Fagan, A. M., et al. (2018). Simultaneously Evaluating the Effect of Baseline Levels and Longitudinal Changes in Disease Biomarkers on Cognition in Dominantly Inherited Alzheimer's Disease. *Alzheimer's Demen. Translational Res. Clin. Interventions* 4, 669–676. doi:10.1016/j.trci.2018.10.009

Watts, G. (2018). Prospects for Dementia Research. *The Lancet* 391, 416. doi:10.1016/s0140-6736(18)30190-9

Wei, S., and Kryscio, R. J. (2016). Semi-Markov Models for Interval Censored Transient Cognitive States with Back Transitions and a Competing Risk. *Stat. Methods Med. Res.* 25, 2909–2924. doi:10.1177/0962280214534412

Williams, O. A., An, Y., Armstrong, N. M., Kitner-Triolo, M., Ferrucci, L., and Resnick, S. M. (2020). Profiles of Cognitive Change in Preclinical and Prodromal Alzheimer's Disease Using Change-Point Analysis. *Jad* 75, 1169–1180. doi:10.3233/jad-191268

Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., et al. (2016). Defeating Alzheimer's Disease and Other Dementias: a Priority for European Science and Society. *Lancet Neurol.* 15, 455–532. doi:10.1016/s1474-4422(16)00062-4

Wolz, R., Aljabar, P., Hajnal, J. V., Hammers, A., and Rueckert, D. (2010). LEAP: Learning Embeddings for Atlas Propagation. *NeuroImage* 49, 1316–1325. doi:10.1016/j.neuroimage.2009.09.069

Wu, Y., Zhang, X., He, Y., Cui, J., Ge, X., Han, H., et al. (2020). Predicting Alzheimer's Disease Based on Survival Data and Longitudinally Measured Performance on Cognitive and Functional Scales. *Psychiatry Res.* 291, 113201. doi:10.1016/j.psychres.2020.113201

Yiannopoulou, K. G., and Papageorgiou, S. G. (2020). Current and Future Treatments in Alzheimer Disease: An Update. *J. Cent. Nervous Syst. Dis.* 12, 1–12. doi:10.1177/1179573520907397

Younes, L., Albert, M., Moghekar, A., Soldan, A., Pettigrew, C., and Miller, M. I. (2019). Identifying Changepoints in Biomarkers during the Preclinical Phase of Alzheimer's Disease. *Front. Aging Neurosci.* 11, 74. doi:10.3389/fnagi.2019.00074

Young, A. L., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., Firth, N. C., et al. (2018). Uncovering the Heterogeneity and Temporal Complexity of Neurodegenerative Diseases with Subtype and Stage Inference. *Nat. Commun.* 9, 1–16. doi:10.1038/s41467-018-05892-0

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A Data-Driven Model of Biomarker Changes in Sporadic Alzheimer's Disease. *Brain* 137, 2564–2577. doi:10.1093/brain/awu176

Zhang, L., Lim, C. Y., Maiti, T., Li, Y., Choi, J., Bozoki, A., et al. (2019). Analysis of Conversion of Alzheimer's Disease Using a Multi-State Markov Model. *Stat. Methods Med. Res.* 28, 2801–2819. doi:10.1177/0962280218786525

Zhang, X., Mormino, E. C., Sun, N., Sperling, R. A., Sabuncu, M. R., Yeo, B. T. T., et al. (2016). Bayesian Model Reveals Latent Atrophy Factors with Dissociable Cognitive Trajectories in Alzheimer's Disease. *Proc. Natl. Acad. Sci. USA* 113, E6535–E6544. doi:10.1073/pnas.1611073113

Zhang, X., Yang, Y., Li, T., Zhang, Y., Wang, H., and Fujita, H. (2021). CMC: A Consensus Multi-View Clustering Model for Predicting Alzheimer's Disease Progression. *Comp. Methods Programs Biomed.* 199, 105895. doi:10.1016/j.cmpb.2020.105895

Author Disclaimer: The views expressed are those of the author(s) and not necessarily those of the NHS, the NIHR or the Department of Health and Social Care.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 13 October 2021
doi: 10.3389/frai.2021.549255

Artificial Intelligence to Analyze the Cortical Thickness Through Age

Sergio Ledesma[1,2], *Mario-Alberto Ibarra-Manzano*[2], *Dora-Luz Almanza-Ojeda*[2], *Pascal Fallavollita*[1] *and Jason Steffener*[1]*

[1]Faculty of Health Sciences, University of Ottawa, Ottawa, ON, Canada, [2]School of Engineering, University of Guanajuato, Guanajuato, Mexico

In this study, Artificial Intelligence was used to analyze a dataset containing the cortical thickness from 1,100 healthy individuals. This dataset had the cortical thickness from 31 regions in the left hemisphere of the brain as well as from 31 regions in the right hemisphere. Then, 62 artificial neural networks were trained and validated to estimate the number of neurons in the hidden layer. These neural networks were used to create a model for the cortical thickness through age for each region in the brain. Using the artificial neural networks and kernels with seven points, numerical differentiation was used to compute the derivative of the cortical thickness with respect to age. The derivative was computed to estimate the cortical thickness speed. Finally, color bands were created for each region in the brain to identify a positive derivative, that is, a part of life with an increase in cortical thickness. Likewise, the color bands were used to identify a negative derivative, that is, a lifetime period with a cortical thickness reduction. Regions of the brain with similar derivatives were organized and displayed in clusters. Computer simulations showed that some regions exhibit abrupt changes in cortical thickness at specific periods of life. The simulations also illustrated that some regions in the left hemisphere do not follow the pattern of the same region in the right hemisphere. Finally, it was concluded that each region in the brain must be dynamically modeled. One advantage of using artificial neural networks is that they can learn and model non-linear and complex relationships. Also, artificial neural networks are immune to noise in the samples and can handle unseen data. That is, the models based on artificial neural networks can predict the behavior of samples that were not used for training. Furthermore, several studies have shown that artificial neural networks are capable of deriving information from imprecise data. Because of these advantages, the results obtained in this study by the artificial neural networks provide valuable information to analyze and model the cortical thickness.

Keywords: modeling, cortical thickness, artificial neural network, derivative, changes with age, adaptive models, neuroimaging

OPEN ACCESS

Edited by:
Holger Fröhlich, Fraunhofer Institute for Algorithms and Scientific Computing (FHG), Germany

Reviewed by:
Shailesh Tripathi, Tampere University of Technology, Finland
Ibrahem Kandel, Universidade NOVA de Lisboa, Portugal

***Correspondence:**
Jason Steffener
jason.steffener@uOttawa.ca

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *09 June 2020*
Accepted: *30 August 2021*
Published: *13 October 2021*

Citation:
Ledesma S, Ibarra-Manzano M-A, Almanza-Ojeda D-L, Fallavollita P and Steffener J (2021) Artificial Intelligence to Analyze the Cortical Thickness Through Age. Front. Artif. Intell. 4:549255. doi: 10.3389/frai.2021.549255

1 BACKGROUND

In the last few years, machine learning techniques have been used in common applications (Alpaydin, 2016). In this paper, we use one technique from Artificial Intelligence to analyze the progress of the cortical thickness with age. This study includes data from 1,100 healthy individuals. The cortical thickness was measured using FreeSurfer which is a fully automated software for

measuring several parameters in the brain including neuroanatomic volume and cortical thickness (McCarthy et al., 2015).

Several studies illustrate the relevance of the analysis of the cortical thickness through the life span. For instance, the authors in (Steffener et al., 2016) indicate that brain aging can be analyzed taking into consideration the inevitable and universal effects of advancing age and the effects resulting from a lifetime of exposures. These effects and a decreased cortical thickness in some regions of the brain may be related to some mental disorders or cognitive decline (Fouche et al., 2017; Razlighi et al., 2017). Thus, some studies have indicated correlations between disease states and cortical thickness, see the references in (Scott et al., 2009).

In the state of the art, there are many studies about the modeling of changes in the cortical thickness. The authors in (Scott et al., 2009) propose a voxel-based method to measure the cortical thickness utilizing inversion recovery anatomical magnetic resonance images. Churchwell et al. use separate hierarchical multiple regressions to analyze changes with age in the cortex thickness in specific zones in the brain (Churchwell and Yurgelun-Todd, 2013). Additionally, it has been suggested that brain aging is a process influenced by degenerative and restorative activities (Fjell et al., 2014). Consequently, the resulting process can be linear and non-linear. Similarly, it has been proposed that cortical thickness changes follow non-linear patterns across childhood and adolescence, and these changes vary to some degree by cortical region (Wierenga et al., 2014; Piccolo et al., 2016; Sowell et al., 2007).

In this sense, the thinning of the cortical thickness has been analyzed. For instance, Tamnes et al. describe the age-related changes in cortical thickness, their findings revealed regional age-related cortical thinning (Tamnes et al., 2010), see also (Salat et al., 2004). The authors in (McGinnis et al., 2011) analyze the thinning of the cerebral cortex in different regions of the brain in the course of aging. Chen et al. demonstrate age-related alterations in the modular organization of the human brain structural networks using regional cortical thickness measurements (Chen et al., 2011). Lemaitre et al. use linear regressions of age, their studies indicate an associated global age-related reduction in cortical thickness, surface area and volume (Lemaitre et al., 2012). On the other hand, it has been indicated that cortical surface area is an increasingly used brain morphology metric that is ontogenetically and phylogenetically distinct from the cortical thickness and offers a separate index of neuro-development and disease (Winkler et al., 2018).

2 ARTIFICIAL NEURAL NETWORKS

An artificial neural network is a computational technique motivated by a specific behavior found in the brain (Marsland, 2015). A neural network is composed of basic units of processing called neurons. Inside the network, the neurons are organized in layers. Artificial neural networks are used for: image classification, image processing, signal processing, prediction, pattern recognition, function approximation, and other applications (Jin et al., 2017; Jordan and Mitchell, 2015). From a practical point of view, artificial neural networks can be used to create a model using only a set of data samples (Russell and Norvig, 2020; Masters, 2015). The main advantage of using an artificial neural network to model the cortical thickness is that the network creates the model that best fits the patterns in the data. In other words, an artificial neural network is capable of learning and modeling non-linear and complex relationships. Additionally, the neural network is immune to noise in the data samples and can infer unseen relationships on unseen data. Therefore, the models obtained are able to generalize and predict on unseen data. Furthermore, research has shown that artificial neural networks have a great capability of deriving information from complex or imprecise data.

3 DATASET DESCRIPTION

The simulations in this study were performed using a dataset with information from approximately 1,100 healthy individuals. This dataset was built by combining data from four different common datasets: IXI, MMRR, NKI, and OASIS. **Table 1** includes a sample from one patient of the cortical thickness for each dataset. These datasets are briefly discussed next.

3.1 IXI Dataset

This dataset contains approximately 600 magnetic resonance images from normal and good health individuals. The data was collected at three different hospitals in London: Hammersmith hospital, Guy's hospital and the Institute of Psychiatry. The IXI dataset was prepared during the project called Information eXtraction from Images, (Information eXtraction from Images, 2019).

3.2 MMRR Dataset

The Multi-Modal MRI Reproducibility Resource dataset was built using information from 21 healthy volunteers. In the MMRR dataset, all volunteers did not have a history of neurological conditions, and therefore, all of them were used in this study. This dataset has 42 records and each record includes information from a 1-h scan session (Landman et al., 2011).

3.3 NKI Dataset

The Nathan Klein Institute - Rockland Sample (NKI-RS) is an attempt to create a large-scale community sample. This dataset includes data from different types of assessments including advanced neuroimaging. The dataset has 186 T1-weighted images from 99 males and 87 females.

3.4 OASIS Dataset

The Open Access Series of Imaging Studies dataset is a set of magnetic resonance images collected from 416 individuals between the ages of 18–96 years (Marcus et al., 2007). This dataset is public and can be used for research. As this study focuses only on healthy individuals, data coming from patients with a mental disease was discarded, and therefore, not used.

TABLE 1 | Cortical thickness in millimeters from one person in each database.

Database	IXI		MMRR		NKI		OASIS	
Age (years)	39		25		41		74	
	Left	**Right**	**Left**	**Right**	**Left**	**Right**	**Left**	**Right**
Caudal anterior cingulate	2.432	2.395	2.981	3.201	2.344	2.545	2.7	2.694
Caudal middle frontal	2.23	2.326	2.634	2.578	2.516	2.422	2.351	2.413
Cuneus	1.895	1.663	1.918	1.761	1.935	1.874	1.682	1.805
Entorhinal	3.356	3.728	4.093	3.868	2.808	2.958	2.876	3.053
Fusiform	2.486	2.558	2.657	2.67	2.457	2.538	2.274	2.199
Inferior parietal	2.426	2.356	2.307	2.303	2.338	2.413	2.221	2.267
Inferior temporal	2.892	2.751	2.777	2.832	2.509	2.519	2.57	2.205
Isthmus cingulate	2.214	2.086	2.702	2.38	2.222	2.356	2.031	2.35
Lateral occipital	2.017	2.097	1.863	1.962	2.005	2.066	2.085	2.001
Lateral orbitofrontal	2.522	2.795	3.085	2.95	2.679	2.497	2.538	2.604
Lingual	1.774	1.762	2.096	2.086	1.961	1.911	1.784	1.837
Medial orbitofrontal	2.53	2.444	2.701	2.628	2.633	2.414	2.159	2.553
Middle temporal	2.856	2.825	2.792	2.845	2.607	2.716	2.561	2.548
Parahippocampal	2.456	2.509	3.339	3.143	2.787	2.608	2.035	2.496
Paracentral	2.108	2	2.579	2.395	2.209	2.253	2.214	2.136
Pars opercularis	2.665	2.307	2.69	2.768	2.549	2.635	2.456	2.528
Pars orbitalis	2.464	2.529	2.893	2.771	2.45	2.332	2.308	2.612
Pars triangularis	2.243	2.4	2.533	2.431	2.287	2.364	2.077	2.243
Pericalcarine	1.441	1.308	1.528	1.642	1.58	1.554	1.482	1.454
Postcentral	1.98	1.901	2.349	2.262	2.144	2.127	2.094	2.039
Posterior cingulate	2.397	2.311	2.79	2.655	2.282	2.229	2.234	2.432
Precentral	2.28	2.339	2.248	2.344	2.574	2.449	2.317	2.231
Precuneus	2.28	2.231	2.625	2.438	2.309	2.22	2.285	2.126
Rostral anterior cingulate	2.69	2.899	3.118	3.406	2.894	2.531	3.041	2.908
Rostral middle frontal	2.224	2.34	2.383	2.356	2.373	2.266	2.283	2.15
Superior frontal	2.558	2.565	2.674	2.812	2.518	2.483	2.592	2.477
Superior parietal	2.135	1.978	2.198	2.084	2.311	2.201	2.128	2.168
Superior temporal	2.774	2.826	2.824	3.023	2.771	2.774	2.614	2.633
Supramarginal	2.482	2.414	2.577	2.577	2.545	2.478	2.302	2.309
Transverse temporal	1.893	1.968	2.713	2.628	2.332	2.364	2.621	2.285
Insula	3.072	2.749	3.169	3.242	3.01	2.915	2.942	3.049

Consequently, data from only 313 individuals were used for the computer simulations and analysis performed in this work.

