A large-scale multi-centre study characterising atrophy heterogeneity in Alzheimer's disease

Vikram Venkatraghavan[1,2], Damiano Archetti[3], Pierrick Bourgeat[4], Chenyang Jiang[1,2], Mara ten Kate[5], Anna C. van Loenhoud[1,2], Rik Ossenkoppele[1,2,6], Charlotte E. Teunissen[2,7], Elsmarieke van de Giessen[5,8], Yolande A.L. Pijnenburg[1,2], Giovanni B. Frisoni[9,10], Béla Weiss[11,12], Zoltán Vidnyánszky[11], Tibor Auer[11,13], Stanley Durrleman[14], Alberto Redolfi[3], Simon M. Laws[15], Paul Maruff[16], for the Australian Imaging Biomarkers and Lifestyle Study[†], for the Alzheimer's Disease Neuroimaging Initiative[‡], for the E-DADS Consortium, Neil P. Oxtoby[17], Andre Altmann[17], Daniel C. Alexander[17], Wiesje M. van der Flier[1,2,18], Frederik Barkhof[5,17,19], Betty M. Tijms[1,2]

[1] Alzheimer Centre Amsterdam, Neurology, Vrije Universiteit, Amsterdam UMC, location VUmc, Amsterdam, the Netherlands

[2] Amsterdam Neuroscience, Neurodegeneration, Amsterdam, the Netherlands

[3] Laboratory of Neuroinformatics, IRCCS Istituto Centro San Giovanni di Dio Fatebenefratelli, Brescia, Italy

[4] The Australian e-Health Research Centre, CSIRO Health and Biosecurity, Brisbane, Queensland, Australia

[5] Department of Radiology and Nuclear Medicine, Amsterdam Neuroscience, Amsterdam University Medical Center, Location VUmc, Amsterdam, the Netherlands

[6] Clinical Memory Research Unit, Lund University, Sweden

[7] Neurochemistry Laboratory, Department of Laboratory Medicine, Vrije Universiteit Amsterdam, Amsterdam UMC location VUmc, Amsterdam, the Netherlands

[8] Amsterdam Neuroscience, Brain Imaging, Amsterdam, the Netherlands

[9] Laboratory of Neuroimaging of Aging (LANVIE), University of Geneva, Geneva, Switzerland

[10] Geneva Memory Center, Department of Rehabilitation and Geriatrics, Geneva University Hospitals, Geneva, Switzerland

[11] Brain Imaging Centre, HUN-REN Research Centre for Natural Sciences, Budapest, Hungary

[12] Machine Perception Research Laboratory, HUN-REN Institute for Computer Science and Control , Budapest, Hungary

[13] School of Psychology, University of Surrey, Guildford, United Kingdom

[14] Sorbonne Université, Institut du Cerveau - Paris Brain Institute – ICM, CNRS, Inria, Inserm, AP-HP, Hôpital Pitié-Salpêtrière, Paris, France

[15] Centre for Precision Health, Edith Cowan University, Joondalup, Western Australia, Australia

[16] Cogstate Ltd., Melbourne, Victoria, Australia

[17] Centre for Medical Image Computing, Department of Medical Physics and Biomedical Engineering and Department of Computer Science, University College London, UK

[18] Department of Epidemiology and Data Science, Vrije Universiteit, Amsterdam UMC, location VUmc, the Netherlands

[19] Queen Square Institute of Neurology, University College London, UK

† Data used in the preparation of this article was obtained from the Australian Imaging Biomarkers and Lifestyle (AIBL) Study. Unless named, the AIBL researchers contributed data but did not participate in analysis or writing of this report. AIBL researchers are listed at https://aibl.org.au

‡ 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:

adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf.

Abstract

Background: Previous studies reported on the existence of atrophy-based Alzheimer's disease (AD) subtypes that associate with distinct clinical symptoms. However, the consistency of AD atrophy subtypes across approaches remains uncertain. This large-scale study aims to assess subtype concordance in individuals using two methods of data-driven subtyping.

Methods: We included $n = 10{,}011$ patients across the clinical spectrum from 10 AD cohorts across Europe, United States, and Australia, and extracted regional volumes using Freesurfer v7.1.1. To characterise atrophy heterogeneity in the AD continuum, we introduced a hybrid two-step approach called Snowphlake (Staging NeurOdegeneration With PHenotype informed progression timeLine of biomarKErs) to identify subtypes and sequence of atrophy-events within each subtype. We compared our results with SuStaIn (Subtype and Stage Inference) which jointly estimates both, and was trained and validated similarly. The training dataset included Aβ+ participants ($n = 1{,}195$), and a control group of Aβ− cognitively unimpaired participants ($n = 1{,}692$). We validated model staging within each subtype, in a held-out clinical-validation dataset ($n = 6{,}362$) comprising patients across the clinical spectrum irrespective of Aβ biomarker status and an independent external dataset ($n = 762$). Furthermore, we validated the clinical significance of the detected subtypes, in a subset of Aβ+ validation datasets with $n = 1{,}796$ in the held-out sample and $n = 159$ in the external dataset. Lastly, we performed concordance analysis to assess the consistency between the methods.

Results: In the AD dementia (AD-D) training data, Snowphlake identified four subtypes: diffuse cortical atrophy ($21.1\%, age\ 67.5 \pm 9.3$), parieto-temporal atrophy ($19.8\%, age\ 60.9 \pm 7.9$), frontal atrophy ($24.8\%, age\ 67.6 \pm 8.8$) and subcortical atrophy ($25.1\%, 68.3 \pm 8.2$). The subtypes assigned in Aβ+ validation datasets were associated with alterations in specific cognitive domains (Cohen's *f:* $[0.15 - 0.33]$), while staging correlated with Mini-Mental State Examination (MMSE) scores (R: $[-0.51\ to\ -0.28]$) in the validation datasets. SuStaIn also identified four subtypes: typical ($55.7\%, age\ 66.7 \pm 7.8$), limbic-predominant ($24.2\%, age\ 72.2 \pm 6.6$), hippocampal-sparing ($14.6\%, age\ 62.8 \pm 6.9$), and subcortical ($0.8\%, age\ 68.2 \pm 7.6$). The subtypes assigned in Aβ+ validation datasets using SuStaIn were also associated with alterations in specific cognitive domains (Cohen's *f*: $[0.17 - 0.34]$), while staging correlated with MMSE scores in the validation datasets (R: $[-0.54\ to\ -0.26]$). However, we observed low concordance between Snowphlake and

SuStaIn, with 39.7% of AD-D patients consistently grouped in concordant subtypes by both the methods.

Conclusion: In this multi-cohort study, both Snowphlake and SuStaIn identified four subtypes that were associated with different symptom profiles and atrophy-severity measures that were associated with global cognition. The low concordance between Snowphlake and SuStaIn suggests that heterogeneity may rather be a spectrum than discretised by subtypes.

Keywords: Alzheimer's disease; heterogeneity; subtypes; data-driven; MRI

Introduction

Alzheimer's disease (AD) is the leading cause of dementia.[1] It is characterised by progressive loss of brain volume (atrophy) and cognitive decline. Across individuals with AD, there is substantial variability in severity and pattern of brain atrophy,[2-5] as well as in the symptoms that AD patients manifest.[6,7] Understanding the variability in brain atrophy between patients, and how they explain differences in cognitive symptoms, could improve tailored patient care management.

One approach to study heterogeneity in atrophy patterns is by data-driven analysis of structural magnetic resonance imaging (MRI) that quantify *in-vivo* atrophy patterns in AD patients. Previous studies taking this approach, using different techniques, identified either three subtypes[8,9,10] or four subtypes[4,11] of AD. The most frequently identified subtypes include a medial temporal lobe (MTL) atrophy subtype and hippocampal-sparing subtype.[2,4,8,9,11] Other subtypes that have been identified include subcortical atrophy subtype[2,9], parieto-occipital atrophy subtype[4], cortical atrophy subtype[2,4,9], and minimal atrophy subtype.[4,11] Although these findings suggest that atrophy-based subtypes may represent robust biological entities, there remains inconsistency in the specific subtypes found, number of subtypes found, and in their associations with clinical symptoms. Possibly, this may be explained by difference in methodology used for subtyping, but so far remains unclear to what extent different subtyping methodologies converge on identifying the same subtypes when performed in the same patient population.

Apart from distinct patterns of atrophy, studies have identified another dimension that contributes to atrophy heterogeneity i.e. severity of atrophy (also referred to as atrophy stage).[2,3] Consequently, it remains a challenge to reliably identify data-driven subtypes that reflect meaningful phenotypic differences independent of disease severity, which might further explain the inconsistencies in atrophy subtypes observed across studies. To overcome this challenge, data-driven disease progression models (DPMs),[12] such as SuStaIn[2] and Disease Course Mapping[13], have been developed to identify subtypes and severity jointly within a single framework. However, these methods remain computationally expensive and thus use a limited number of volumetric (or thickness) markers. Other studies have used regular machine-learning (ML) approaches for subtyping by selecting patients within the same clinical stage of AD[4,9]. While the regular ML approaches are computationally efficient as compared to DPMs and thus scalable to large cohorts and markers with greater spatial

resolution, regular ML methods do not account for atrophy severity. To address this drawback, in this work, we combined a well-validated ML approach for AD subtyping using non-negative matrix factorization (NMF)[4,14] with a scalable disease progression model called discriminative event-based model (DEBM)[15,16] for estimating severity. The resulting hybrid-method was termed Snowphlake (Staging NeurOdegeneration With PHenotype informed progression timeLine of biomarKErs) and this was used to study AD heterogeneity and compare our results with those obtained using SuStaIn.