4 METHODOLOGY

In this study, the cortical thickness of the images provided in (Tustison et al., 2014) was used for the training and validation of 62 artificial neural networks. The total number of records in this dataset was approximately 1,100. Each record had the sex and age of each individual. Additionally, each record included the values of the cortical thickness in 31 regions in the left hemisphere of the brain and 31 regions in the right hemisphere, see Fischl (2012) and Klein and Tourville (2012).

To create the neural network models, several steps were performed. First, the input data, the age of each person in the dataset, was linearly scaled so that all the values at the input of the network were in the range of −1 to 1. Second, the cortical thickness values were also scaled using a linear transformation so that all target values at the output of the network were in the range of −1 to 1. Third, each neural network was trained in two steps. In the first step, a non-greedy optimization method called simulated annealing was used to find initial values of the weights connecting the neurons in the network. Then, a gradient-based method was used to quickly optimize the values of the weights by moving the weights in the opposite direction of the gradient of the error. Once the networks were trained, we validated the performance of the network by measuring the mean squared error between the predicted value and the observed data from the validation set.

4.1 Training and Validation of the Artificial Neural Networks

Once the dataset was ready, 62 multilayer neural networks were created using the Neural Lab software (Ledesma et al., 2017). All 62 networks had three layers: the input layer, the hidden layer, and the output layer as shown in **Figure 1**. All neurons in the network were designed to use the hyperbolic tangent as their activation functions. The neurons were connected with weights, these are denoted by h and w in **Figure 1**. Each network had one input, the age, and one output, the cortical thickness of one specific region of the brain as in **Figure 1**. Thus, each neural network had one neuron in the output layer. The number of neurons in the hidden layer was iteratively determined as follows. First, the complete dataset with the 1,100 cases was split into two datasets: the training set and the validation set. Second, each network was trained with zero neurons in the

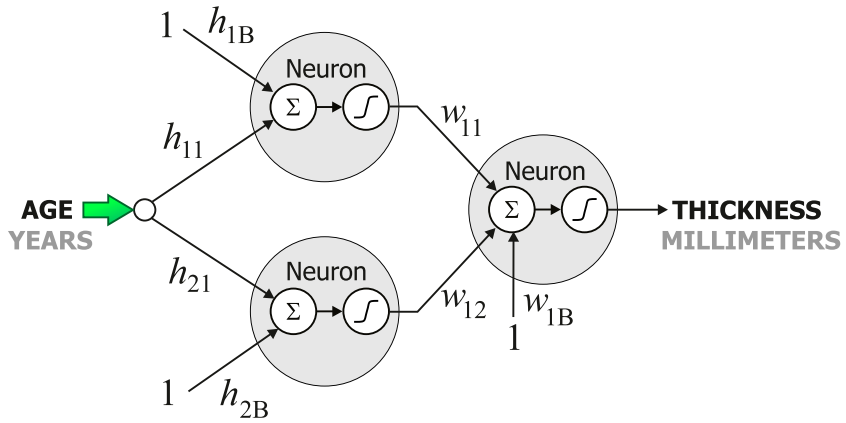


FIGURE 1 | Structure of the artificial neural network used to model the cortical thickness.

TABLE 2 | Methods and parameters used for training.

Simulated annealing	
Initial temperature	15
Final temperature	0.001
Number of temperatures	100
Iterations per temperature	100
Cooling schedule	Linear
Levenberg-Marquardt	
Number of iterations	1,000
Goal (mean squared error)	1×10^{-5}

hidden layer. Both the mean squared error for training and the mean squared error for validation were computed. Then, the number of neurons in the hidden layer was increased by one. Again, the mean squared error for training and the mean squared for validation were computed. This iterative process was stopped when the mean squared error during validation did not decrease. The main conclusion obtained from this iterative process was that only two neurons in the hidden layer were necessary to model the cortical thickness.

In this case, 80% of the cases were included in the training set, and the 20% remaining cases were used to build the validation set. The training of the 62 artificial neural networks was performed in two steps using the parameters shown in **Table 2**. The training of each neural network began using simulated annealing. Then, the method of Levenberg–Marquardt was used to improve the training.

4.2 Derivative Computation

In the field of numerical differentiation, there are some methods to estimate the numerical value of the derivative of a function. One common method to approximate the derivative of a function is based on finite differences. There are three types of differences: forward difference, backward difference, and central difference. These differences are associated with a stencil or kernel. A stencil s (or kernel) is a set of N points that are arranged in the vicinity of a point of interest (Hassan et al., 2012). For instance, the stencil

$$\mathbf{s} = [-1, 0, 1] \tag{1}$$

is used to describe a stencil with three points (N = 3) in the vicinity of the point of interest. The numbers in the stencil indicate the time steps, 0 represents the current value, − 1 represents the previous value, and 1 represents the next value. In general, a stencil with N points is represented as

$$\mathbf{s} = [s_1, s_2, s_3, \cdots s_N]. \tag{2}$$

For instance, when N = 5, the derivative is computed using five points in the vicinity of the point of interest. Consequently, when the value of N is increased, the accuracy of the derivative also increases. However, when working in the upper or lower ends of the data, it is important to use different stencils to compute the derivative for each point. That is, the point of interest must be dynamically located inside the stencil to compensate for the missing data, see (Hassan et al., 2012). For the stencil s in **Equation 2**, the finite difference coefficients c_1, c_2, ⋯, c_N, can be obtained by solving the system of linear equations

$$\begin{pmatrix} (s_1)^0 & (s_2)^0 & \cdots & (s_N)^0 \\ (s_1)^1 & (s_2)^1 & \cdots & (s_N)^1 \\ \vdots & \vdots & \ddots & \vdots \\ (s_1)^{N-1} & (s_2)^{N-1} & \cdots & (s_N)^{N-1} \end{pmatrix} \begin{pmatrix} c_1 \\ c_2 \\ \vdots \\ c_N \end{pmatrix} = d! \begin{pmatrix} \delta_{0,d} \\ \delta_{1,d} \\ \vdots \\ \delta_{N-1,d} \end{pmatrix} \tag{3}$$

where d is the order of derivative and $\delta_{i,j}$ is the Kronecker delta, see (Hassan et al., 2012). The main advantage of using this method is that different stencils can be used to estimate the derivative at different points of interest increasing the accuracy of the computation. It is important to note that **Equation 3** cannot be used to estimate the derivative in a non-differentiable region. However, as it can be seen from databases in the state of the art, changes in the cortical thickness are slow and non-differentiable regions were not found in the four databases used in this study.

5 COMPUTER SIMULATIONS AND RESULTS

The computer simulations performed in **Section 4.1** were used to determine the proper number of neurons in the hidden layer and to validate the performance of the models. However, once the validation process was finished, it was convenient to create new models by performing the training of the networks using all samples in the data set. Therefore, all the 62 artificial neural networks were again trained, but in this case, all the 1,100 cases (instead of only 80% of the cases) were used. The training was performed as before using the parameters in **Table 2**. According to the results of the computer simulations performed in **Section 4.1**, all neural networks had two neurons in the hidden layer.

5.1 Cortical Thickness Progress With Age

As it is well known, artificial neural networks may be used to create a model when there is not a mathematical equation to represent the data (Kelleher et al., 2015; Goodfellow et al., 2016). In this study, artificial neural networks were used to model the changes in cortical thickness in the brain at different ages. Specifically, for each region in the brain, one artificial neural network was used to model the cortical thickness in that region. Thus, a total of 62 artificial neural networks were trained and validated to model the cortical thickness of the brain. There are several approaches that can be used to model the different regions of the brain. For instance, instead of using 62 neural networks, it is possible to design a single neural network with 62 outputs. However, computer simulations showed that the performance of the single neural network was very similar to the performance of the 62 neural networks.

The results of the computer simulations indicated that the mean squared error during the training of the artificial neural networks was from 0.016 to 0.031. During the validation of the models, the computer simulations indicated that the variations between the observed data and the predicted results had errors from 0.016 to 0.033. Finally, to build the models, a new set of artificial neural networks was trained using the whole dataset. In this case, the mean squared error was in the range of 0.017–0.034. To our knowledge, this is the first study to use this type of approach to analyze changes in the cortical thickness.

To ease the presentation of the computer simulations, the models obtained by the artificial neural networks were organized manually in clusters. In this sense, each cluster included those regions which exhibit similar behavior through age. A total of six clusters were created based on the patterns observed in the cortical thickness. We chose this number of clusters because most of the patterns observed in the 62 regions of the brain were represented using only six clusters. However, it is important to mention that if more clusters are used, each cluster will include very few regions. These clusters are described next.

5.1.1 Changes in Cortical Thickness Around 25 years of Age

Figure 2 shows the behavior of the models created by the artificial neural networks in twelve different regions in the brain. Each graph was built using one artificial neural network. All networks in this study had the configuration shown in **Figure 1**. However, each network had a different set of weights, h and w. These weights were adjusted during the training process to model one single region of the brain, and thus, discover and learn hidden patterns in the data. To build the graph, a set of uniformly distributed values for the age was applied to the input of the neural network. Then, an estimate for the cortical thickness in millimeters was produced at the output of the artificial neural network. Finally, the respective input and output values were used to build each graph in **Figures 2–7**.

All regions in **Figure 2** exhibit a similar pattern for the changes in cortical thickness with age. Specifically, all these regions present an abrupt change in the cortical thickness speed around the age of 25 years. This abrupt change is observed by a change in the direction (line slope) of the graph for each region. As it was mentioned before, those regions of the brain with similar behavior in their cortical thickness were manually selected, and then presented in the same figure.

The first row in **Figure 2** displays the cortical thickness in millimeters for the left insula and the right insula as a function of age. From this figure, it can be seen that the thickness of the left insula constantly reduces during the first 20 years of life. A similar behavior is also observed in the right insula. From age 20 to 30, the cortical thickness remains almost constant in these two regions. Then, starting at age 30, the thickness of the left and right insula starts decreasing with age at a low rate. Thus, it can be observed that both regions the left insula and the right insula exhibit a somehow similar pattern for the changes in cortical thickness with age. In the next row in **Figure 2**, the graphs show the cortical thickness models created using the artificial neural networks for the left superior parietal and the right superior parietal. The next row shows the models for the left precentral and right precentral. The next rows in the figure show the behavior of the cortical thickness with age in other regions of the brain; all these regions follow a similar pattern with age. However, it is important to note that the left rostral anterior cingulate and the right rostral anterior cingulate present a more abrupt change at 25 years of age than the other regions in **Figure 2**.

It is important to note that each artificial neural network was trained separately without using data from the same region in the other hemisphere of the brain. However, as it has been concluded by other researchers, some regions in the brain did not present the same behavior for the cortical thickness in both hemispheres. Consequently, some of the graphs in the figures do not present the results for the left hemisphere on the column on the left, and the results for the right hemisphere on the column on the right. For instance, the fifth row in **Figure 2** shows the results for the left transverse temporal and the right caudal anterior cingulate.

5.1.2 Changes in Cortical Thickness Around 40 years of Age

Figure 3 shows eight regions in the brain that have a special behavior in cortical thickness around 40 years of age. The first row in **Figure 3** displays the cortical thickness for the left poscentral and the right poscentral. Observe that both

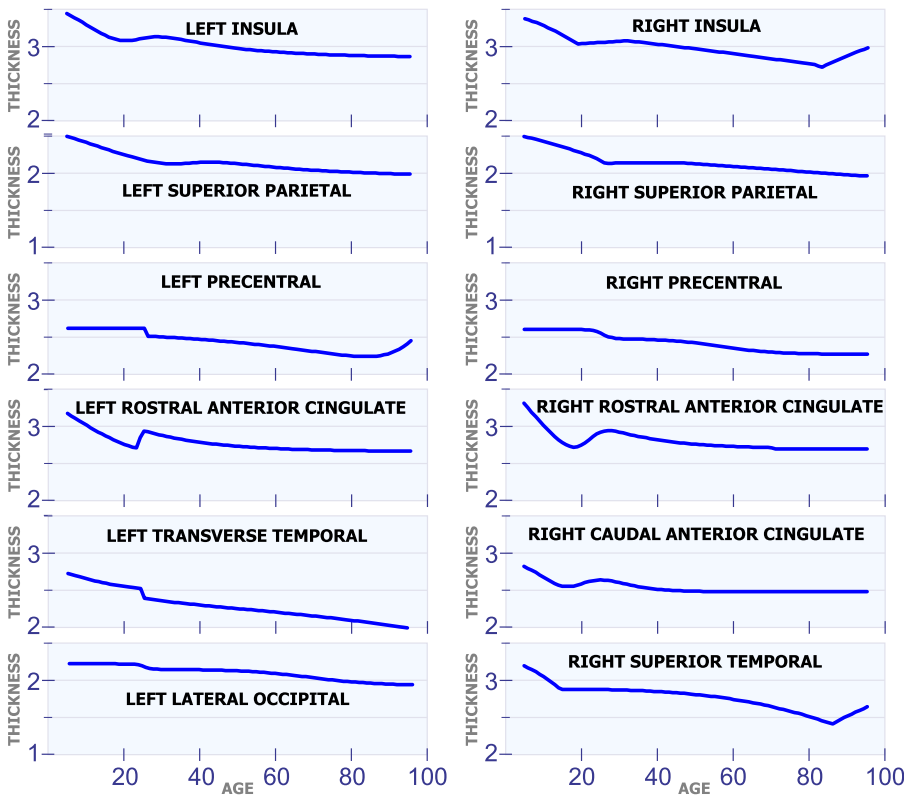


FIGURE 2 | Regions with changes in cortical thickness around 25 years of age.