In this large-scale multi-centre study including $n = 10{,}011$ participants from 10 cohorts across Europe, United States, and Australia, we first characterised atrophy heterogeneity in the AD continuum using Snowphlake and compared our results with SuStaIn, trained and validated similarly. Second, we studied how the data-driven estimates of atrophy heterogeneity for each method were related to the cognitive symptoms that patients experience. Finally, we examined the concordance between the subtype assigned by Snowphlake and SuStaIn.

Methods

Study participants

We selected participants with a clinical diagnosis of AD dementia (AD-D), mild cognitive impairment (MCI), subjective cognitive decline (SCD), or were cognitively normal (CN) when they had a 3D T1w MRI scan available, from 10 cohorts across Europe, United States of America, and Australia. The included cohorts were: Amsterdam Dementia Cohort (ADC)[17], Alzheimer's Disease Neuroimaging Initiative (ADNI)[18], Australian Imaging Biomarker & Lifestyle Flagship Study of Ageing (AIBL)[19], National Alzheimer's Coordinating Center (NACC)[20], Open Access Series of Imaging Studies (OASIS)[21], Alzheimer's Repository Without Borders (ARWiBo)[22], European DTI Study on Dementia (EDSD)[23], Italian Alzheimer's Disease Neuroimaging Initiative (I-ADNI)[24], European Alzheimer's Disease Neuroimaging Initiative (also known as PharmaCOG)[25], and the Geneva memory-centre cohort (GMC)[26]. The characteristics of each cohort are summarized in Supplementary Table 1. ADNI data used in the preparation of this article were obtained from the database adni.loni.usc.edu. Further details about ADNI are mentioned in the Supplementary methods section S1.1. The institutional review boards of all participating institutes approved the protocol for data collection and its subsequent use in retrospective analyses. The clinical diagnosis of participants in each cohort was performed by the different study teams according to international criteria and have been described in detail in each of those cohorts. In the present study we grouped the CN and SCD participants together as cognitively unimpaired (CU).

Study data, MRI processing and harmonization

Across cohorts, baseline 3D T1w MRI scans were acquired with 44 different MRI scanners, with varied image acquisition protocols. Supplementary Table 2 gives an overview of the scanners included in this study. Cortical reconstruction and volumetric segmentation were performed with a Docker container of Freesurfer v7.1.1 in 3 different centres (ADC and NACC in Amsterdam, ADNI and AIBL in Brisbane, and the rest in Brescia) to extract volumes of 68 cortical regions as per the Desikan-Killiany atlas and 14 subcortical brain regions. Automated quality control for Freesurfer segmentations utilized the Euler number,[27,28] with outlier thresholds determined independently for each scanner. These thresholds were based on the interquartile range (IQR) specific to each scanner, where outliers were identified as 1.5×IQR below the first quartile.[27,28] Furthermore, subjects with total intracranial volume (TIV) greater than the threshold of 1.5×IQR above the third quartile

computed independently for males and females, were identified as outliers. These outliers were excluded from further analysis in this study. The number of participants excluded based on these two criteria were $n = 1{,}198\ (10.7\%)$, leaving a total number of scans of $n = 10{,}011$ participants included for subsequent analyses.

We harmonized cortical and subcortical volumes by removing scanner related batch effects while preserving the effects of age, sex, and clinical stage. In our analysis, we used ComBat harmonization[29] with empirical Bayes optimization to remove batch-related effects, with the largest single-scanner data from the ADNI cohort (Siemens TrioTim 3T scanner, $n = 257$) used as a reference batch. Finally, because SuStaIn is a computationally intensive algorithm and prior subtyping studies using SuStaIn have used between 12 to 21 input features[2,30], we reduced the number of cortical areas by constructing 24 composite regions, comprising 17 composite cortical ROIs (details of the mapping to derive these composite ROI volume from Freesurfer cortical parcellation are tabulated in Supplementary Table 3) and 7 subcortical regions (namely: Thalamus, Caudate, Putamen, Pallidum, Hippocampus, Amygdala, and Accumbens-area). We corrected for the effects of total intracranial volume and normal aging by regressing out the effects that were estimated in the Aβ− CU participants (see the next section for details on determining amyloid status). The harmonized volumes were combined using the sum of left and right counterparts. These volumes were converted to w-scores (covariate-adjusted z-score) based on the mean and standard deviation of Aβ− CU participants in the study.

Amyloid Status

Where information about amyloid markers was available, individuals were labelled as having a normal or abnormal amyloid biomarker (Aβ−/ Aβ+ for normal/abnormal respectively) based on either cerebrospinal fluid (CSF, available in ADC, ADNI, ARWiBo, EDSD, PharmaCog, and in NACC after 2015), positron emission tomography (PET) images, or pathological examination (NACC). CSF testing and PET imaging performed during the baseline visit (within a timeframe of 90 days of MRI) were considered for this purpose. Positivity in PET images was determined by either visual readouts by radiologists (available in ADC and GMC), centiloid values (available in ADNI, AIBL, cut-off = 30),[31] or a combination of the two (in NACC after 2015). The cut-off points for Aβ positivity based on CSF were defined for each cohort independently based on $A\beta_{1-42}$ concentrations. The details of cut-off point selection and assays used are in Supplementary Section S1.2. Details of the Aβ PET processing pipeline and the tracers used are in Supplementary Section S1.3. In ADC,

ADNI, and NACC, participants were considered Aβ+ if either one of CSF or PET were positive. In pre-2015 NACC cohort, due to the absence of either of these biomarkers, autopsy-confirmed AD-related neuro-pathologic change (ADNC) based on ABC summary score[32] (comprising Aβ plaque score, modified Braak stage, modified CERAD score) was used to define Aβ positivity in patients, when available. These scores were categorized as either non-AD, or graded as low, intermediate, or high ADNC in the NACC cohort. In this study, MCI and AD-D participants with low to high ADNC were termed Aβ positive. Participants for whom amyloid status was unavailable were excluded from training the models, and their inclusion in the validation experiments is detailed in the study design.

Cognitive-data preparation

Neuropsychological test batteries assessing the cognitive domains of episodic memory, attention and executive function, language, and visuospatial function were used to compute composite scores for these domains. Cognitive tests performed during the baseline visit (within a timeframe of 90 days of MRI) were considered for this purpose. Aβ− CU participants' data were used as a reference group for computing these composite scores. The methodological details of computing the cognitive domain scores in each of our cohorts are included in Supplementary material Section S1.4. We computed the domain scores in the cohorts of ADC, ADNI, AIBL, NACC, and GMC in our analysis. Cognitive test data in the remaining cohorts were not available to us. In the GMC cohort, the language domain score was not computed as the cognitive test battery in that cohort did not include any associated tests for assessing language. Since the different cohorts had different neuropsychological tests to assess the patients, we computed the domain scores independently in the different cohorts, with the Aβ− CU participants in that cohort serving as a reference group to compute z-scores for individual tests. Subsequently, for each domain, multiple test scores belonging to a specific domain were averaged to compute the domain score.

Study design

We divided our combined cohorts into three different datasets: training dataset, held out clinical-validation dataset, and an independent external dataset. A subset of the clinical-validation dataset and the external dataset with Aβ+ participants was further selected for a few experiments (Aβ+ validation dataset). Figure 1 gives a graphical overview of the study design described here.

The training dataset comprised 40% of the combined Aβ+ participants randomly selected from six cohorts (ADC, ADNI, AIBL, NACC, ARWiBO, EDSD). With the aim of creating atrophy-based subtyping models that are equally generalizable to AD patients across all ages, we ensured the training set had a uniform age distribution. Hence the participants were selected in the training dataset based on weighted random sampling without replacement, with weights inversely proportional to the age distribution in each clinical stage. Moreover, we also included Aβ− CU participants in all the cohorts except GMC to serve as a reference group for training the models.

The held-out clinical-validation dataset consisted of all the participants not included in the training dataset or the reference group from ADC, ADNI, AIBL, NACC, ARWiBO, EDSD, I-ADNI, OASIS, and PharmaCOG. The GMC cohort was chosen as the independent external-validation dataset. The difference between the held-out clinical-validation dataset and the independent external-validation dataset is that for the training dataset all the Aβ+ participants from the GMC cohort were excluded.

The Aβ+ validation datasets comprised the remaining 60% Aβ+ participants not included in training from the aforementioned six cohorts (ADC, ADNI, AIBL, NACC, ARWiBO, EDSD) and 100% Aβ+ participants in the external dataset.

Characterising atrophy heterogeneity

We used two data-driven approaches for estimating atrophy subtypes and severity: Snowphlake and SuStaIn. Snowphlake is a hybrid method we introduce using non-negative matrix factorization (NMF)[4,14] for subtyping followed by DEBM[15,16] for estimating sequence of atrophy-events *within* each subtype. Although each component of this approach has independently been validated before, this is the first study to jointly use them for the purpose of subtype and severity estimation. To ensure easy reproducibility of this approach, we built a python software toolbox: https://github.com/snowphlake-dpm/snowphlake. SuStaIn is a disease progression modelling technique developed previously[2], with an existing python software package.[33] The key conceptual difference is that Snowphlake is a two-step subtype-then-stage approach, while SuStaIn estimates both subtype and stage jointly.