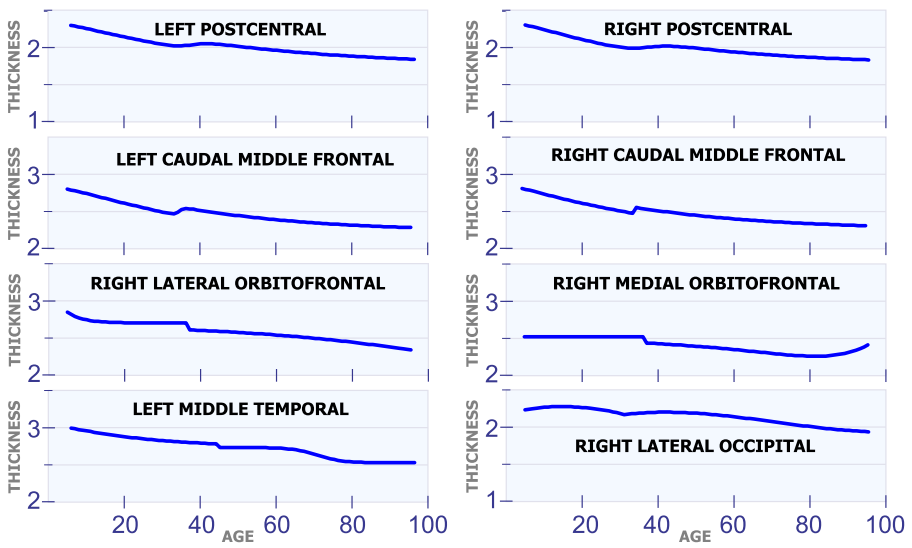


FIGURE 3 | Regions with changes in cortical thickness around 40 years of age.

regions exhibit a constant reduction in cortical thickness during the first 35 years of life. From 35 to 45 years of age, the cortical thickness remains almost constant in both regions. Then, starting at age 45, the cortical thickness begins to slowly decrease. The second row in **Figure 3** shows the model for the left caudal middle frontal and the right caudal middle frontal. For these two regions, it can be observed a sudden and small increase in cortical thickness around age 35. In the

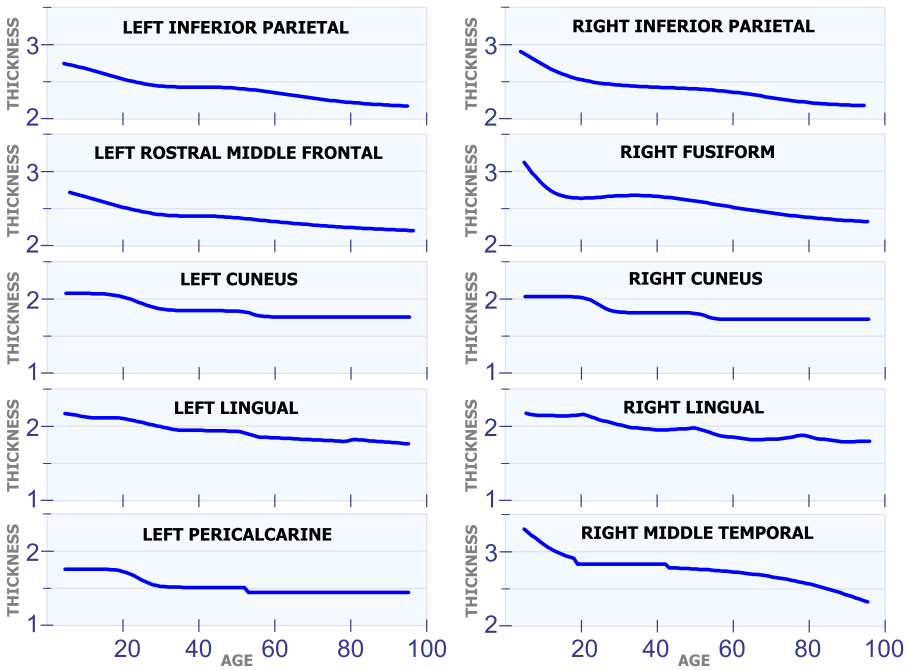


FIGURE 4 | Regions with changes in cortical thickness around 50 years of age.

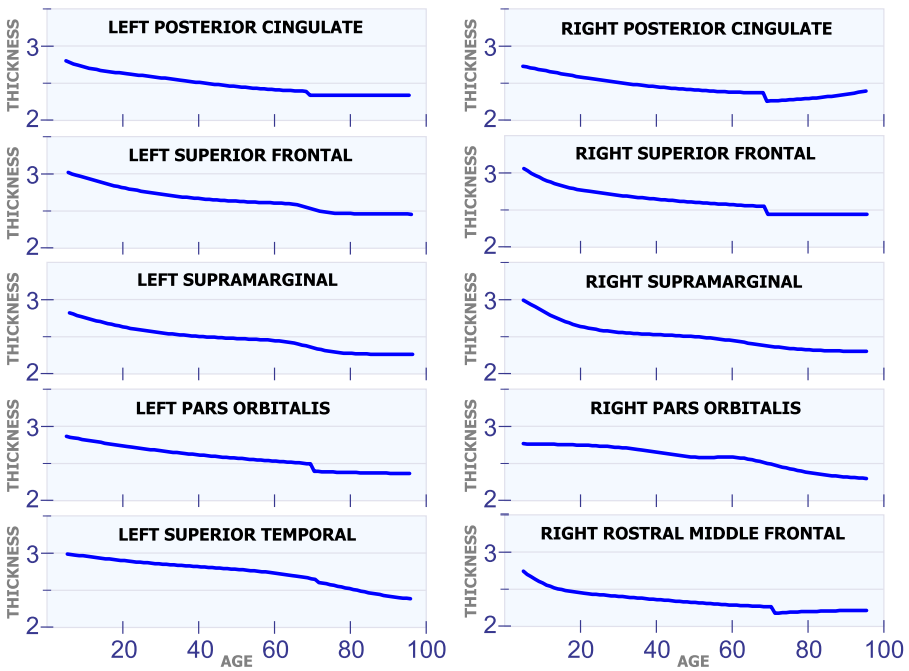


FIGURE 5 | Regions with changes in cortical thickness around 70 years of age.

same sense, an unexpected reduction around age 38 is present in the right lateral orbitofrontal and the right medial orbitofrontral. The last row in **Figure 3** shows the behavior of the cortical thickness in the left middle temporal and the right lateral occipital. Observe that the left middle temporal exhibits an abrupt transition around are age 45, while the right lateral occipital exhibits a transition around age 32.

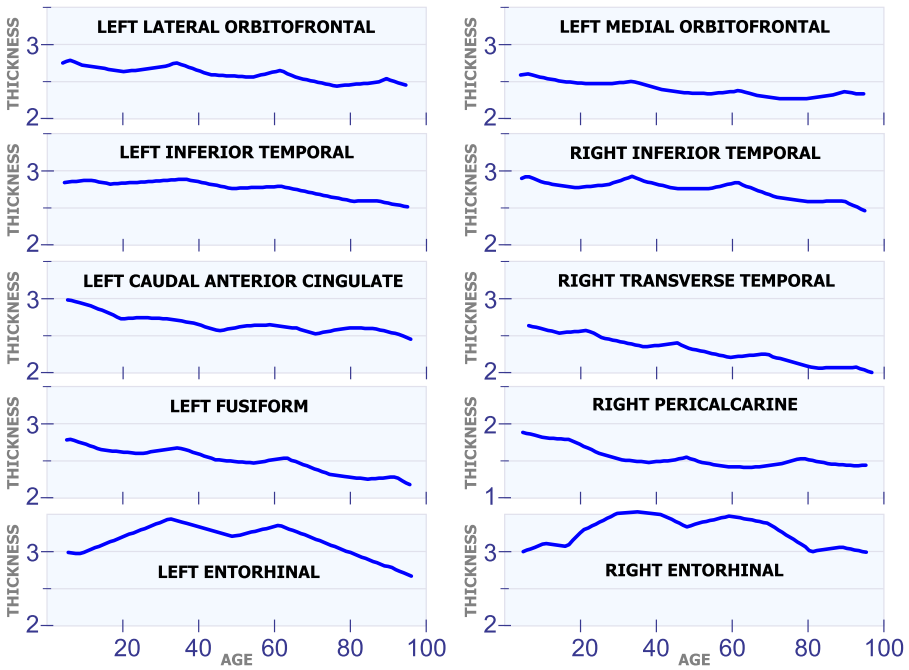


FIGURE 6 | Regions with multiple changes in cortical thickness through age.

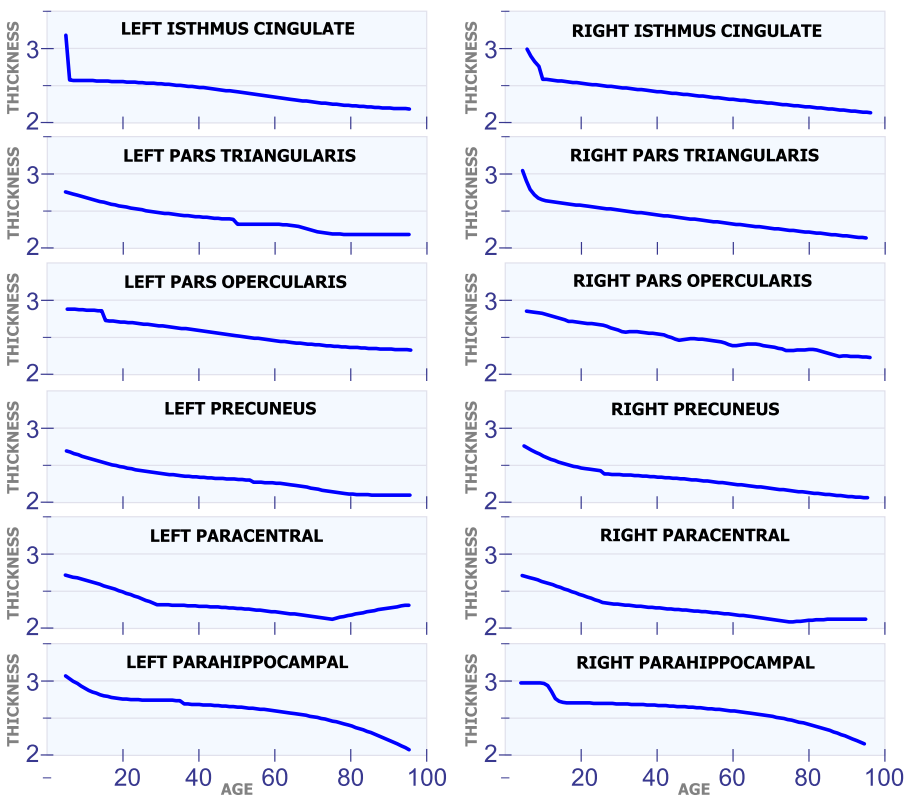


FIGURE 7 | Regions with constant changes in cortical thickness.

5.1.3 Changes in Cortical Thickness Around 50 years of Age

Figure 4 shows 10 regions that have changes in cortical thickness around 50 years of age. The first row in **Figure 4** displays the model for the cortical thickness in millimeters for the left inferior parietal and the right inferior parietal. In these two regions, the cortical thickness remains almost constant from age 25 to 50. Then, these regions present a slow and constant reduction in cortical starting at age 50. The graphs in the second row in **Figure 4** displays the cortical thickness for the left rostral middle frontal and the right fusiform. From age 25 to 50 the cortical thickness remains approximately constant in both regions. Then, starting at age 50 there is slow a constant reduction in cortical thickness. The third row in **Figure 4** shows the cortical thickness for the left cuneus and the right cuneus. Both regions present a sudden cortical thickness reduction at age 30 and 50. The fourth row in **Figure 4** displays the cortical thickness behavior for the left lingual and the right lingual. An abrupt change in cortical thickness is clearly observed in both regions at age 50 years old. The last row in **Figure 4** illustrates the behavior of the cortical thickness in the left pericalcarine and the right middle temporal. Notice that both regions exhibit a sudden change in cortical thickness in two different periods of life. The left pericalcarine exhibits the first change in cortical thickness around 20 years of age and the second change around 55 years of age. On the other hand, the right middle temporal has the first abrupt change at 20 years of age, while the second change is present around 45 years of age.

5.1.4 Changes in Cortical Thickness Around 70 years of Age

Figure 5 shows ten regions in the human brain that present changes in cortical thickness around 70 years of age. The first row in **Figure 5** illustrates these changes for the left posterior cingulate and the right posterior cingulate. These two regions exhibit a steady and non-linear reduction in cortical thickness during all stages of life. However, they have an abrupt reduction in cortical thickness around 70 years of age. All regions of the brain in **Figure 5** present a very similar behavior as the ones in the first row. They have a constant and slow reduction in cortical thickness with age. They also have a sudden reduction in cortical thickness around 70 years of age.

5.1.5 Regions With Changes at Multiple Ages

Figure 6 shows ten different regions that exhibit multiple cortical changes during the human lifespan. The first row in **Figure 6** shows the development of the left lateral orbitofrontal and the left medial orbitofrontal. These two regions have a non-linear relation with age, and they both have a sudden increase in cortical thickness at 35 and 62 years of age. The second row in **Figure 6** shows the left inferior temporal and the right inferior temporal. Again, these two regions present an abrupt increase in cortical thickness around 35 and 62 years of age. The graphs in the third row of **Figure 6** include the left caudal anterior cingulate and the right transverse temporal. Both regions have inflection points at 20, 45 and 70 years of age. The graphs in the fourth row in **Figure 6** include the behavior in the left fusiform and the right pericalcarine. The last row in **Figure 6** shows the cortical thickness development in the left entorhinal and the right entorhinal. These are the only two regions in the brain that have very big changes in cortical thickness through the lifespan. The cortical thickness in these two regions reaches a maximum value at ages 35 and 60.

5.1.6 Regions With a Constant Rate

All the regions in **Figure 7** exhibit a mostly steady reduction in cortical thickness through age. The first row in **Figure 7** shows the cortical thickness in millimeters for the left isthmus cingulate and the right isthmus cingulate. With the exception at the beginning of life, both of these two regions exhibit a mostly linear reduction in cortical thickness through life. The second row in **Figure 7** shows the cortical thickness in millimeters for the left pars triangularis and the right pars triangularis. From the graph, it can be observed that the left pars triangularis presents an abrupt transition in cortical thickness around 50 years of age. While the right pars triangularis exhibits a linear reduction in cortical thickness for most of the human life span. The third row in **Figure 7** includes the left pars opercularis and the right pars opercularis. Both of these two regions have an almost linear reduction in cortical thickness. The fourth row in **Figure 7** shows the behavior of the cortical thickness in the left precuneus and the right precuneus. The left precuneus exhibits a small transition around 55 years of age, while the right precuneus exhibits a minor transition in cortical thickness around 25 years of age. The fifth row in **Figure 7** shows the cortical thickness changes for the left paracentral and the right paracentral. Both of these regions have two inflection points, one at 30 of age and another at 75 years of age. The last row in **Figure 7** shows the models for the left parahippocampal and the right parahippocampal. The cortical thickness for both of these regions follows a non-linear reduction through life.

5.2 Cortical Thickness Changes Through Life

The study of changes in cortical thickness with age is very important because it provides information about the individual. For instance, a reduction in cortical thickness has been associated with some neurodegenerative diseases (Oertel-Knöchel et al., 2015). Additionally, it has been suggested that age-related non-linear changes in cortical thickness are influenced by family income and parental education (Piccolo et al., 2016). In the same sense, Plessen et al. evaluated the connection between measures of asymmetry in cortical thickness with age, sex, and cognitive performance (Plessen et al., 2014).

In this section, we compute the derivative of the cortical thickness using **Equation 3** and the models created using the artificial neural networks. The computer simulations were performed using stencils (kernels) with seven points, $N = 7$ in **Equation 2**. Additionally, the stencils were dynamically computed at the beginning and at the end of the lifespan to improve accuracy, see (Hassan et al., 2012). The computer simulations in **Section 5.1** focused on the value and progress of the cortical thickness through different ages. On the other hand, the simulations in this section

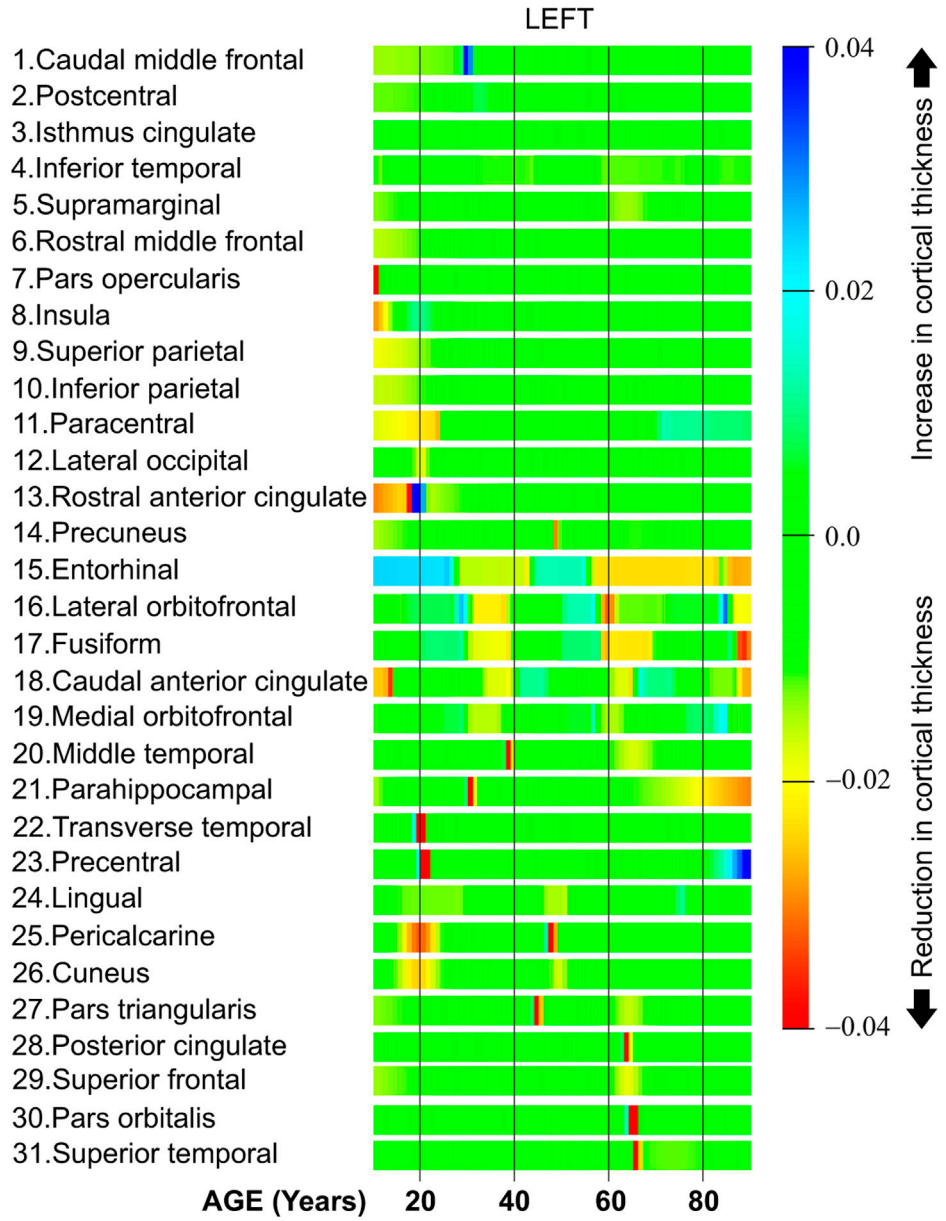


FIGURE 8 | Cortical thickness derivative with respect to age, left hemisphere.

focused on the speed of the cortical thickness during the life span. Thus, when the derivative is positive, the speed is also positive and this implies that there is an increase in the cortical thickness during this part of life. When the derivative is negative the speed is also negative implying that there is a reduction in cortical thickness for that part of life. In the same sense, when the derivative is almost zero, the speed is also close to zero, and therefore, the cortical thickness does not change.

Figure 8 shows the cortical thickness derivative with respect to age. Observe that the figure includes the results only for the left hemisphere of the brain. Observe also that the results are organized in clusters, that is, those brain regions with similar derivatives are displayed next to each other. The thickness derivative is represented using the color scale displayed on the right part of **Figure 8**. Starting at the top of the scale, the blue dark color is used to display a significant increase in cortical thickness. In the middle of the scale, the green color is used to indicate no changes in cortical thickness, 0.0. At the bottom of the scale, the red color is used to indicate an important reduction in cortical thickness.

Row one in **Figure 8** shows the derivative for the caudal middle frontal. As it can be seen this band is mostly green with a blue band around 30. Therefore, this region exhibits a constant derivative with an abrupt increase in the cortical thickness speed around 30 years of age. The bands from row two (postcentral) to row six (rostral middle frontal) in **Figure 8** are mostly green with

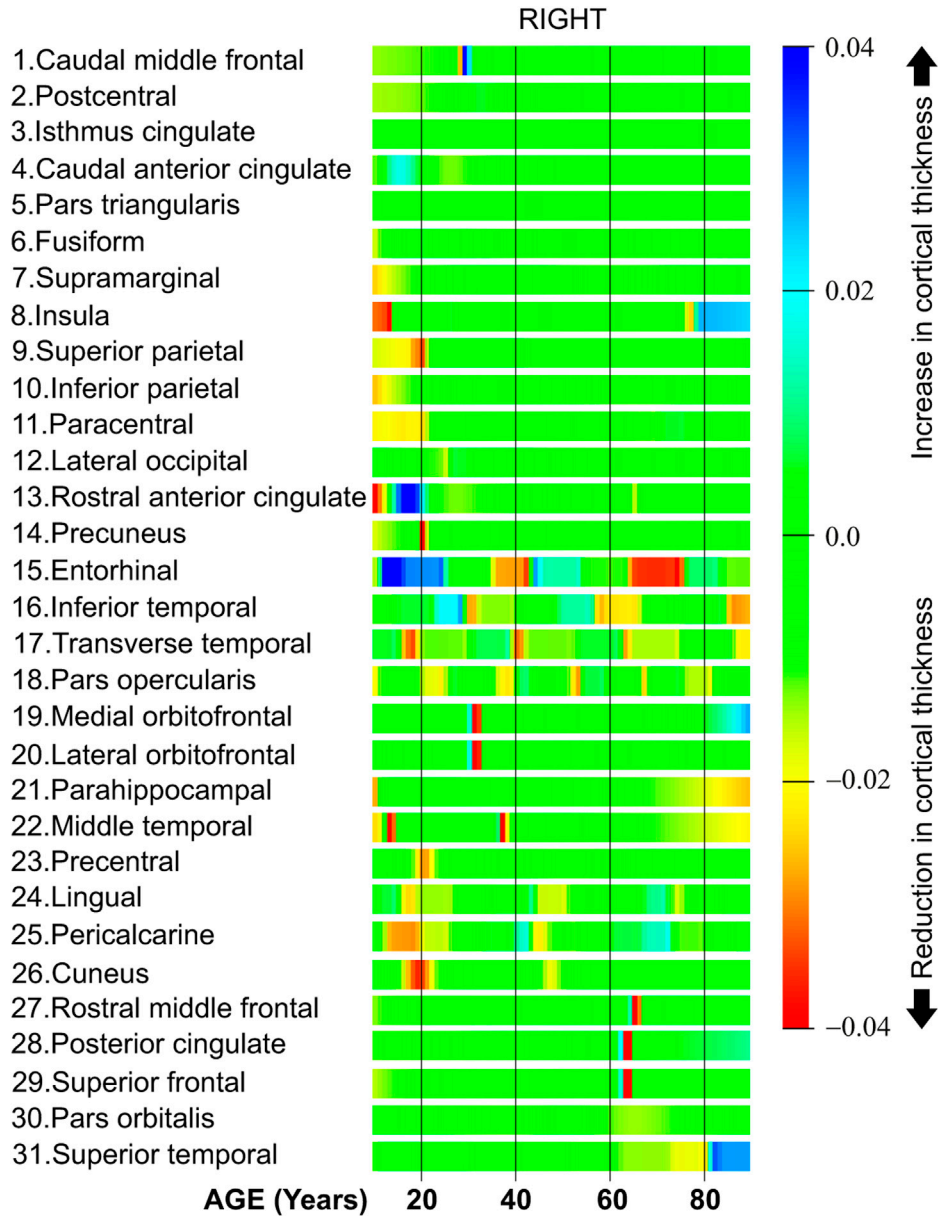


FIGURE 9 | Cortical thickness derivative with respect to age, right hemisphere.

some soft yellow zones. Thus, these brain regions exhibit an almost constant cortical thickness derivative during the lifespan. From row seven in **Figure 8** (pars opercularis) to row twelve (lateral occipital) all these bands have red and yellow zones at the beginning of life. Thus, these brain regions lose cortical thickness at high speed around the first 20 years of age. Row 13 in **Figure 8** shows the behavior of the rostral anterior cingulate. There are red, yellow and blue color bands in the first 20 years of life. This implies that the cortical thickness speed considerably changes during the first 2 decades of life. From row 15 (entorhinal) to row 21 (parahippocampal), all these brain regions present different cortical thickness speeds at diverse parts of life. Both the transverse temporal in row 22 and the precentral in row 23 have a red zone around 20 years of age. This implies that the human brain presents a period with great reductions in cortical thickness for these two regions at age 20.

All regions from row 24 (lingual) to row 26 (cuneus) exhibit a red or yellow band around 20 and 50 years. This means that during this age, the derivative is negative, and therefore, the cortical thickness is quickly reduced during these two parts of life. The last regions in **Figure 8** starting in row 28 (posterior cingulate) have a red band around 65 years of age. Thus, these regions exhibit a fast reduction in cortical thickness at 65 years.

Figure 9 shows the derivative of the cortical thickness for the right hemisphere. The regions in **Figure 9** are organized in clusters as in the regions in the left hemisphere.

TABLE 3 | Modeling of the cortical thickness through life.

	Lowest thickness	Highest thickness
Through all life	Left pericalcarine	Right entorhinal
age ≥40	Left pericalcarine	Right entorhinal
age ≥60	Right pericalcarine	Right entorhinal

TABLE 4 | Variability of the cortical thickness through life.

	Lowest variability	Highest variability
Through all life	Right caudal anterior cingulate	Left parahippocampal
age ≥40	Right caudal anterior cingulate	Right transverse temporal
age ≥60	Left pericalcarine	Left enthorhinal

TABLE 5 | Linearity of the cortical thickness through life.

	Lowest linearity	Highest linearity
Through all life	Right entorhinal	Left lateral occipital
age ≥40	Right entorhinal	Right isthmus cingulate
age ≥60	Left lateral orbitofrontal	Left pericalcarine

The first six regions in **Figure 9**, from caudal middle frontal to fusiform, have a constant cortical thickness speed for most of the life. Regions from row 7 (supramarginal) to 11 (paracentral) present a high cortical thickness reduction during the first 20 years of life. The regions located in the cluster in the middle of **Figure 9**, from row 15 (entorhinal) to row 18 (pars opercularis), have several abrupt changes in the cortical thickness speed at different parts of life. Both the medial orbitofrontal in row 19 and the lateral orbitofrontal in row 20 have a negative cortical thickness speed around 33 years of age. The regions from row 24 (lingual) to 26 (cuneus) have a negative cortical thickness speed around 20 and 50 years of age. Finally, the regions from row 27 (rostral middle frontal) to 31 (superior temporal) present a negative cortical thickness speed around 70 years of age.

Tables 3–**5** show some of the main results from this study. The first row in **Table 3** indicates that the left pericalcarine is the region with the lowest cortical thickness throughout all life. As it can be seen from the third column in **Table 3**, the right entorhinal is the region with the highest thickness throughout all life, after 40 years of age, and after 60 years of age. However, the value in the second column in the last row in **Table 3** indicates that the right pericalcarine is the region with the lowest thickness after 60 years of age.

Table 4 shows the variability of the cortical thickness. Through all life, the region with the lowest variability is the right caudal anterior cingulate, and the region with the highest variability is the left parahippocampal. For a person 40 years and older, the region with the lowest variability is again the right caudal anterior cingulate, while the region with the highest variability is the right transverse temporal. For a person 60 years and older, the left enthorhinal is the region with the highest variability, and the left pericalcarine is the region with the lowest variability.