Snowphlake: The subtyping model was trained on Aβ+ AD-D participants, using the non-smooth variant of non-negative matrix factorization (ns-NMF)[4,34] with KL-divergence as a distance metric. Ns-NMF is a stochastic dual-clustering approach that is designed to estimate sparse clusters in the data. With different random initializations resulting in slightly different

subtypes, ns-NMF was run n_{run} times on the training data, where $n_{run} = 25 \times n_{AD-D}$. Here, n_{AD-D} was the number of Aβ+ AD-D participants in the training data. The run with the least residual of subtyping (res_k) was chosen as the optimal solution, where k is the number of subtypes. For choosing the optimal number of subtypes (n_{opt}), a random permutation of the training data was subsequently subtyped. k_{opt} is chosen such that $\Delta residual = res_{k-1} - res_k$ for the training data is higher than that in the random permutations. On subtyping, each participant is assigned a weight for each subtype. These subtype weights were further used to detect outliers within each subtype based on minimum covariance determinant algorithm[35] with Mahalanobis distance metric.

Next, based on the identified subtypes, we assigned the subtypes of Aβ+ MCI and Aβ+ CU participants in the training dataset. We then determined the sequences of atrophy-events for each subtype using co-initialized discriminative event-based model (DEBM).[15,16] Briefly, Gaussian mixture modelling (GMM) was used to estimate the probabilities for each region to be abnormal for each participant, with Aβ− CU group considered as a reference group for GMM. These probabilistic abnormality values were used to infer a sequence for each Aβ+ participant in the training data. These individual estimates were aggregated using generalized Mallows model[15] to estimate the sequence of atrophy-events for each subtype. Further details about training DEBM are in Supplementary Section S1.5.

SuStaIn: We trained SuStaIn on the same training data as used in Snowphlake, with the pySuStaIn toolbox[33]. We used the cross-validation information criteria (CVIC) for selecting optimum number of subtypes, with $w = -1$ and $w = -2$ chosen as event thresholds. The methodological details of the SuStaIn approach has been described in detail in *Young et al*[2]. For the sake of completeness, the method has been briefly described in Supplementary Section S1.6.

For both Snowphlake and SuStaIn, the trained models were used to assign atrophy-based subtype and stage to participant data in the different validation datasets.

Statistical analysis to characterise subtypes and evaluate concordance of assigned subtypes

The subtype and staging measures assigned in the Aβ+ validation dataset, clinical-validation dataset, and external-validation dataset by both methods, were used further for investigating if these measures were associated with symptom profile and severity respectively.

Experiment 1: Validating the estimated staging

To evaluate the staging system of Snowphlake and SuStaIn, the trained models were used to assign the subtypes and stages of all participants in the Aβ+ validation dataset, clinical-validation dataset, and external-validation dataset. The assigned stage within each subtype was used to compute Pearson's correlation with Mini-Mental Status Examination (MMSE), as a proxy for disease severity.

Experiment 2: Comparison of subtypes on cognitive symptoms

In the absence of ground-truth in data-driven subtyping, we used the association of the identified subtypes with the patients' cognitive-symptom profile, to determine their validity. We performed analysis of variance (ANOVA) tests in MCI and AD-D patients in Aβ+ validation dataset and Aβ+ subset of the external-validation dataset to determine if subtypes differed in terms of deficits in specific cognitive domains, after correction for confounding effects of age, sex, and level of education. These statistical tests were performed for both subtyping methods in the validation datasets, independently in each of ADC, ADNI, NACC, AIBL, and GMC. Lastly, we performed random-effect meta-analysis pooling the results of independent cohorts and accounted for multiple testing using false discovery rate (FDR) correction.

Experiment 3: Concordance between Snowphlake and SuStaIn

The motivation to investigate concordance between the methods was to go beyond group-level definitions of subtypes to individuals assigned to these subtypes. High concordance between the two methods would indicate individual patients in different subtypes have distinct atrophy pattern much like their group-level definitions, while low concordance would indicate individual atrophy patterns vary substantially even within each subtype. To quantify the concordance between the two methods, we constructed contingency matrices of participant subtypes by Snowphlake and SuStaIn for Aβ+ CU, MCI, and AD-D in the training and in the validation dataset. Concordant subtype-pairs are defined as the Snowphlake subtype most frequently co-occurring with SuStaIn subtypes identified in Aβ+ AD-D patients.

We performed additional analyses for identifying atrophy pattern determinants for participants to not be grouped into concordant subtype-pairs. To this end, we created average w-score volume maps for all AD-D participants in the concordant subtype-pairs and their less frequently co-occurring counterparts. We performed a t-test between these average w-score

maps, to investigate if there are significant differences in atrophy patterns that influence concordance between the two methods. The null hypothesis in this test is that there is no difference in average w-scores of participants within a specific SuStaIn subtype, grouped in different subtypes of Snowphlake. Lastly, we estimated the sequence of atrophy-events in the concordant subtype-pairs using DEBM, the methodological equivalent of Snowphlake with 1-subtype and w-score EBM, the methodological equivalent of SuStaIn with 1-subtype.

Results

The demographics of participants and their amyloid status are summarized in Table 1. Overall, our combined dataset (from 10 cohorts) consisted of $n = 3{,}150$ Aβ+ participants ($n_{ADD} = 1{,}525$; $n_{MCI} = 1{,}150$; $n_{CU} = 475$), $n = 2{,}568$ Aβ− participants ($n_{ADD} = 131; n_{MCI} = 706;\ n_{CU} = 1{,}731$), and $n = 4{,}293$ participants with unknown Aβ status ($n_{ADD} = 1{,}264; n_{MCI} = 1{,}360;\ n_{CU} = 1{,}669$). This combined dataset was divided into a training dataset, held-out validation dataset, and an external-validation dataset. The training dataset consisted of $n = 1{,}195$ Aβ+ participants ($n_{ADD} = 596;\ n_{MCI} = 416; n_{CU} = 183$) and $n = 1{,}692$ Aβ− CU reference group participants. The held-out validation dataset consisted of $n = 6{,}362$ participants across the clinical spectrum ($n_{ADD} = 2{,}187;\ n_{MCI} = 2{,}381;\ n_{CU} = 1{,}794$) and the external dataset consisted of $n = 723$ patients ($n_{ADD} = 137;\ n_{MCI} = 419;\ n_{CU} = 167$) and $n = 39$ Aβ− CU reference group participants. A subset of participants in the validation datasets with Aβ+ status (Aβ+ validation dataset) consisted of $n = 1{,}796$ participants in the internal cohorts ($n_{ADD} = 894;\ n_{MCI} = 626;\ n_{CU} = 276$) and $n = 159$ participants in the external cohort.

The age of the $n = 1{,}525$ Aβ+ AD-D patients included in our study was 66.8 ± 8.7 years (see Supplementary Figure 1), with ADC predominantly being a young-onset AD cohort, while the rest being predominantly late-onset AD cohorts. Supplementary Figure 1 also shows the age distribution in the different clinical stages within the Aβ+ patient population and in our selected training dataset. $52.2\%\ (5226/10{,}011)$ of the included participants were women, while $47.6\%\ (569/1195)$ of the Aβ+ patients included in the training dataset were women. Furthermore, all the imaging markers used in this study except Pallidum volume were different between the Aβ+ AD-D patients and Aβ− CU reference group with $p > 0.05$ for Pallidum and $p < 10^{-5}$ for all other markers, after correcting for multiple testing with FDR.

Subtypes identified with Snowphlake and SuStaIn

Snowphlake and SuStaIn each identified four subtypes. Supplementary Figure 2 shows the criteria used for selecting the optimum number of subtypes for each modelling technique ($\Delta residual$ for Snowphlake, CVIC for SuStaIn). For SuStaIn, the CVIC value for the 5-subtype solution was marginally better than the 4-subtype solution. However, only $3/1{,}195$ Aβ+ patients in the training data belonged to 5th subtype. We hence chose the 4-subtype solution for our further analysis.

The atrophy subtypes identified by Snowphlake, along with the prevalence of each subtype and age distribution among AD-D Aβ+ patients in the training and Aβ+ validation datasets were: Diffuse cortical atrophy subtype ($Training$: 21.6% ($n = 129/596$), $age = 66.5 \pm 7.8$; Aβ + $Validation$: 21.1% ($n = 189/894$), $age = 67.5 \pm 9.4$), Parieto-temporal atrophy subtype ($Training$: 19.2% ($n = 115/596$), $age = 63.1 \pm 6.9$; Aβ + $Validation$: 19.7% ($n = 177/896$), $age = 60.9 \pm 7.9$), Frontal atrophy subtype ($Training$: 25.5% ($n = 152/596$), $age = 68.3 \pm 7.9$; Aβ + $Validation$: 24.8% ($n = 222/894$), $age = 67.6 \pm 8.9$), and Subcortical atrophy subtype ($Training$: 24.8% ($n = 148/596$), $age = 70.0 \pm 7.2$; Aβ + $Validation$: 25.2% ($n = 225/894$), $age = 68.3 \pm 8.3$), with prominent temporal lobe atrophy in each of the identified subtypes. Apart from these subtypes, an additional outlier group not assigned to any subtype was detected (Training: 8.7%; Aβ + $Validation$: 9.1%). Figure 2 depicts the sequence of atrophy-events estimated for each subtype by Snowphlake. Supplementary Figure 3 shows the uncertainty in these estimates.