Table 5 measures the linearity of the cortical thickness with age. Throughout life, the left lateral occipital is the region that exhibits the highest linearity. For an age of 40 years and older, the right isthmus cingulate is the region with the highest linearity. For an age of 60 years and older, the left pericalcarine is the region with the highest linearity. In this sense, the cortical thickness in those regions in the third column of **Table 5** can be estimated using a simple linear model. On the other hand, the cortical thickness of those regions in the second column of **Table 5** cannot be accurately predicted using a simple linear model. In summary, the models created with artificial neural networks adapt to the patterns in the data. Therefore, the performance of a neural network model or a linear model is very similar in those regions that exhibit a linear tendency in its cortical thickness with time. For those regions that have a linear behavior, the mean squared error was 0.016 for both models. However, the performance of the neural network models was better than the performance of linear models in those regions with complex patterns through age. For those regions that do not have a linear behavior with time, the mean squared error for the neural network models was 0.03 while the mean squared error for the linear models was 3.0.

In this publication, we propose the use of artificial neural networks to model the thickness of the cortical thickness through life for different regions in the brain. Once the neural networks are trained, it is possible to validate the performance of the model using new datasets. One important feature of artificial neural networks is their capacity to generalize. This means that a neural network has been trained, it should be able to predict the cortical thickness of data that the network has not seen before (Masters, 2015). Future work may include the study on how to utilize the artificial neural network models to understand various cognitive functions through life.

6 CONCLUSION

This work analyzes the progress of the cortical thickness with age using Artificial Intelligence. A set of artificial neural networks was trained and validated using a dataset with information from 1,100 healthy individuals. Each neural network was designed to model one single region in the human brain. Thus, 31 artificial neural networks were created to model the cortical thickness in each region in the left hemisphere of the brain. Similarly, 31 networks were created to model the cortical thickness for the regions in the right hemisphere. Furthermore, computer simulations were used to adjust the number of neurons in the hidden layer of the artificial neural networks, and thus, obtain the best model given the amount of data available.

The models created by the artificial neural networks were, then, organized in clusters. Each cluster included those regions that followed a similar pattern for the cortical thickness through age. The results from the computer simulations show that the models allow the detection of abrupt changes in cortical thickness. The simulations also provide an age estimate of when these changes may happen.

Additionally, the neural networks were used with numerical differentiation techniques to estimate the derivative of the cortical

thickness with respect to age. Dynamic stencils were used to improve the accuracy of the derivative at the beginning and the end of life. Then, color bands were created to display the speed of the cortical thickness. A color scale was designed to locate and visualize those parts of life with a positive or a negative speed. A positive speed is obtained when there is an increase in cortical thickness. On the other hand, a negative speed is present when there is a reduction in cortical thickness during that part of life. Therefore, the color bands allowed the detection of those parts of life with a reduction or an increase in cortical thickness. Finally, these graphs were organized in clusters. Each cluster included those regions with similar behavior through life.

After examining the results, it was concluded that some regions in the left hemisphere do not present the same progress with age as the counterpart regions in the right hemisphere. Some regions in the brain exhibit very particular patterns in their cortical thickness; one of these regions is the entorhinal. One advantage of the methodology proposed in this paper is that the models created using the artificial neural networks do not assume a linear or non-linear model. Instead, the artificial neural network is capable of dynamically adapt to the required complexity of each region in the human brain. Additionally, artificial neural networks are insensitive to noise present in the data and learn the patterns relevant to the specific application. Most importantly, neural networks are capable of generalizing, that is, they are able to predict patterns that are present in other datasets that were not used for training.

DATA AVAILABILITY STATEMENT

Publicly available datasets were analyzed in this study. This data can be found here: OASIS: http://oasis-brains.org/ IXI Dataset https://brain-development.org/ixi-dataset/ NKI-RS http://fcon_1000.projects.nitrc.org/indi/enhanced/.

ETHICS STATEMENT

Ethical approval was not required as per local legislation. All data used was in the public domain and has been deidentified. References to the original datasets are included in the manuscript where additional information about the original data collection may be found.

AUTHOR CONTRIBUTIONS

SL organized and performed the training of artificial neural networks. SL prepared the figures. M-AI-M performed the validation of the models created by artificial neural networks. D-LA-O assisted with the preparation for the data so that the dataset could be used for the Artificial Intelligence algorithms. PF drafted and organized the tests and the text in the document. JS provided assistance about the interpretation of the results and improve the quality of the manuscript by providing key information.

ACKNOWLEDGMENTS

We acknowledge DAIP, University of Guanajuato and the University of Ottawa for their sponsorship in the realization of this work. This work was developed during the sabbatical stay of SL at the Faculty of Health Sciences in the University of Ottawa, Canada.

REFERENCES

Alpaydin, E. (2016). *Machine Learning: The New AI (The MIT Press Essential Knowledge Series)*. Massachusetts, London, England: The MIT Press Cambridge.

Chen, Z. J., He, Y., Rosa-Neto, P., Gong, G., and Evans, A. C. (2011). Age-Related Alterations in the Modular Organization of Structural Cortical Network by Using Cortical Thickness from MRI. *NeuroImage.* 56, 235–245. doi:10.1016/j.neuroimage.2011.01.010

Churchwell, J. C., and Yurgelun-Todd, D. A. (2013). Age-Related Changes in Insula Cortical Thickness and Impulsivity: Significance for Emotional Development and Decision-Making. *Developmental Cogn. Neurosci.* 6, 80–86. doi:10.1016/j.dcn.2013.07.001

Fischl, B. (2012). FreeSurfer. *NeuroImage.* 62, 774–781. doi:10.1016/j.neuroimage.2012.01.021

Fjell, A. M., Westlye, L. T., Grydeland, H., Amlien, I., Espeseth, T., Reinvang, I., et al. (2014). Accelerating Cortical Thinning: Unique to Dementia or Universal in Aging?. *Cereb. Cortex.* 24, 919–934. doi:10.1093/cercor/bhs379

Fouche, J.-P., du Plessis, S., Hattingh, C., Roos, A., Lochner, C., Soriano-Mas, C., et al. (2017). Cortical Thickness in Obsessive-Compulsive Disorder: Multisite Mega-Analysis of 780 Brain Scans From Six Centres. *Br. J. Psychiatry.* 210, 67–74. doi:10.1192/bjp.bp.115.164020

Goodfellow, I., Bengio, Y., and Courville, A. (2016). *Deep Learning (Adaptive Computation and Machine Learning Series)*. Massachusetts, London, England: The MIT Press Cambridge.

Hassan, H. Z., Mohamad, A. A., and Atteia, G. E. (2012). An Algorithm for the Finite Difference Approximation of Derivatives With Arbitrary Degree and Order of Accuracy. *J. Comput. Appl. Mathematics.* 236, 2622–2631. doi:10.1016/j.cam.2011.12.019

Information eXtraction Images (2019). *IXI Dataset, Biomedical Image Analysis Group, Information eXtraction from Images*. London: Imperial College London, South Kensington Campus. Available at: https://brain-development.org/ixi-dataset/ (Accessed December 4, 2019).

Jin, B., Jing, Z., and Zhao, H. (2017). Incremental and Decremental Extreme Learning Machine Based on Generalized Inverse. *IEEE Access.* 5, 20852–20865. doi:10.1109/access.2017.2758645

Jordan, M. I., and Mitchell, T. M. (2015). Machine Learning: Trends, Perspectives, and Prospects. *Science.* 349, 255–260. doi:10.1126/science.aaa8415

Kelleher, J. D., Namee, B. M., and D'Arcy, A. (2015). *Fundamentals of Machine Learning for Predictive Data Analytics: Algorithms, Worked Examples, and Case Studies*. Cambridge, MA: Massachusetts Institute of Technology.

Klein, A., and Tourville, J. (2012). 101 Labeled Brain Images and a Consistent Human Cortical Labeling Protocol. *Front. Neurosci.* 6, 171–212. doi:10.3389/fnins.2012.00171

Landman, B. A., Huang, A. J., Gifford, A., Vikram, D. S., Lim, I. A. L., Farrell, J. A. D., et al. (2011). Multi-Parametric Neuroimaging Reproducibility: A 3-T Resource Study. *NeuroImage.* 54, 2854–2866. doi:10.1016/j.neuroimage.2010.11.047

Ledesma, S., Ibarra-Manzano, M. A., Garcia-Hernandez, M. G., and Almanza-Ojeda, D. L. (2017). *Neural Lab a Simulator for Artificial Neural Networks*. London, United Kingdom: IEEE Computing Conference, 716–721.

Lemaitre, H., Goldman, A. L., Sambataro, F., Verchinski, B. A., Meyer-Lindenberg, A., Weinberger, D. R., et al. (2012). Normal Age-Related Brain Morphometric Changes: Nonuniformity Across Cortical Thickness, Surface Area and Gray Matter Volume? *Neurobiol. Aging.* 33, 617–619. doi:10.1016/j.neurobiolaging.2010.07.013

Marcus, D. S., Wang, T. H., Parker, J., Csernansky, J. G., Morris, J. C., and Buckner, R. L. (2007). Open Access Series of Imaging Studies (OASIS): Cross-Sectional MRI Data in Young, Middle Aged, Nondemented, and Demented Older Adults. *J. Cogn. Neurosci.* 19, 1498–1507. doi:10.1162/jocn.2007.19.9.1498

Marsland, S. (2015). *Machine Learning: An Algorithmic Perspective*. Second Edition. Boca Raton, London, New York: Chapman & Hall/Crc Machine Learning & Pattern Recognition, CRC Press, Taylor & Francis Group.

Masters, T. (2015). *Deep Belief Nets in C++ and CUDA, Restricted Boltzmann Machines*. Lexington, KY, USA: Masters.

McCarthy, C. S., Ramprashad, A., Thompson, C., Botti, J. A., Coman, I. L., and Kates, W. R. (2015). A Comparison of FreeSurfer-Generated Data With and Without Manual Intervention. *Front. Neurosci.* 9, 379–418. doi:10.3389/fnins.2015.00379

McGinnis, S. M., Brickhouse, M., Pascual, B., and Dickerson, B. C. (2011). Age-Related Changes in the Thickness of Cortical Zones in Humans. *Brain Topogr.* 24, 279–291. doi:10.1007/s10548-011-0198-6

Oertel-Knöchel, V., Reuter, J., Reinke, B., Marbach, K., Feddern, R., Alves, G., et al. (2015). Association Between Age of Disease-Onset, Cognitive Performance and Cortical Thickness in Bipolar Disorders. *J. Affective Disord.* 174, 627–635. doi:10.1016/j.jad.2014.10.060

Piccolo, L. R., Merz, E. C., He, X., Sowell, E. R., and Noble, K. G. (2016). Age-Related Differences in Cortical Thickness Vary by Socioeconomic Status. *PLOS one.* 11, e0162511–18. doi:10.1371/journal.pone.0162511

Plessen, K. J., Hugdahl, K., Bansal, R., Hao, X., and Peterson, B. S. (2014). Sex, Age, and Cognitive Correlates of Asymmetries in Thickness of the Cortical Mantle Across the Life Span. *J. Neurosci.* 34, 6294–6302. doi:10.1523/jneurosci.3692-13.2014

Razlighi, Q. R., Habeck, C., Barulli, D., and Stern, Y. (2017). Cognitive Neuroscience Neuroimaging Repository for the Adult Lifespan. *NeuroImage.* 144, 294–298. doi:10.1016/j.neuroimage.2015.08.037

Russell, S., and Norvig, P. (2020). *Artificial Intelligence: A Modern Approach*. 4th Edition. Hoboken, NJ: Pearson International.

Salat, D. H., Buckner, R. L., Snyder, A. Z., Greve, D. N., Desikan, R. S. R., Busa, E., et al. (2004). Thinning of the Cerebral Cortex in Aging. *Cereb. Cortex.* 14, 721–730. doi:10.1093/cercor/bhh032

Scott, M. L. J., Bromiley, P. A., Thacker, N. A., Hutchinson, C. E., and Jackson, A. (2009). A Fast, Model-Independent Method for Cerebral Cortical Thickness Estimation Using MRI. *Med. Image Anal.* 13, 269–285. doi:10.1016/j.media.2008.10.006

Sowell, E. R., Peterson, B. S., Kan, E., Woods, R. P., Yoshii, J., Bansal, R., et al. (2007). Sex Differences in Cortical Thickness Mapped in 176 Healthy Individuals Between 7 and 87 Years of Age. *Cereb. Cortex.* 17, 1550–1560. doi:10.1093/cercor/bhl066

Steffener, J., Habeck, C., O'Shea, D., Razlighi, Q., Bherer, L., and Stern, Y. (2016). Differences Between Chronological and Brain Age Are Related to Education and Self-Reported Physical Activity. *Neurobiol. Aging.* 40, 138–144. doi:10.1016/j.neurobiolaging.2016.01.014

Tamnes, C. K., Østby, Y., Fjell, A. M., Westlye, L. T., Due-Tønnessen, P., and Walhovd, K. B. (2010). Brain Maturation in Adolescence and Young Adulthood: Regional Age-Related Changes in Cortical Thickness and White Matter Volume and Microstructure. *Cereb. Cortex.* 20, 534–548. doi:10.1093/cercor/bhp118

Tustison, N. J., Cook, P. A., Klein, A., Song, G., Das, S. R., Duda, J. T., et al. (2014). Large-Scale Evaluation of ANTs and FreeSurfer Cortical Thickness Measurements. *NeuroImage.* 99, 166–179. doi:10.1016/j.neuroimage.2014.05.044

Wierenga, L. M., Langen, M., Oranje, B., and Durston, S. (2014). Unique Developmental Trajectories of Cortical Thickness and Surface Area. *NeuroImage.* 87, 120–126. doi:10.1016/j.neuroimage.2013.11.010

Winkler, A. M., Greve, D. N., Bjuland, K. J., Nichols, T. E., Sabuncu, M. R., Ha˚berg, A. K., et al. (2018). Joint Analysis of Cortical Area and Thickness as a Replacement for the Analysis of the Volume of the Cerebral Cortex. *Cereb. Cortex.* 28, 738–749. doi:10.1093/cercor/bhx308

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

frontiers | Frontiers in Artificial Intelligence

ORIGINAL RESEARCH
published: 26 May 2022
doi: 10.3389/frai.2022.660581

Targeted Screening for Alzheimer's Disease Clinical Trials Using Data-Driven Disease Progression Models

Neil P. Oxtoby[1], Cameron Shand[1], David M. Cash[2], Daniel C. Alexander[1] and Frederik Barkhof[1,3] for the Alzheimer's Disease Neuroimaging Initiative and the Alzheimer's Disease Cooperative Study*

[1] Centre for Medical Image Computing, Department of Computer Science, University College London, London, United Kingdom, [2] Dementia Research Centre, Department of Neurodegenerative Disease, UCL Queen Square Institute of Neurology, University College London, London, United Kingdom, [3] Amsterdam University Medical Center, Amsterdam, Netherlands

Heterogeneity in Alzheimer's disease progression contributes to the ongoing failure to demonstrate efficacy of putative disease-modifying therapeutics that have been trialed over the past two decades. Any treatment effect present in a subgroup of trial participants (responders) can be diluted by non-responders who ideally should have been screened out of the trial. How to identify (screen-in) the most likely potential responders is an important question that is still without an answer. Here, we pilot a computational screening tool that leverages recent advances in data-driven disease progression modeling to improve stratification. This aims to increase the sensitivity to treatment effect by screening out non-responders, which will ultimately reduce the size, duration, and cost of a clinical trial. We demonstrate the concept of such a computational screening tool by retrospectively analyzing a completed double-blind clinical trial of donepezil in people with amnestic mild cognitive impairment (clinicaltrials.gov: NCT00000173), identifying a data-driven subgroup having more severe cognitive impairment who showed clearer treatment response than observed for the full cohort.

Keywords: disease progression modeling, Alzheimer's disease, mild cognitive impairment, clinical trials, screening, dementia, biomarkers, donepezil

OPEN ACCESS

Edited by:
Kathiravan Srinivasan, Vellore Institute of Technology, India

Reviewed by:
Pengwei Hu, Merck, Germany
Kewei Chen, Banner Alzheimer's Institute, United States

****Correspondence:***
Neil P. Oxtoby
n.oxtoby@ucl.ac.uk

Specialty section:
This article was submitted to Medicine and Public Health, a section of the journal Frontiers in Artificial Intelligence

Received: *29 January 2021*
Accepted: *25 April 2022*
Published: *26 May 2022*

Citation:
Oxtoby NP, Shand C, Cash DM, Alexander DC and Barkhof F (2022) Targeted Screening for Alzheimer's Disease Clinical Trials Using Data-Driven Disease Progression Models. Front. Artif. Intell. 5:660581. doi: 10.3389/frai.2022.660581

1. INTRODUCTION

Alzheimer's Disease (AD) is one of the most important socioeconomic challenges of the twenty-first century, being the leading cause of age-related dementia in an aging global population. Despite decades of research and clinical trials of potential therapies (Cummings et al., 2018b), no trials have been able to prove disease-modifying efficacy (Cummings et al., 2014, 2016, 2017, 2018a, 2019, 2020). There are multiple possible explanations for this. For example, potentially targeting the "wrong" pathology at the wrong time—typically amyloid protein pathogens are the target but if a treatment is given to symptomatic individuals, it may be too late to halt or reverse any damage done. Notwithstanding this, enrolling the right people at the right time (disease stage) into a clinical trial remains a considerable challenge because of undetected heterogeneity in phenotype/presentation (Firth et al., 2020)

and/or ensuring the underlying pathology is present (Salloway et al., 2014), which can be a general problem because clinical trials often cannot adapt their designs to accommodate research discoveries made after they have begun. This can result in enrolment of non-responders into a clinical trial that wash out treatment effect in any subgroup of responders. Identification of non-responders typically occurs in *post hoc* subgroup analysis, which does not confer the benefits of a reduced trial size, and requires careful analysis to infer conclusions which can be misleading (Wang et al., 2007; Cummings, 2018). Given the breadth of evidence in support of the amyloid hypothesis (Hardy and Higgins, 1992) that has driven this clinical research for two decades, albeit with some controversies (Morris et al., 2014), here we focus on the aforementioned challenges of screening to identify the right participants at the right time. The good news is that there has been a swell of computational research into unraveling the heterogeneity of Alzheimer's disease progression over the past decade (e.g., see Oxtoby et al., 2017), driven largely by the increasing availability of large open medical datasets (Marinescu et al., 2018).

Computational approaches for aging and age-related diseases have been designed to fuse multimodal data into a quantitative template (Bilgel and Jedynak, 2019) of disease progression. These signatures often include a patient staging mechanism (Young et al., 2014) that provides a quantitative tool for fine-grained, individualized inference based on disease severity that goes above and beyond standard clinical phenotyping using patient symptoms. A recent innovation of data-driven disease progression modeling incorporates unsupervised machine learning, i.e., clustering, to provide both subtype and stage inference (Young et al., 2018). A frequent occurrence in this literature are claims of how these data-driven models can benefit clinical trials in Alzheimer's disease, but we are yet to find any evidence of studies actually analyzing clinical trial data to demonstrate the claimed benefit.

In this work we demonstrate the potential of data-driven models of disease progression to enhance clinical trials in Alzheimer's disease *via* targeted screening. We achieve this by example, using a particular modeling approach—the event-based model (Fonteijn et al., 2012)—in a *post hoc* subgroup analysis of a particular completed clinical trial that concluded without evidence of efficacy (Petersen et al., 2005).

2. MATERIALS AND METHODS

This section describes the data, the computational model, and the statistical analysis used in our study. Overall, our analysis includes three steps. First, we fit a data-driven disease progression model of cognitive decline in AD to data from a large multicentre observational study, the Alzheimer's Disease Neuroimaging Initiative (ADNI; *training set*). Second, we use this computational model to score disease progression at baseline for participants in the completed "MCI" clinical trial from the Alzheimer's Disease Cooperative Study (ADCS-MCI; *test set*). Finally, this disease progression score is used to stratify the ADCS-MCI Trial participants for a *post hoc* analysis of subgroup treatment effect.

2.1. Data

Our reference model fit to data from the ADNI observational study is used to stage participants from the ADCS-MCI clinical trial (clinicaltrials.gov: NCT00000173; Petersen et al., 2005). For this we use a set of features common to both data sets, which is a subset of cognitive instruments used in the ADCS-MCI trial (see the vertical axis of Results, **Figure 1**), taking care to exclude ADAS-Cog (being a secondary outcome of the trial).[1] For simplicity, we included only ADNI participants having complete data for this feature set.

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD).

Additional data used in the preparation of this article were obtained from the Alzheimer's Disease Cooperative Study (ADCS) database (adcs.org). Specifically, we analyse data from the completed ADCS-MCI clinical trial of donepezil and vitamin E, reported in Petersen et al. (2005). The ADCS-MCI trial aimed to assess the efficacy of vitamin E and donepezil in subjects with amnestic MCI. The primary end point was the time to the development of possible or probable AD dementia, with secondary outcomes on cognition and function. Measurements were taken at 6-month intervals until the end of the trial (36 months). At screening, 769 subjects were included in the trial, randomized into 259, 257, and 253 subjects for the placebo, vitamin E, and donepezil arms, respectively—reducing to 174, 158, and 145 by the end of the trial.

2.2. Event-Based Model

The event-based model (EBM) (Fonteijn et al., 2012; Young et al., 2014) estimates the most likely sequence, and uncertainty in this sequence, of observable cumulative abnormality events in the pathophysiological cascade (Jack et al., 2010) of a progressive disease. In this context, an event constitutes deviation of a biomarker measurement from those typical of healthy controls, toward those typical of patients. Events, and the overall sequence of events, are probabilistic entities. The EBM sequence of cumulative abnormality is estimated from cross-sectional data. This is made possible by combining data from a cohort of individuals at different stages of cumulative abnormality. The EBM sequence estimation is achieved directly from the data distributions in diseased and healthy groups and without *a priori*-defined disease stages or biomarker cutpoints /thresholds. The EBM, in its various versions, has been applied to a variety of diseases since 2011 (e.g., Fonteijn et al., 2012; Eshaghi et al., 2018; Oxtoby et al., 2018, 2021; Wijeratne et al., 2018; Firth et al., 2020). For a detailed intuitive description of the EBM, we refer the reader to Oxtoby et al. (2021).

[1] Results with ADAS-Cog included can be found in the **Supplementary Material**.

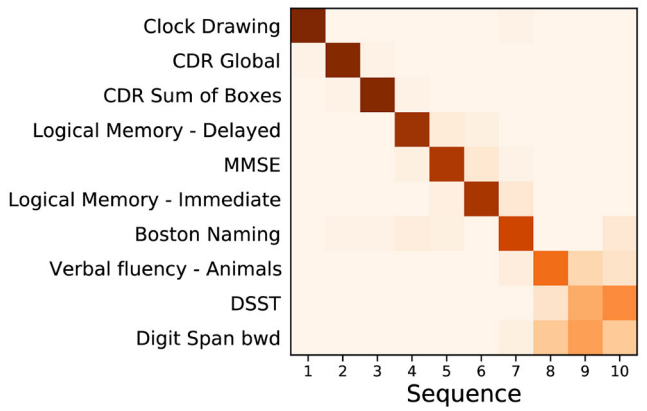


FIGURE 1 | Event based model of cognitive decline (ADNI). Positional density/variance diagram showing the sequence (top to bottom) and uncertainty (left to right) under five-fold cross-validation (repeated 10 times). CDR, clinical dementia rating; MMSE, mini-mental state examination; bwd, backward; DSST, digit symbol substitution test.

Here, we employ the recently-developed kernel density estimation (KDE) EBM that copes naturally with the ceiling/floor effects seen in cognitive data (Firth et al., 2020), and gives a cleaner interpretation of the model by exploiting prior information on disease direction (Oxtoby et al., 2021). To improve generalizability, we perform repeated five-fold cross-validation (10 repeats) and combine all 50 sets of posterior samples of the EBM into a cross-validated positional density map (Oxtoby et al., 2021).

The EBM affords us a screening tool by way of the patient staging mechanism introduced by Young et al. (2014). This process assigns a model stage (disease progression score) that maximizes the likelihood given an individual's set of measurements. Here, we use the ADNI-trained EBM to stage baseline data from the ADCS-MCI clinical trial, then stratify subjects into strata based on disease progression scores for *post hoc* subgroup analyses. In future, this process could be performed as part of the screening process to homogenize the clinical trial cohort.

2.3. Statistical Analysis

Our hypothesis is that AD clinical trial cohorts are likely to contain undetected heterogeneity that washes out treatment effects which may exist in an independently identifiable subgroup of responders. Accordingly, in order to examine whether our proposed screening tool can detect this heterogeneity and reveal such a subgroup of responders, our *post hoc* subgroup analysis of the ADCS-MCI clinical trial closely follows the primary analyses in Petersen et al. (2005). We describe the key steps below.

Primary Outcome: We use Kaplan–Meier estimators to estimate the rate of progression from MCI to AD over the course of the trial. Additionally, Cox proportional-hazards models were constructed to compare the risk for progression in each treatment arm with the placebo (using baseline age, MMSE, and APOE-ϵ4 carrier status as covariates). This intention-to-treat analysis in the trial was conducted for both placebo vs. vitamin E and placebo vs. donepezil, but in this paper we focus on the latter.

To correct for multiple comparisons in the Cox proportional-hazards model (for the two treatment arms), the Hochberg method was used. As our introduction of subgroups increases the number of comparisons made, we extend this adjustment for the total number of subgroups, regardless of whether a single subgroup is the focus of analysis.

Secondary Outcome: We compare ADAS-Cog 13 scores between placebo and donepezil arms in subgroups at each 6-month interval to assess the difference in longitudinal cognitive decline. A two-sided Mann–Whitney U-test is used to compare the treatment groups at each time point for each subgroup, correcting for multiple comparisons using the Hochberg method.

3. RESULTS

3.1. Reference Model

Figure 1 shows a positional variance diagram for an event-based model (Firth et al., 2020) of cognitive decline due to probable Alzheimer's disease, across a set of cognitive instruments from $N = 810$ (of 2,040) ADNI participants [229 cognitively normal (CN), 181 AD, 400 MCI] having complete data (see Section 2). The cross-validated model's confidence in the sequence is higher where the positional variance is reduced—a dark diagonal corresponds to strong confidence in the data-driven ordering. The estimated sequence of cognitive decline starts from the Clock Drawing test and Clinical Dementia Rating (CDR), through tests of memory recall (Logical Memory) and general cognition (MMSE), to verbal fluency (Boston Naming; Animals), working memory (Digit span backwards), and executive function (Digit Symbol Substitution Test, DSST).

Figure 2 shows a key component of the EBM—the normal/abnormal mixture models for each cognitive instrument (blue/orange solid lines, respectively), and the resulting cumulative probability of an event having occurred (dashed lines) (Fonteijn et al., 2012). These sigmoidal event probabilities quantify divergence from normality (Oxtoby et al., 2021) and provide a visualization of the data-driven event threshold (akin to a data-driven biomarker cutpoint). Histograms show the AD (orange) and CN (blue) data from ADNI. Early/late events are, respectively, those that have occurred in many/few patients and thus show greater/smaller separation between the group histograms.