The prevalence, age and MMSE distribution, and the percentage of *APOE4* carriers in each of these atrophy subtypes across the different clinical stages in the pooled validation datasets (held-out validation and external validation pooled together) are summarized in Table 2 and these results in each cohort independently are reported in Supplementary Table 4. Age of onset of AD-D differed significantly ($p < 0.05$) between the four identified subtypes in the pooled validation datasets, as well as in each of the cohorts independently, except EDSD ($p = 0.11$), with Parieto-temporal atrophy subtype consisting of the youngest AD-D patients (61.2 ± 8.1) and subcortical atrophy subtype the oldest (68.3 ± 8.6). In ADNI, AIBL, ARWiBo, I-ADNI and OASIS cohorts, MMSE of the AD-D patients in different subtypes were not significantly different ($p > 0.05$), indicating the identified subtypes and severity were disentangled. In ADC, EDSD and NACC cohorts, MMSE of AD-D patients was significantly different ($p < 0.05$) between subtypes, with the Parieto-temporal atrophy subtype having the lowest MMSE among the four subtypes. Percentage of *APOE4* carriers was significantly different ($p < 0.05$) in the AD-D dementia patients in the pooled validation datasets. The percentage of outliers across all Aβ+ validation datasets decreased with clinical stage (CU: 25.0%, MCI: 12.4%, AD-D: 9.1%) indicating that characteristic atrophy patterns emerge as the disease progresses.

Supplementary Figure 4 depicts the atrophy subtypes and sequence of atrophy-events estimated by SuStaIn and Supplementary Figure 5 shows the posterior probability distribution

of these sequences using Markov chain Monte Carlo (MCMC) sampling, interpreted as the uncertainty in this estimation. The identified subtypes were Typical subtype (with early hippocampus and temporal lobe atrophy), Limbic predominant subtype, Hippocampal sparing subtype, and Subcortical subtype. The prevalence of these subtypes and age distribution among AD-D participants in the training and Aβ+ validation dataset were: Typical ($Training$: 55.7% ($n = 332/596$), $age = 66.7 \pm 7.8$; Aβ + $Validation$: 56.0% ($n = 501/894$), $age = 65.8 \pm 9.3$), Limbic predominant ($Training$: 24.1% ($n = 144/596$), $age = 72.2 \pm 6.6$; Aβ + $Validation$: 24.0%(215/894), $age = 69.8 \pm 8.2$), Hippocampal sparing ($Training$: 14.5% ($n = 87/596$), $age = 62.8 \pm 6.9$; Aβ + $Validation$: 12.9% ($n = 115/894$) $age = 60.9 \pm 7.0$), Subcortical atrophy ($Training$: 0.8% ($n = 5/596$), $age = 68.2 \pm 7.6$; Aβ + $Validation$: 0.5% ($n = 5/894$), $age = 70.4 \pm 9.3$). Apart from these subtypes, an outlier group (defined as AD-D patients in stage 0) was detected (Training: 4.7%; Aβ + $Validation$: 6.5%) The prevalence, age and MMSE distribution, and the percentage of *APOE4* carriers in each of these subtypes across the different diagnostic categories in the pooled validation datasets have been summarized in Table 2 and these results in each cohort independently are reported in Supplementary Table 5. Age of onset of AD dementia and *APOE4* carriership percentage differed significantly ($p < 0.05$) between the four subtypes identified by SuStaIn with Hippocampal sparing subtype consisting of the youngest AD-D patients (61.0 ± 7.1) and Limbic-predominant and Subcortical atrophy subtypes the oldest (69.9 ± 8.2 and 70.4 ± 9.3 respectively).

Experiment 1: Atrophy-based model stage correlates with MMSE

Figure 3 depicts the correlation between the atrophy-based patient stage assigned by Snowphlake for the clinical-validation dataset and external dataset, with MMSE, a clinical screening tool for measuring disease severity of the patient. The atrophy-based stage showed significant correlation with MMSE within all four subtypes, with higher atrophy stage related to worse MMSE scores ($R = -0.51\ to\ -0.28$), with $p < 0.0001$ in the clinical-validation dataset and $p < 0.05$ in the external-validation dataset. The distribution of atrophy-based stages for the different diagnostic groups (of CU, MCI, AD-D) were different ($p < 0.0001$) and are also shown in Figure 3. Supplementary Figure 6 depicts a similar plot for these correlations for the Aβ+ validation dataset and the Aβ+ subset of the external dataset. Supplementary Figure 7 shows the correlation between the atrophy-based patient stage assigned by SuStaIn for the clinical-validation dataset and external dataset, with MMSE. The

atrophy-based stage assigned by SuStaIn showed significant correlation with MMSE within all subtypes except in the subcortical atrophy subtype ($R = -0.54\ to\ -0.26$) with $p < 0.0001$ in the clinical-validation dataset and $p < 0.01$ in the external-validation dataset and $p > 0.05$ for the subcortical atrophy subtype in both validation datasets.

Experiment 2: Cognitive domain characteristics of the subtypes

Figure 4 shows the effect sizes (Cohen's f-statistic) of cognitive domain score differences between subtypes identified by Snowphlake and SuStaIn, for the diagnostic groups of MCI and AD-D. These subtype differences are computed for participants in the Aβ+ validation cohorts of ADC, ADNI, NACC, AIBL, and GMC for the cognitive domains of memory, executive function and attention, language, and visuospatial function. For Snowphlake, the mean effect sizes for the four cognitive domains were between $f = 0.15\ to\ 0.33$ in the AD-D group, and were between $f = 0.15\ to\ 0.24$ in the MCI group. For SuStaIn, the mean effect sizes for the four cognitive domains were between $f = 0.17\ to\ 0.34$ in the AD-D group and were between $f = 0.08\ to\ 0.20$ in the MCI group. There were no significant differences between the effect sizes of Snowphlake and SuStaIn for AD-D patients, while Snowphlake was significantly better at detecting differences in the language domain for MCI patients (FDR-corrected $p = 0.016$) than SuStaIn's subtypes. There was significant heterogeneity (based on Cochran's Q statistic) observed between cohorts for both the methods, for both the diagnostic groups. Supplementary Figure 8, depicts the subtype differences in the distribution of cognitive domain scores independently in each cohort for Snowphlake subtypes, which further highlights the differences in associations across different cohorts. Supplementary Figure 9 shows similar associations between the subtypes and cognitive domain scores in different cohorts for the subtypes identified by SuStaIn.

Experiment 3: Concordant subtype-pairs

When comparing how participants were clustered, we observed a low concordance between Snowphlake and SuStaIn. Figure 5 shows the contingency matrices between Snowphlake and SuStaIn subtype assignments in different clinical stages of Aβ+ participants, in the training and validation datasets.

Of the $n = 501$ AD-D individuals assigned to the typical subtype (with prominent hippocampal and temporal lobe atrophy) of SuStaIn in the Aβ+ validation dataset, $n = 183$ (36.5%) were also assigned to the frontal atrophy subtype (with prominent frontal and temporal lobe atrophy) of Snowphlake, which is referred to as concordant subtype-pair #1. Of

the $n = 215$ AD-D individuals assigned to the limbic-predominant subtype (with prominent thalamus, hippocampus, and amygdala atrophy) of SuStaIn in the Aβ+ validation dataset, $n = 127$ (59.1%) were also assigned to the subcortical-atrophy subtype of Snowphlake, which is referred to as concordant subtype-pair #2. Of the $n = 115$ AD-D individuals assigned to the hippocampal-sparing subtype of SuStaIn in the Aβ+ validation dataset, $n = 52$ (45.2%) were also assigned to the parieto-temporal atrophy subtype of Snowphlake, which is referred to as concordant subtype-pair #3. The Subcortical atrophy subtype of SuStaIn was too small to be compared. The concordant subtype-pairs accounted only for 38.6% $(n = 230/596)$ of Aβ+ AD-D participants in the training dataset and 40.5% $(n = 362/894)$ in the Aβ+ validation dataset. Cohort-wise contingency matrix shown in Supplementary Figure 10 further added to the evidence that low concordance was consistent across cohorts.

Average w-score maps of the concordant subtype-pairs #1 and #2 and their less frequently co-occurring counterparts showed significant differences $(p_{FDR} < 0.05)$ within the Typical and Limbic-predominant subtypes of SuStaIn in 14 and 18 of the 24 regions respectively, adding further evidence for the spectrum of differences in atrophy within SuStaIn subtypes. Similar analysis for the concordant subtype-pairs #3 showed significant differences $(p_{FDR} < 0.05)$ within the Hippocampal-sparing atrophy subtype of SuStaIn in 5 of the 24 regions. Supplementary Figure 11 visualizes these regional differences.

Lastly, progression modelling in the three concordant subtype-pairs using DEBM and w-score EBM showed that the estimated atrophy-event sequences using the two methods were largely similar. The normalized Kendall's Tau (KT) metric measuring the dissimilarity between the sequences estimated by SuStaIn and Snowphlake were: $KT = 0.11$ for concordant subtype-pair #1, $KT = 0.14$ for concordant subtype-pair #2, and $KT = 0.12$ for concordant subtype-pair #3. These values are within the expected error ranges of each model,[15] indicating that the estimated sequences in concordant subtype-pairs using the two methods agree with each other. The sequences of atrophy-events estimated using DEBM and z-score EBM are shown in Figure 6.

Discussion

In this large-scale multi-cohort study of atrophy-heterogeneity in AD, we used a novel methodology, Snowphlake, that couples a previously-validated ML approach for disease subtyping (NMF)[4,14] with data-driven disease progression modelling (DEBM), to estimate sequences of atrophy-events in four atrophy-based subtypes of AD. We compared our results with those obtained using SuStaIn and used the trained models to assign subtypes and atrophy stage in patient populations not included in training them. The assigned subtypes in validation datasets were associated with distinct cognitive profiles and the atrophy stage with the subtypes correlated with global cognition level of patients. We have made the trained models of both SuStaIn and Snowphlake openly available at https://snowphlake-dpm.github.io, along with the associated code. The source code for Snowphlake has also been made available at: https://github.com/snowphlake-dpm/snowphlake, while the source code for SuStaIn was previously made available by Aksman et al.[33] A thorough comparison of Snowphlake's subtype assignments with that of SuStaIn's provided evidence for a spectrum of differences in atrophy among AD patients, rather than discretised by distinct subtypes.

The identified atrophy-based subtypes were consistent with literature

Snowphlake identified a parieto-temporal atrophy subtype where the AD-D patients were consistently the youngest and had worse visuospatial function, attention and executive function consistent with prior studies on young-onset AD patients.[4,6,36] This subtype also had a significantly lower percentage of *APOE*4 carriers in the ADC cohort, also observed in a previous study[36], as well as in the and the ARWiBo cohort. Still, *APOE*4 carriership did not differ significantly in other cohorts in our study, which may be because those cohorts predominantly consisted of late-onset AD patients. The subcortical atrophy subtype (also referred to as "mild atrophy" in literature) patients had the least affected cognition across all domains when compared to the other subtypes.[4,9,11] The diffuse cortical atrophy subtype (or cortical atrophy subtype) and frontal atrophy subtype have also been identified in previous studies[10,37,38]. Moreover, the subtypes identified by SuStaIn in this study (typical, hippocampal-sparing, limbic-predominant) were aligned with the neuropathological subtypes of AD reported in literature[3,39] and largely aligned with the previous studies of atrophy-subtypes using SuStaIn.[2,10,40]

Comparing Snowphlake and SuStaIn subtypes

A novel approach in our study was that we compared two data-driven AD subtyping techniques directly on the same patient population, while the comparisons in the previous studies so far have been based on the identified atrophy characteristics or patient characteristics.[2,4,9] The subtypes identified by the two methods in our analysis also showed some similarities in patient characteristics, for example, parieto-temporal atrophy subtype of Snowphlake and hippocampal-sparing subtype of SuStaIn both consisted of significantly younger-onset AD-D patients. Nevertheless, our direct comparison showed low concordance between the subtype assignments of the two methods, highlighting the limitations of indirect comparisons based on read-outs.

While comparing average w-score maps of patients within a specific SuStaIn subtype, but assigned to different Snowphlake subtype, we saw significant differences in atrophy profiles, providing further evidence that atrophy patterns might vary substantially between individuals within a data-driven subtype. The three concordant subtype-pairs that accounted for approximately 40% of individuals with AD-D were the typical subtype with temporal and frontal lobe atrophy, the limbic predominant subtype with severe subcortical atrophy, and the hippocampal sparing subtype with parieto-temporal atrophy. The sequence of atrophy-events estimated by the two methods in these concordant subtype-pairs agreed with each other, showing that in spite of the methodological differences, similar inferences could be made in these concordantly subtyped individuals. Although these concordant subtype-pairs are in line with previous literature[3,5], future work on synthetic data simulating a spectrum of atrophy differences would be crucial for understanding more about concordant subtype-pairs. However, the notion that not all patients were clustered similarly, suggests that group estimates of atrophy subtypes may be driven by a particular subset of patients, and may not capture heterogeneity of all patients. Future studies should further investigate more continuous measures of subtyping that may be able to better capture such nuance and heterogeneity.

The differences in estimated subtypes by the two methods arise from the differences in the objective-functions being optimized by the methods. While SuStaIn optimizes a non-linear objective-function to jointly estimate subtypes and atrophy-stage, Snowphlake uses linear objective-function in NMF to identify subtypes. Each of them have been shown before to identify true subtypes in the presence of distinct subtypes.[2,41] The low concordance between the atrophy-subtype assignments of the two methods can hence be seen as evidence of a

spectrum of atrophy differences between individuals with AD. This spectrum could either consist of distinct prototypical subtypes coupled with a lot of variations in a large number of AD patients, or it could be a continuum of atrophy-variations with the Snowphlake and SuStaIn identifying different variations depending on the objective-function used for their optimization. While the non-linear objective-function of SuStaIn identifies non-uniform distribution of the identified subtypes, Snowphlake's linear objective-function identifies four subtypes that were roughly uniformly distributed. In the absence of ground-truth in data-driven subtyping, the ability of the identified subtypes to associate with distinct cognitive profiles determines their validity.

Differences in cognitive domain profiles

The subtypes identified by Snowphlake and SuStaIn each showed significant differences in cognitive domain scores in both Aβ+ MCI and AD-D patients. While the effect sizes were comparable for Snowphlake and SuStaIn for AD-D patients, Snowphlake showed marginally stronger effect sizes for MCI patients, potentially indicating that Snowphlake's subtypes are more sensitive at associating with different symptom profiles at the prodromal stage of the disease. While some of the differences between subtypes (by either method) assigned were consistent across the multiple cohorts in our study, we also observed significant heterogeneity in associations across cohorts. These differences could potentially indicate genuine cohort-wise differences in how atrophy causes symptoms or could be due to using different cognitive tests to compute cognitive domain scores in different cohorts. Future work on studying these associations could focus on working with harmonized cognitive data across multiple cohorts.[42,43] Notwithstanding these inconsistencies, the significant differences in cognitive domain profiles between subtypes indicate that data-driven subtyping models have the potential to identify personalized end-points in future interventions to boost statistical power.[44,45]

Methodological considerations and limitations

A potential limitation of our approach is that while our algorithms allow estimation of sequences of atrophy events, these remain inferences based on cross-sectional data. While there have been prior studies that validated these inferences on longitudinal datasets[46,47], future studies could focus on a similar large-scale validation on multi-cohort longitudinal datasets to confirm if these subtypes remain consistent in preclinical and prodromal AD patients as the disease develops. One of the strengths of our study is that we have made the trained models and source code openly available and validated the subtype assignments in

external datasets. Future studies can hence use these trained models to identify proteomic profiles, genetic and lifestyle factors driving these subtypes in large external cohorts. Another important feature of this study is that our combined multi-cohort data had many patients with young-onset AD-D. This could potentially be a strength of our study since young-onset AD-D patients have less comorbidity or it could be a limitation with the identified subtypes being an over-representation of young-onset AD-D patients. Lastly, by decoupling atrophy-based subtyping from disease progression modelling in the Snowphlake framework, we pave the way for the inclusion of high-dimensional imaging features (such as voxel-based measures) in data-driven subtyping and staging analysis.

Conclusion

In conclusion, in this large-scale multi-centre study, we identified four atrophy-based subtypes using Snowphlake and SuStaIn. Subtype assignments in independent validation datasets were associated with different cognitive symptoms, and estimated atrophy-severity measures were associated with global cognition. The low concordance of subtypes between the two methods indicates that atrophy differences between individuals may be a spectrum rather than strictly delineated subtypes. Based on our findings, future research should prioritize developing novel approaches to capture and analyse this spectrum of heterogeneity in atrophy patterns to help us further understand the biological-basis for the observed variability in atrophy patterns between individuals.

Data availability

The ADNI data used is this study were obtained from the ADNI database (adni.loni.usc.edu). The ADC data used in this study are available from the corresponding author, upon reasonable request. The AIBL imaging data used in this study were obtained from the AIBL LONI database (https://ida.loni.usc.edu/login.jsp?project=AIBL), while cognitive and genetic data can be requested from the AIBL management team, upon reasonable request by submitting an Expression of Interest (EOI) form available on the AIBL website (https://aibl.org.au/collaboration/). The NACC data used in this study were obtained from https://naccdata.org/. The OASIS data used in this study were obtained from https://sites.wustl.edu/oasisbrains/ website. The data of the other cohorts used in this study can be requested from the neuGRID (https://www.neugrid2.eu/) and GAAIN (https://www.gaain.org) platforms after registration.

Funding

This study was supported by the Early Detection of Alzheimer's Disease Subtypes (E-DADS) project, an EU Joint Programme - Neurodegenerative Disease Research (JPND) project (see www.jpnd.eu). The project is supported under the aegis of JPND through the following funding organizations: United Kingdom, Medical Research Council (MR/T046422/1); Netherlands, ZonMW (733051106); France, Agence Nationale de la Recherche (ANR-19-JPW2–000); Italy, Italian Ministry of Health (MoH); Australia, National Health & Medical Research Council (1191535); Hungary, National Research, Development and Innovation Office (2019–2.1.7-ERA-NET-2020–00008).

This work used the Dutch national e-infrastructure with the support of the SURF Cooperative using grant no. EINF-5353. W.F. and B.T. are recipients of TAP-dementia (www.tap-dementia.nl), receiving funding from ZonMw (#10510032120003). F.B. is supported by the NIHR biomedical research centre at UCLH. B.W. and Z.V. was supported by project no. RRF-2.3.1-21-2022-00015, which has been implemented with the support provided by the European Union. B.W. was also supported by project no. RRF-2.3.1-21-2022-00004 that has been implemented with the support of the European Union within the framework of the Artificial Intelligence National Laboratory.