3.2. Patient Staging: Re-screening the ADCS-MCI Trial

Figure 3 shows the distribution of patient stages assigned to participants in the (**Figure 3A**) ADNI study and (**Figure 3B**) ADCS-MCI trial, using the ADNI-trained EBM shown in **Figure 3**. The MCI distributions show considerable heterogeneity, with a notable late-stage ADCS-MCI subgroup beyond stage 8 in **Figure 3B**, delineated by a red dashed line. **Table 1** compares the whole ADCS-MCI cohort and 2 subgroups

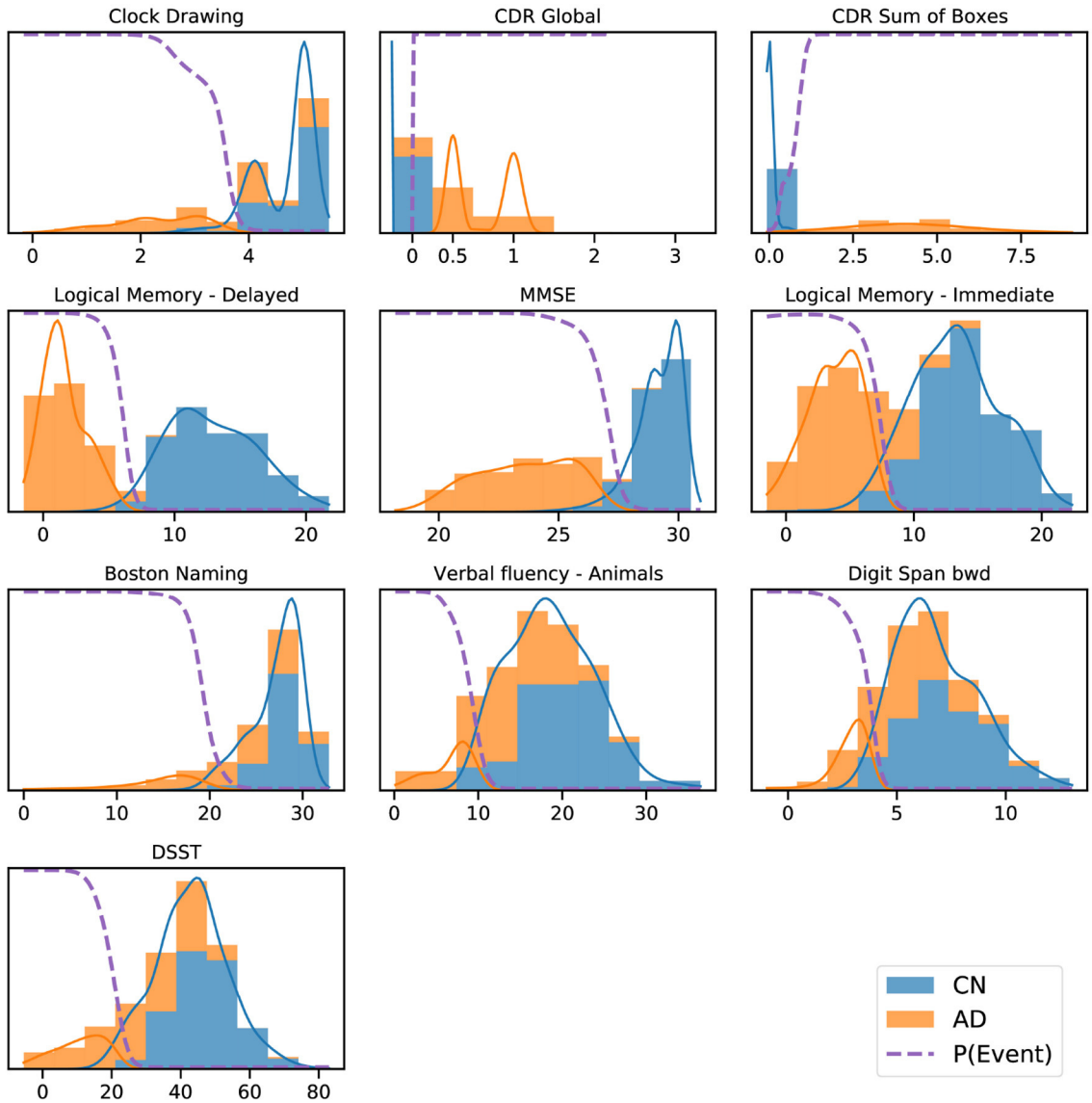


FIGURE 2 | ADNI data histograms (adjusted for age and education level) and EBM mixture models for each feature. Orange bars corresponds to AD patient data, blue bars to data from CN participants, showing the "normal" and "abnormal" distributions and the determined probability of the event having occurred (dashed line).

("Late-stage" and "Others") on demographic and cognitive measures at baseline.

Primary Outcome: Figure 4 shows Kaplan–Meier curves for the whole ADCS-MCI cohort (**Figure 4A**), the early-to-middle "Others" subgroup (**Figure 4B**) and the "Late-stage" subgroup (**Figure 4C**) in the placebo and donepezil arms, illustrating the change in survival rates (specifically, not progressing to probable AD dementia) during the trial. For each survival function estimate, 95% confidence intervals are shown in the shaded area. **Figure 5** shows the corresponding hazard ratios and 95% confidence intervals for Cox proportional-hazards models quantifying the risk of progression from MCI to AD. Although there are no significant differences between all subjects (hazard ratio 0.80; 95% CI 0.57–1.13; $p = 0.42$), the estimated effect seems larger than in the early-to-middle stage subgroup (hazard ratio 1.00; 95% CI 0.67–1.51; $p = 0.99$), or the late-stage subgroup (hazard ratio 0.55; 95% CI 0.28–1.07; $p = 0.24$).

Figure 6 shows ADAS-Cog 13 scores at 6-month intervals throughout the ADCS-MCI trial separately for the two subgroups. Conducting a two-sided Mann–Whitney U-test at each time point, no significant difference (in adjusted p-values) was found in either subgroup, despite the apparent trend toward treatment effect in the late-stage subgroup.

4. DISCUSSION

We fit an event-based model of cognitive decline in Alzheimer's disease using a reference data set (ADNI), which was then used to score disease progression in subjects at baseline in a completed clinical trial (ADCS-MCI). This disease progression score was

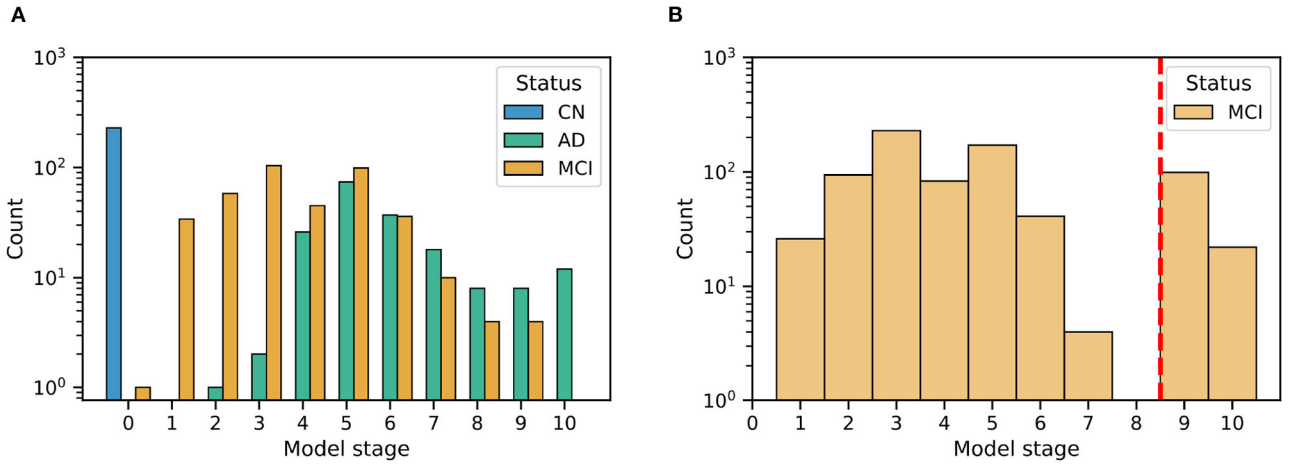


FIGURE 3 | Histograms of model stage for subjects in the ADNI dataset **(A)** and ADCS-MCI trial **(B)**.

TABLE 1 | Demographic and cognitive comparison of ADCS-MCI trial participants (All) and the model-determined subgroups thereof ("Late-stage" and "Others").

	Group		
Measure	**All (*N* = 769)**	**Others (*N* = 648)**	**Late-stage (*N* = 121)**
Age (years)	72.9 (7.3)	73.0 (7.2)	72.4 (7.9)
Education (years)	14.6 (3.1)	14.6 (3.1)	15.0 (3.0)
Sex (% female)	352 (45.8%)	290 (44.8%)	62 (51.2%)
APOE-ϵ4 carrier (%)	424 (55.1%)	352 (54.3%)	72 (59.5%)
Donepezil arm (%)	253 (32.9%)	219 (33.8%)	34 (28.1%)
Vitamin E arm (%)	257 (33.4%)	216 (33.3%)	41 (33.9%)
Placebo arm (%)	259 (33.7%)	213 (32.9%)	46 (38.0%)
ADAS-Cog 11	11.3 (4.4)	10.8 (4.2)	14.1 (4.0)
ADAS-Cog 13	17.7 (6.1)	17.0 (5.9)	21.6 (5.6)
ADAS-Cog Q4	6.3 (2.2)	6.1 (2.2)	7.3 (2.0)
Boston naming	6.9 (2.4)	7.3 (2.2)	5.1 (2.5)
CDR global	0.5 (0.0)	0.5 (0.0)	0.5 (0.0)
CDR sum of boxes	1.8 (0.8)	1.8 (0.8)	2.2 (0.8)
Clock drawing	4.3 (0.9)	4.5 (0.8)	3.4 (1.0)
Digit span bwd	6.2 (2.1)	6.4 (2.1)	5.1 (1.9)
DSST	31.5 (10.9)	33.4 (10.2)	21.1 (8.0)
Logical memory - delayed	3.3 (2.4)	3.5 (2.5)	2.2 (2.0)
Logical memory - immediate	6.2 (3.1)	6.5 (3.1)	4.7 (2.7)
MMSE	27.3 (1.8)	27.5 (1.8)	26.2 (1.7)
Verbal fluency - animals	15.8 (5.2)	16.8 (5.0)	10.5 (3.0)

used to stratify trial participants for a *post hoc* subgroup analysis of treatment effect.

The event-based model of cognitive decline in **Figure 1** is representative of typical (memory-led) Alzheimer's disease, with CDR and impaired memory recall occurring before decline in verbal fluency, working memory, and executive function. Indeed, the estimated sequence shares similarities with results in Firth et al. (2020), which involved an independent cohort. We deliberately excluded ADAS-Cog scores from the model to avoid circularity with the corresponding secondary outcome of the trial (and also to avoid having to perform the relatively arduous ADAS-Cog test at a screening visit). **Supplementary Figure 1** shows that the sequence is largely unchanged with ADAS-Cog features included. Notably, Clock Drawing appears as the first event (before even CDR features), albeit with an additional component of positional density around stages 7–9, supporting the presence of additional heterogeneity among individuals. This result warrants further investigation but is beyond the scope of our study.

The event-based model patient staging mechanism (Young et al., 2014) revealed considerable heterogeneity in the cognitive impairment of MCI participants in both the ADNI observational study (**Figure 3A**) and the ADCS-MCI clinical trial (**Figure 3B**). Such clinical heterogeneity is likely to mask treatment response in clinical trials, particularly if the underlying source is biological heterogeneity relevant to the experimental treatment. The biological underpinnings here are unknown due to the absence of biomarker data in the ADCS-MCI trial, and we need access to such individual-level biomarker data from more recent clinical trials if we are to assess the value of EBM screening vs. biomarker screening. Regardless, we found promising trends in our *post hoc* subgroup analyses (discussed below). Of course, the reduced sample size increases screen-in cost of a clinical trial and potentially diminishes the treatable patient group (affecting also the drug label). This is mostly positive. Pros: a medicine that is effective on a subgroup is better than no medicine at all; not treating non-responders reduces the occurrence of unnecessary side-effects. Con: the smaller group of potential responders limits the treatable

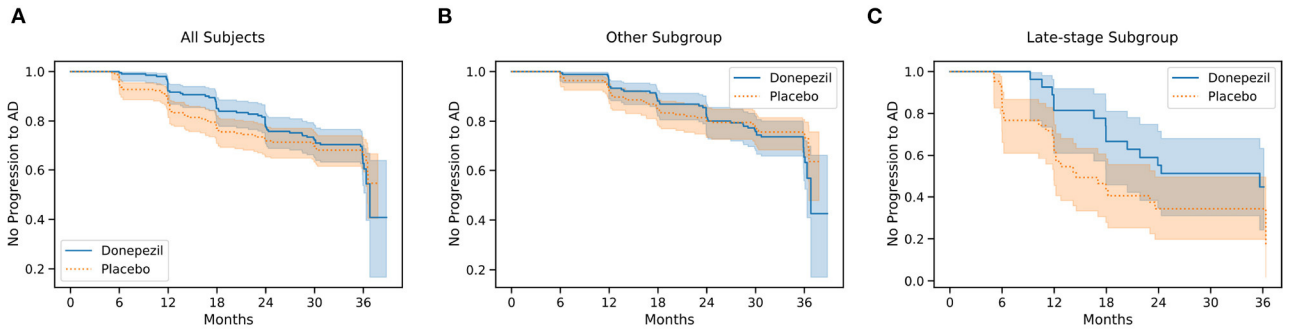


FIGURE 4 | Kaplan–Meier survival curves for all 769 participants **(A)**, the "Others" subgroup **(B)**, and the "Late-stage" subgroup **(C)** in the ADCS-MCI trial.

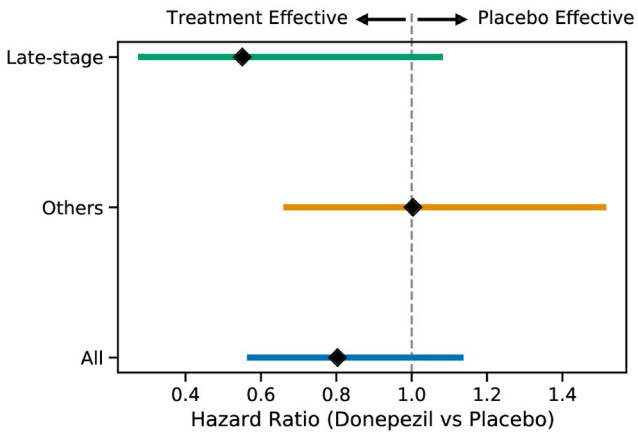


FIGURE 5 | Hazard ratios (with 95% confidence intervals) for the progression to AD for the two subgroups and all subjects when comparing the placebo and donepezil arms.

patient population (but at least those treated are likely to benefit).

In the ADCS-MCI trial we found encouraging trends toward improved survival (**Figure 4**), preserved cognition (**Figure 6**), and a lower hazard ratio (**Figure 5**) in the more severely affected "Late-stage" MCI subgroup ($N = 121$) compared to the less affected "Others" subgroup ($N = 648$). These results suggest that the treatment (donepezil) may protect cognition and provide more protection against MCI conversion to dementia for late-stage MCI. This result concurs with the fact that donepezil is approved for symptomatic relief in more severely affected groups—specifically, dementia patients. Additionally, a recent re-analysis of the ADCS-MCI trial unmasked beneficial effects of donepezil (Edmonds et al., 2018) in a more severely affected subgroup by screening out false-positive MCI participants using hierarchical clustering by Ward's method.

There are multiple possible explanations for why more severely impaired individuals with MCI seem to benefit from donepezil preferentially over less impaired individuals. For one, donepezil may have less cognitive benefit earlier in the disease. Another is that ADAS-Cog might be inadequate to detect such a benefit. Regardless, the key finding is that our approach was able to stratify a clinical trial population into potential responders and non-responders using only baseline/screening data. This supports the notion that computational, data-driven screening can substantially reduce the size (and cost) of a clinical trial, without sacrificing statistical power (see also Franzmeier et al., 2020).