N.P.O. is supported by a UKRI Future Leaders Fellowship (UK Medical Research Council MR/S03546X/1). Research of C.E.T. is supported by the European Commission (Marie Curie International Training Network, grant agreement No 860197 (MIRIADE), Innovative Medicines Initiatives 3TR (Horizon 2020, grant no 831434) EPND (IMI 2 Joint Undertaking (JU), grant No. 101034344) and JPND (bPRIDE), European Partnership on Metrology, co-financed from the European Union's Horizon Europe Research and Innovation Programme and by the Participating States ((22HLT07 NEuroBioStand), CANTATE project funded by the Alzheimer Drug Discovery Foundation, Alzheimer Association, Health Holland, the Dutch Research Council (ZonMW), Alzheimer Drug Discovery Foundation, The Selfridges Group Foundation, Alzheimer Netherlands. CT is recipient of ABOARD, which is a public-private partnership receiving funding from ZonMW (#73305095007) and Health~Holland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106). CT is recipient of TAP-dementia, a ZonMw funded project (#10510032120003) in the context of the Dutch National Dementia Strategy.

Acknowledgments

The authors would like to thank all the research participants and their families for donating their data for scientific research.

Data collection and sharing for ADNI 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 (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.

The AIBL study (https://aibl.org.au) is a consortium between Austin Health, CSIRO, Edith Cowan University, the Florey Institute (The University of Melbourne), and the National Ageing Research Institute. The study has received partial financial support from the Alzheimer's Association (US), the Alzheimer's Drug Discovery Foundation, an Anonymous foundation, the Science and Industry Endowment Fund, the Dementia Collaborative Research Centres, the Victorian Government's Operational Infrastructure Support program, the Australian Alzheimer's Research Foundation, the National Health and Medical Research Council (NHMRC), and The Yulgilbar Foundation. Numerous commercial interactions have supported data collection and analyses. In-kind support has also been provided by Sir Charles Gairdner Hospital, Cogstate Ltd, Hollywood Private Hospital, The University of Melbourne, and St Vincent's Hospital. The AIBL team wishes to thank all clinicians who referred patients with AD and/or MCI to the study. We also thank all those who took part as subjects in the study for their commitment and dedication to helping advance research into the early detection and causation of AD. We thank all the investigators within the AIBL who contributed to the design and implementation of the resource and/or provided data but did not actively participate in the development, analysis, interpretation or writing of this current study.

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30

AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici, MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

Competing interests

F.B. is on the steering committee or Data Safety Monitoring Board member for Biogen, Merck, ATRI/ACTC and Prothena. F.B. has been a consultant for Roche, Celltrion, Rewind Therapeutics, Merck, IXICO, Jansen, Combinostics and has research agreements with Merck, Biogen, GE Healthcare, Roche. F.B and D.C.A. are also co-founders and shareholders of Queen Square Analytics Ltd. N.P.O. is a consultant for Queen Square Analytics Ltd.

Research programs of W.F. have been funded by ZonMW, NWO, EU-JPND, EU-IHI, Alzheimer Nederland, Hersenstichting CardioVascular Onderzoek Nederland, Health~Holland, Topsector Life Sciences & Health, stichting Dioraphte, Gieskes-Strijbis fonds, stichting Equilibrio, Edwin Bouw fonds, Pasman stichting, stichting Alzheimer & Neuropsychiatrie Foundation, Philips, Biogen MA Inc, Novartis-NL, Life-MI, AVID, Roche BV, Fujifilm, Eisai, Combinostics. W.F. holds the Pasman chair. W.F. is recipient of ABOARD, which is a public-private partnership receiving funding from ZonMW (#73305095007) and Health~Holland, Topsector Life Sciences & Health (PPP-allowance; #LSHM20106). W.F. is recipient of TAP-dementia (www.tap-dementia.nl), receiving funding from ZonMw (#10510032120003). TAP-dementia receives co-financing from Avid Radiopharmaceuticals and Amprion. All funding is paid to her institution.

W.F. has been an invited speaker at Biogen MA Inc, Danone, Eisai, WebMD Neurology (Medscape), NovoNordisk, Springer Healthcare, European Brain Council. All funding is paid to her institution. W.F. is consultant to Oxford Health Policy Forum CIC, Roche, Biogen MA Inc, and Eisai. All funding is paid to her institution. W.F. participated in advisory boards of

Biogen MA Inc, Roche, and Eli Lilly. W.F. is member of the steering committee of EVOKE/EVOKE+ (NovoNordisk). All funding is paid to her institution. W.F. is member of the steering committee of PAVE, and Think Brain Health. W.F. was associate editor of Alzheimer, Research & Therapy in 2020/2021. W.F. is associate editor at Brain.

C.E.T. has research contracts with Acumen, ADx Neurosciences, AC-Immune, Alamar, Aribio, Axon Neurosciences, Beckman-Coulter, BioConnect, Bioorchestra, Brainstorm Therapeutics, Celgene, Cognition Therapeutics, EIP Pharma, Eisai, Eli Lilly, Fujirebio, Instant Nano Biosensors, Novo Nordisk, Olink, PeopleBio, Quanterix, Roche, Toyama, Vivoryon. She is editor in chief of Alzheimer Research and Therapy, and serves on editorial boards of Molecular Neurodegeneration, Neurology: Neuroimmunology & Neuroinflammation, Medidact Neurologie/Springer, and serves on committee to define guidelines for Cognitive disturbances, and one for acute Neurology in the Netherlands. She had consultancy/speaker contracts for Aribio, Biogen, Beckman-Coulter, Cognition Therapeutics, Eli Lilly, Merck, Novo Nordisk, Olink, Roche and Veravas.

Supplementary material

Supplementary material is available online.

References

1. 2023 Alzheimer's disease facts and figures. *Alzheimers Dement*. Apr 2023;19(4):1598-1695. doi:10.1002/alz.13016
2. Young AL, Marinescu RV, Oxtoby NP, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nat Commun*. Oct 15 2018;9(1):4273. doi:10.1038/s41467-018-05892-0
3. Ferreira D, Nordberg A, Westman E. Biological subtypes of Alzheimer disease: A systematic review and meta-analysis. *Neurology*. Mar 10 2020;94(10):436-448. doi:10.1212/WNL.0000000000009058
4. Ten Kate M, Dicks E, Visser PJ, et al. Atrophy subtypes in prodromal Alzheimer's disease are associated with cognitive decline. *Brain*. Dec 1 2018;141(12):3443-3456. doi:10.1093/brain/awy264
5. Zhang B, Lin L, Wu S. A Review of Brain Atrophy Subtypes Definition and Analysis for Alzheimer's Disease Heterogeneity Studies. *J Alzheimers Dis*. 2021;80(4):1339-1352. doi:10.3233/JAD-201274
6. Scheltens NME, Tijms BM, Koene T, et al. Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts. *Alzheimers Dement*. Nov 2017;13(11):1226-1236. doi:10.1016/j.jalz.2017.03.002
7. Geifman N, Kennedy RE, Schneider LS, Buchan I, Brinton RD. Data-driven identification of endophenotypes of Alzheimer's disease progression: implications for clinical trials and therapeutic interventions. *Alzheimers Res Ther*. Jan 15 2018;10(1):4. doi:10.1186/s13195-017-0332-0
8. Risacher SL, Anderson WH, Charil A, et al. Alzheimer disease brain atrophy subtypes are associated with cognition and rate of decline. *Neurology*. Nov 21 2017;89(21):2176-2186. doi:10.1212/WNL.0000000000004670
9. Zhang X, Mormino EC, Sun N, et al. Bayesian model reveals latent atrophy factors with dissociable cognitive trajectories in Alzheimer's disease. *Proc Natl Acad Sci U S A*. Oct 18 2016;113(42):E6535-E6544. doi:10.1073/pnas.1611073113
10. Chen H, Young A, Oxtoby NP, et al. Transferability of Alzheimer's disease progression subtypes to an independent population cohort. *Neuroimage*. May 1 2023;271:120005. doi:10.1016/j.neuroimage.2023.120005
11. Ferreira D, Verhagen C, Hernandez-Cabrera JA, et al. Distinct subtypes of Alzheimer's disease based on patterns of brain atrophy: longitudinal trajectories and clinical applications. *Sci Rep*. Apr 18 2017;7:46263. doi:10.1038/srep46263
12. Young AL, Oxtoby NP, Garbarino S, et al. Data-driven modelling of neurodegenerative disease progression: thinking outside the black box. *Nat Rev Neurosci*. Feb 2024;25(2):111-130. doi:10.1038/s41583-023-00779-6
13. Poulet P-E, Durrleman S. Mixture Modeling for Identifying Subtypes in Disease Course Mapping. Springer International Publishing; 2021:571-582.
14. Tijms BM, Vromen EM, Mjaavatten O, et al. Cerebrospinal fluid proteomics in patients with Alzheimer's disease reveals five molecular subtypes with distinct genetic risk profiles. *Nat Aging*. Jan 2024;4(1):33-47. doi:10.1038/s43587-023-00550-7
15. Venkatraghavan V, Bron EE, Niessen WJ, Klein S, Alzheimer's Disease Neuroimaging I. Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *Neuroimage*. Feb 1 2019;186:518-532. doi:10.1016/j.neuroimage.2018.11.024

16. Venkatraghavan V, Klein S, Fani L, et al. Analyzing the effect of APOE on Alzheimer's disease progression using an event-based model for stratified populations. *Neuroimage*. Feb 15 2021;227:117646. doi:10.1016/j.neuroimage.2020.117646
17. van der Flier WM, Pijnenburg YA, Prins N, et al. Optimizing patient care and research: the Amsterdam Dementia Cohort. *J Alzheimers Dis*. 2014;41(1):313-27. doi:10.3233/JAD-132306
18. Jack CR, Jr., Bernstein MA, Fox NC, et al. The Alzheimer's Disease Neuroimaging Initiative (ADNI): MRI methods. *J Magn Reson Imaging*. Apr 2008;27(4):685-91. doi:10.1002/jmri.21049
19. Ellis KA, Bush AI, Darby D, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr*. Aug 2009;21(4):672-87. doi:10.1017/S1041610209009405
20. Beekly DL, Ramos EM, Lee WW, et al. The National Alzheimer's Coordinating Center (NACC) database: the Uniform Data Set. *Alzheimer Dis Assoc Disord*. Jul-Sep 2007;21(3):249-58. doi:10.1097/WAD.0b013e318142774e
21. Marcus DS, Wang TH, Parker J, Csernansky JG, Morris JC, Buckner RL. Open Access Series of Imaging Studies (OASIS): cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *J Cogn Neurosci*. Sep 2007;19(9):1498-507. doi:10.1162/jocn.2007.19.9.1498
22. Frisoni GB, Prestia A, Zanetti O, et al. Markers of Alzheimer's disease in a population attending a memory clinic. *Alzheimers Dement*. Jul 2009;5(4):307-17. doi:10.1016/j.jalz.2009.04.1235
23. Brueggen K, Grothe MJ, Dyrba M, et al. The European DTI Study on Dementia - A multicenter DTI and MRI study on Alzheimer's disease and Mild Cognitive Impairment. *Neuroimage*. Jan 2017;144(Pt B):305-308. doi:10.1016/j.neuroimage.2016.03.067
24. Cavedo E, Redolfi A, Angeloni F, et al. The Italian Alzheimer's Disease Neuroimaging Initiative (I-ADNI): validation of structural MR imaging. *J Alzheimers Dis*. 2014;40(4):941-52. doi:10.3233/JAD-132666
25. Galluzzi S, Marizzoni M, Babiloni C, et al. Clinical and biomarker profiling of prodromal Alzheimer's disease in workpackage 5 of the Innovative Medicines Initiative PharmaCog project: a 'European ADNI study'. *J Intern Med*. Jun 2016;279(6):576-91. doi:10.1111/joim.12482
26. Ribaldi F, Chicherio C, Altomare D, et al. Brain connectivity and metacognition in persons with subjective cognitive decline (COSCODE): rationale and study design. *Alzheimers Res Ther*. May 25 2021;13(1):105. doi:10.1186/s13195-021-00846-z
27. Monereo-Sanchez J, de Jong JJA, Drenthen GS, et al. Quality control strategies for brain MRI segmentation and parcellation: Practical approaches and recommendations - insights from the Maastricht study. *Neuroimage*. Aug 15 2021;237:118174. doi:10.1016/j.neuroimage.2021.118174
28. Archetti D, Venkatraghavan V, Weiss B, et al. A machine-learning model to harmonize brain volumetric data for quantitative neuro-radiological assessment of Alzheimer’s disease. *medRxiv*. 2024:2024.02.01.24302048. doi:10.1101/2024.02.01.24302048
29. Fortin JP, Cullen N, Sheline YI, et al. Harmonization of cortical thickness measurements across scanners and sites. *Neuroimage*. Feb 15 2018;167:104-120. doi:10.1016/j.neuroimage.2017.11.024
30. Young AL, Vogel JW, Robinson JL, et al. Data-driven neuropathological staging and subtyping of TDP-43 proteinopathies. *Brain*. Jul 3 2023;146(7):2975-2988. doi:10.1093/brain/awad145

31. Salvado G, Molinuevo JL, Brugulat-Serrat A, et al. Centiloid cut-off values for optimal agreement between PET and CSF core AD biomarkers. *Alzheimers Res Ther*. Mar 21 2019;11(1):27. doi:10.1186/s13195-019-0478-z
32. Hyman BT, Phelps CH, Beach TG, et al. National Institute on Aging-Alzheimer's Association guidelines for the neuropathologic assessment of Alzheimer's disease. *Alzheimers Dement*. Jan 2012;8(1):1-13. doi:10.1016/j.jalz.2011.10.007
33. Aksman LM, Wijeratne PA, Oxtoby NP, et al. pySuStaIn: a Python implementation of the Subtype and Stage Inference algorithm. *SoftwareX*. Dec 2021;16doi:10.1016/j.softx.2021.100811
34. Pascual-Montano A, Carazo JM, Kochi K, Lehmann D, Pascual-Marqui RD. Nonsmooth nonnegative matrix factorization (nsNMF). *IEEE Trans Pattern Anal Mach Intell*. Mar 2006;28(3):403-15. doi:10.1109/TPAMI.2006.60
35. Rousseeuw PJ, Van Driessen K. A fast algorithm for the minimum covariance determinant estimator. *Technometrics*. Aug 1999;41(3):212-223. doi:Doi 10.2307/1270566
36. van der Flier WM, Pijnenburg YA, Fox NC, Scheltens P. Early-onset versus late-onset Alzheimer's disease: the case of the missing APOE varepsilon4 allele. *Lancet Neurol*. Mar 2011;10(3):280-8. doi:10.1016/S1474-4422(10)70306-9
37. Alladi S, Xuereb J, Bak T, et al. Focal cortical presentations of Alzheimer's disease. *Brain*. Oct 2007;130(Pt 10):2636-45. doi:10.1093/brain/awm213
38. Sawyer RP, Rodriguez-Porcel F, Hagen M, Shatz R, Espay AJ. Diagnosing the frontal variant of Alzheimer's disease: a clinician's yellow brick road. *J Clin Mov Disord*. 2017;4:2. doi:10.1186/s40734-017-0052-4
39. Murray ME, Graff-Radford NR, Ross OA, Petersen RC, Duara R, Dickson DW. Neuropathologically defined subtypes of Alzheimer's disease with distinct clinical characteristics: a retrospective study. *Lancet Neurol*. Sep 2011;10(9):785-96. doi:10.1016/S1474-4422(11)70156-9
40. Baumeister H, Vogel JW, Insel PS, et al. A generalizable data-driven model of atrophy heterogeneity and progression in memory clinic settings. *Brain*. Jul 5 2024;147(7):2400-2413. doi:10.1093/brain/awae118
41. Lee DD, Seung HS. Learning the parts of objects by non-negative matrix factorization. *Nature*. Oct 21 1999;401(6755):788-791. doi:Doi 10.1038/44565
42. Gavett BE, Ilango SD, Koscik R, et al. Harmonization of cognitive screening tools for dementia across diverse samples: A simulation study. *Alzheimers Dement (Amst)*. Apr-Jun 2023;15(2):e12438. doi:10.1002/dad2.12438
43. Boccardi M, Monsch AU, Ferrari C, et al. Harmonizing neuropsychological assessment for mild neurocognitive disorders in Europe. *Alzheimers Dement*. Jan 2022;18(1):29-42. doi:10.1002/alz.12365
44. Evans S, McRae-McKee K, Wong MM, Hadjichrysanthou C, De Wolf F, Anderson R. The importance of endpoint selection: How effective does a drug need to be for success in a clinical trial of a possible Alzheimer's disease treatment? *Eur J Epidemiol*. Jul 2018;33(7):635-644. doi:10.1007/s10654-018-0381-0
45. Doherty T, Yao Z, Khleifat AAL, et al. Artificial intelligence for dementia drug discovery and trials optimization. *Alzheimers Dement*. Dec 2023;19(12):5922-5933. doi:10.1002/alz.13428
46. Wijeratne PA, Eshaghi A, Scotton WJ, et al. The temporal event-based model: Learning event timelines in progressive diseases. *Imaging Neuroscience*. 2023;1:1-19. doi:10.1162/imag_a_00010
47. Venkatraghavan V, Vinke EJ, Bron EE, et al. Progression along data-driven disease timelines is predictive of Alzheimer's disease in a population-based cohort. *Neuroimage*. Sep 2021;238:118233. doi:10.1016/j.neuroimage.2021.118233

Figures and Tables

Table 1: Participant Demographics. Values indicated in this table are calculated after automated quality control.

Cohort	Age [years]	Sex (F/M)	CN and SCD	MCI	AD-D
			$A\beta$ Status: $-$ / $+$ / unknown	$A\beta$ Status: $-$ / $+$ / unknown	$A\beta$ Status: $-$ / $+$ / unknown
ADC	63.9 ± 9.2	1,675 / 1,952	687 / 184 / 456	256 / 328 / 199	79 / 1053 / 385
ADNI	72.1 ± 7.0	875 / 909	378 / 184 / 113	254 / 397 / 177	21 / 192 / 68
AIBL	72.7 ± 6.5	298 / 224	268 / 91 / 28	27 / 49 / 7	6 / 43 / 3
ARWiBo	55.1 ± 16.0	482 / 293	1 / 0 / 593	0 / 14 / 89	4 / 10 / 64
EDSD	70.4 ± 7.3	191 / 174	0 / 0 / 136	24 / 45 / 43	0 / 1 / 116
I-ADNI	72.1 ± 8.0	105 / 64	0 / 0 / 7	0 / 0 / 35	0 / 0 / 127
NACC	71.2 ± 10.2	882 / 688	358 / 0 / 0	39 / 126 / 463	18 / 191 / 375
OASIS	71.9 ± 10.8	197 / 105	0 / 0 / 185	0 / 0 / 90	0 / 0 / 27
PharmaCog	69.0 ± 7.4	78 / 57	0 / 0 / 0	52 / 83 / 0	0 / 0 / 0
GMC	71.5 ± 10.5	443 / 319	39 / 16 / 151	54 / 108 / 257	3 / 35 / 99
Total	67.6 ± 10.9	5,226 / 4,785	1,731 / 475 / 1,669	706 / 1,150 / 1,360	131 / 1,525 / 1,264

Abbreviations: CN: cognitively normal; SCD: subjective cognitive decline; MCI: mild cognitive impairment; AD-D: Clinical diagnosis of AD Dementia.