Our work motivates using event-based model staging as a screening tool to enrich clinical trials, but the general principle can be applied using other models that can calculate disease progression scores (e.g., Jedynak et al., 2012; Leoutsakos et al., 2016; Stallard et al., 2017; Wang et al., 2020). While many such works mention the potential application to analyzing clinical trial data, fewer suggest incorporating this into the screening stage, and none (to our knowledge) have actually applied such models in clinical trials, nor in *post hoc* analyses that follow the original analysis protocol to retrospectively determine subgroup treatment effects. Closest to this work is the aforementioned study of the ADCS-MCI trial data by Edmonds et al. (2018), and the work of Schneider et al. (2016), but the approaches used in these studies do not provide an interpretable disease progression signature, nor do they allow for future extension to seamlessly incorporate imaging data and other biomarkers.

In summary, the ADCS-MCI trial was an attempt to test whether donepezil, an approved symptomatic treatment of dementia patients, could slow progression from MCI to dementia. This placebo-controlled, double-blind, phase 3 study found no significant treatment effects (Petersen et al., 2005). Here, we reanalyzed the trial in a *post hoc* subgroup analysis, with the subgroups defined by a data-driven disease progression model: the event-based model (Fonteijn et al., 2012; Young et al., 2014; Firth et al., 2020). Our two key findings are: (1) there was considerable heterogeneity in cognitive impairment in the ADCS-MCI trial, suggesting an inadequate screening protocol; (2) this heterogeneity masked a possible treatment effect in a sample of more severely impaired late-stage MCI participants, despite the likelihood of this smaller sample being under-powered to detect an effect of this magnitude. Our study has highlighted a potential mechanism for improving clinical trial design but the general applicability will require

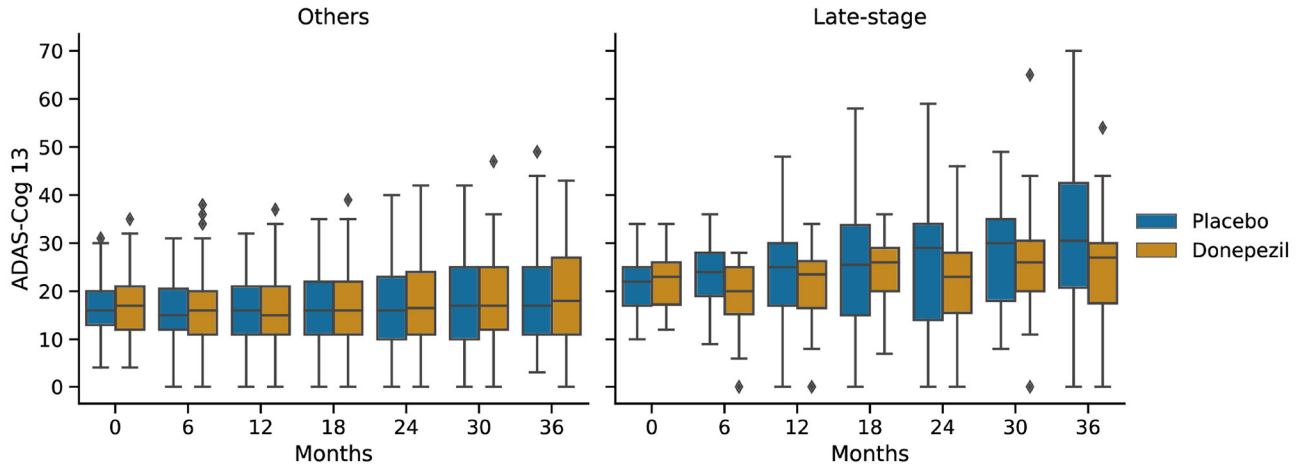


FIGURE 6 | Progression of ADAS-Cog 13 scores in the placebo and donepezil arms throughout the trial for each of the two subgroups.

broader verification, ideally in more recent trials having biomarker data.

In conclusion, our findings support the use of our proposed data-driven screening method to enhance targeting and efficiency of future clinical trials in Alzheimer's disease. What is perhaps most exciting in the immediate future is the prospect of performing similar *post hoc* analyses in other "failed" clinical trials, which could resurrect some Alzheimer's disease drug research programs, saving billions of dollars and years of research. This work is continuing.

AUTHOR'S NOTE

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Data used in the preparation of this manuscript were obtained from the Alzheimer's Disease Cooperative Study legacy database.

DATA AVAILABILITY STATEMENT

The original contributions presented in the study are included in the article/**Supplementary Material**, further inquiries can be directed to the corresponding author.

ETHICS STATEMENT

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

NO conceived of the study and obtained funding. NO and CS performed the data analysis and drafted the manuscript. All authors contributed to interpretation of results and manuscript writing. All authors contributed to the article and approved the submitted version.

FUNDING

NO is a UKRI Future Leaders Fellow. NO and CS acknowledge funding from the UKRI Medical Research Council (MRC MR/S03546X/1). NO, DA, and FB acknowledge funding from the EuroPOND project. This project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under grant agreement no. 666992—and the National Institute for Health Research University College London Hospitals Biomedical Research Centre.

ACKNOWLEDGMENTS

The authors are extremely grateful to the participants of the ADNI observational study and ADCS-MCI trial, without whom this research would not have been possible.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI was funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery

Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd. and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Data collection and sharing for this project was also funded by the Alzheimer's Disease Cooperative Study (ADCS), funded by the National Institutes of Health Grant U19 AG010483.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/frai.2022.660581/full#supplementary-material

REFERENCES

Bilgel, M., and Jedynak, B. M. (2019). Predicting time to dementia using a quantitative template of disease progression. *Alzheimers Dement*. 11, 205–215. doi: 10.1016/j.dadm.2019.01.005

Cummings, J. (2018). Lessons learned from Alzheimer disease: clinical trials with negative outcomes. *Clin. Transl. Sci*. 11, 147–152. doi: 10.1111/cts.12491

Cummings, J., Lee, G., Mortsdorf, T., Ritter, A., and Zhong, K. (2017). Alzheimer's disease drug development pipeline: 2017. *Alzheimers Dement*. 3, 367–384. doi: 10.1016/j.trci.2017.05.002

Cummings, J., Lee, G., Ritter, A., Sabbagh, M., and Zhong, K. (2019). Alzheimer's disease drug development pipeline: 2019. *Alzheimers Dement*. 5, 272–293. doi: 10.1016/j.trci.2019.05.008

Cummings, J., Lee, G., Ritter, A., Sabbagh, M., and Zhong, K. (2020). Alzheimer's disease drug development pipeline: 2020. *Alzheimers Dement*. 6, e12050. doi: 10.1002/trc2.12050

Cummings, J., Lee, G., Ritter, A., and Zhong, K. (2018a). Alzheimer's disease drug development pipeline: 2018. *Alzheimers Dement*. 4, 195–214. doi: 10.1016/j.trci.2018.03.009

Cummings, J., Morstorf, T., and Lee, G. (2016). Alzheimer's drug-development pipeline: 2016. *Alzheimers Dement*. 2, 222–232. doi: 10.1016/j.trci.2016.07.001

Cummings, J., Ritter, A., and Zhong, K. (2018b). Clinical trials for disease-modifying therapies in Alzheimer's disease: a primer, lessons learned, and a blueprint for the future. *J. Alzheimers Dis*. 64, S3–S22. doi: 10.3233/JAD-179901

Cummings, J. L., Morstorf, T., and Zhong, K. (2014). Alzheimer's disease drug-development pipeline: few candidates, frequent failures. *Alzheimers Res. Therapy* 6, 37. doi: 10.1186/alzrt269

Edmonds, E. C., Ard, M. C., Edland, S. D., Galasko, D. R., Salmon, D. P., and Bondi, M. W. (2018). Unmasking the benefits of donepezil via psychometrically precise identification of mild cognitive impairment: a secondary analysis of the ADCS vitamin e and donepezil in MCI study. *Alzheimers Dement*. 4, 11–18. doi: 10.1016/j.trci.2017.11.001

Eshaghi, A., Marinescu, R. V., Young, A. L., Firth, N. C., Prados, F., Jorge Cardoso, M., et al. (2018). Progression of regional grey matter atrophy in multiple sclerosis. *Brain* 141, 1665–1677. doi: 10.1093/brain/awy088

Firth, N. C., Primativo, S., Brotherhood, E., Young, A. L., Yong, K. X., Crutch, S. J., et al. (2020). Sequences of cognitive decline in typical Alzheimer's disease and posterior cortical atrophy estimated using a novel event-based model of disease progression. *Alzheimers Dement*. 16, 965–973. doi: 10.1002/alz.12083

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., et al. (2012). An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *Neuroimage* 60, 1880–1889. doi: 10.1016/j.neuroimage.2012.01.062

Franzmeier, N., Koutsouleris, N., Benzinger, T., Goate, A., Karch, C. M., Fagan, A. M., et al. (2020). Predicting sporadic Alzheimer's disease progression via inherited Alzheimer's disease-informed machine-learning. *Alzheimers Dement*. 16, 501–511. doi: 10.1002/alz.12032

Hardy, J., and Higgins, G. (1992). Alzheimer's disease: the amyloid cascade hypothesis. *Science* 256, 184–185. doi: 10.1126/science.1566067

Jack, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., et al. (2010). Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *Lancet Neurol*. 9, 119–128. doi: 10.1016/S1474-4422(09)70299-6

Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., et al. (2012). A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease neuroimaging initiative cohort. *Neuroimage* 63, 1478–1486. doi: 10.1016/j.neuroimage.2012.07.059

Leoutsakos, J.-M., Gross, A., Jones, R., Albert, M., and Breitner, J. (2016). "Alzheimer's progression score": development of a biomarker summary outcome for ad prevention trials. *J. Prev. Alzheimers Dis*. 3, 229. doi: 10.14283/jpad.2016.120

Marinescu, R. V., Oxtoby, N. P., Young, A. L., Bron, E. E., Toga, A. W., Weiner, M. W., et al. (2018). TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease. *arXiv preprint arXiv:1805.03909*. doi: 10.48550/arXiv.1805.03909

Morris, G. P., Clark, I. A., and Vissel, B. (2014). Inconsistencies and controversies surrounding the amyloid hypothesis of Alzheimer's disease. *Acta Neuropathol. Commun*. 2, 135. doi: 10.1186/PREACCEPT-1342777270140958

Oxtoby, N., Leyland, L., Aksman, L., Thomas, G., Bunting, E., Wijeratne, P., et al. (2021). Sequence of clinical and neurodegeneration events in parkinson's disease progression. *Brain J. Neurol*. 144, 975–988. doi: 10.1093/brain/awaa461

Oxtoby, N. P., Alexander, D. C., and EuroPOND Consortium (2017). Imaging plus X: multimodal models of neurodegenerative disease. *Curr. Opin. Neurol*. 30, 371–379. doi: 10.1097/WCO.0000000000000460

Oxtoby, N. P., Young, A. L., Cash, D. M., Benzinger, T. L. S., Fagan, A. M., Morris, J. C., et al. (2018). Data-driven models of dominantly-inherited Alzheimer's disease progression. *Brain* 141, 1529–1544. doi: 10.1093/brain/awy050

Petersen, R. C., Thomas, R. G., Grundman, M., Bennett, D., Doody, R., Ferris, S., et al. (2005). Vitamin E and donepezil for the treatment of mild cognitive impairment. *N. Engl. J. Med*. 352, 2379–2388. doi: 10.1056/NEJMoa050151

Salloway, S., Sperling, R., Fox, N. C., Blennow, K., Klunk, W., Raskind, M., et al. (2014). Two phase 3 trials of bapineuzumab in mild-to-moderate Alzheimer's disease. *N. Engl. J. Med*. 370, 322–333. doi: 10.1056/NEJMoa1304839

Schneider, L. S., Frangakis, C., Drye, L. T., Devanand, D., Marano, C. M., Mintzer, J., et al. (2016). Heterogeneity of treatment response to citalopram for patients with Alzheimer's disease with aggression or agitation: the citad randomized clinical trial. *Am. J. Psychiatry* 173, 465–472. doi: 10.1176/appi.ajp.2015.15050648

Stallard, E., Kinosian, B., and Stern, Y. (2017). Personalized predictive modeling for patients with Alzheimer's disease using an extension of Sullivan's life table model. *Alzheimers Res. Therapy* 9, 75. doi: 10.1186/s13195-017-0302-6

Wang, R., Lagakos, S. W., Ware, J. H., Hunter, D. J., and Drazen, J. M. (2007). Statistics in medicine–reporting of subgroup analyses in clinical trials. *N. Engl. J. Med*. 357, 2189–2194. doi: 10.1056/NEJMsr077003

Wang, Z., Tang, Z., Zhu, Y., Pettigrew, C., Soldan, A., Gross, A., et al. (2020). Ad risk score for the early phases of disease based on unsupervised machine learning. *Alzheimers Dement*. 16, 1524–1533. doi: 10.1002/alz.12140

Wijeratne, P. A., Young, A. L., Oxtoby, N. P., Marinescu, R. V., Firth, N. C., Johnson, E. B., et al. (2018). An image-based model of brain volume biomarker changes in huntington's disease. *Ann. Clin. Transl. Neurol*. 5, 570–582. doi: 10.1002/acn3.558

Young, A. L., Marinescu, R. V., Oxtoby, N. P., Bocchetta, M., Yong, K., Firth, N. C., et al. (2018). Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat. Commun*. 9, 4273. doi: 10.1038/s41467-018-05892-0

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., et al. (2014). A data-driven model of biomarker changes in sporadic Alzheimers disease. *Brain* 137, 2564–2577. doi: 10.1093/brain/awu176

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Advantages of publishing in Frontiers

OPEN ACCESS
Articles are free to read for greatest visibility and readership

FAST PUBLICATION
Around 90 days from submission to decision

HIGH QUALITY PEER-REVIEW
Rigorous, collaborative, and constructive peer-review

TRANSPARENT PEER-REVIEW
Editors and reviewers acknowledged by name on published articles

Frontiers
Avenue du Tribunal-Fédéral 34
1005 Lausanne | Switzerland

Visit us: www.frontiersin.org
Contact us: frontiersin.org/about/contact

REPRODUCIBILITY OF RESEARCH
Support open data and methods to enhance research reproducibility

DIGITAL PUBLISHING
Articles designed for optimal readership across devices

FOLLOW US
@frontiersin

IMPACT METRICS
Advanced article metrics track visibility across digital media

EXTENSIVE PROMOTION
Marketing and promotion of impactful research

LOOP RESEARCH NETWORK
Our network increases your article's readership