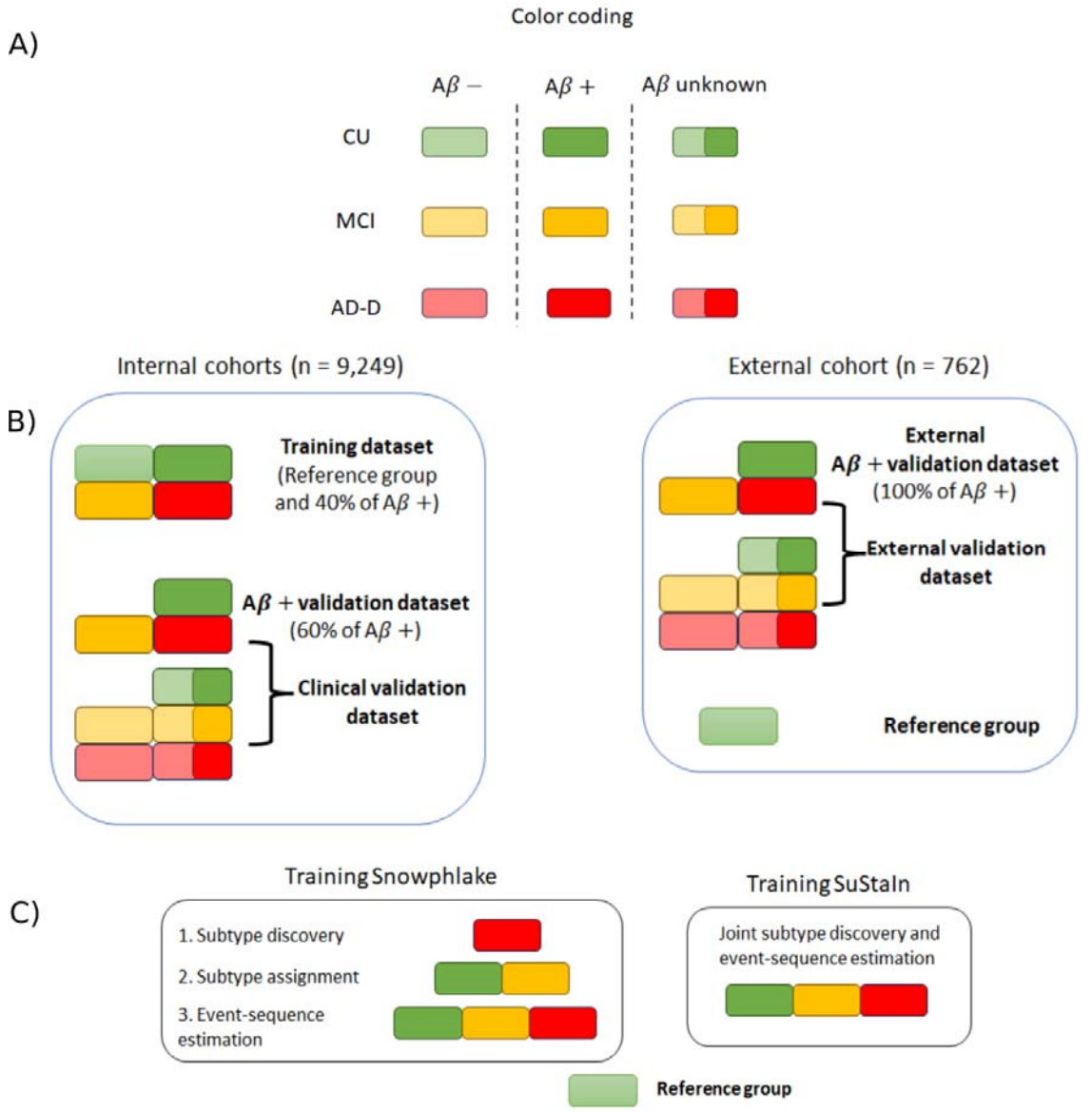

Figure 1: Graphical overview of this study. A) Shows the color coding used in the graphical over to denote participants in different clinical stages of the disease as well as their Aβ status. **B)** Overview of the data partitioning into the training dataset, Aβ+ validation dataset, clinical-validation dataset, and external validation datasets, including the inclusion criteria for participants in each dataset. **C)** Overview of the steps involved in training the Snowphlake and SuStaIn models. The reference group shown here is used in both the methods for creating a reference distribution and for w-scoring the imaging biomarkers. Abbreviations: CU: Cognitively unimpaired consisting of both cognitively normal (CN) individuals and subjective cognitive decliners (SCD); MCI: mild cognitively impaired; AD-D: individuals with clinical diagnosis of AD Dementia; + denotes Aβ positivity; - denotes Aβ negativity

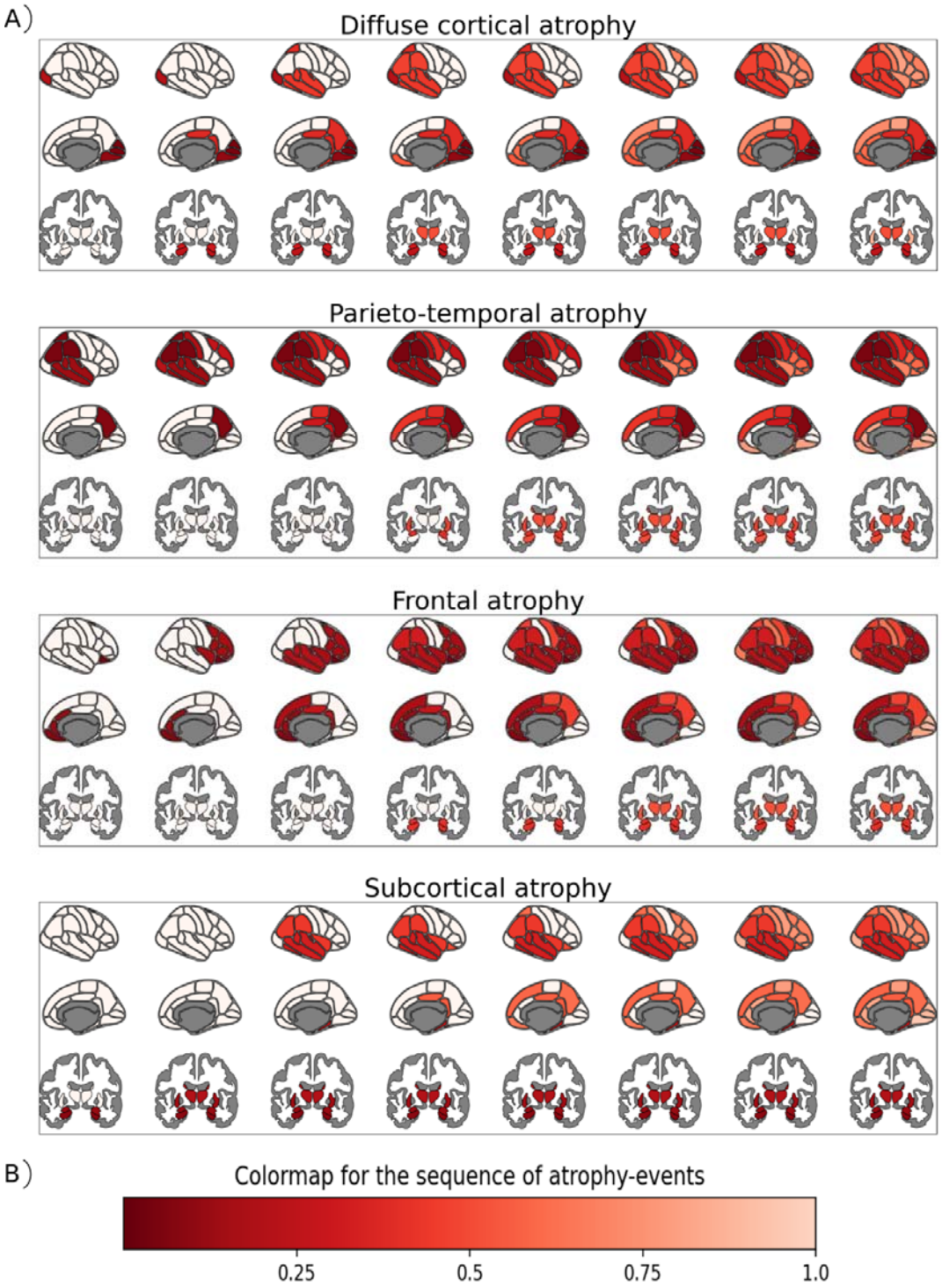

Figure 2: Snowphlake modelling in the Aβ+ participants in the multi-cohort harmonized training dataset. A) These plots depict the subtypes and sequence of atrophy-events for each subtype estimated. Within each subtype, the x-axis corresponds to the stage of the disease. Each column shows the brain in its lateral, medial, and subcortical views, with the regions that is expected to be abnormal at this stage for the subtype in shades of red and unaffected regions in white. **B)** The scale for the colour map goes from 0 to 1, the normalized staging scale for Snowphlake, where 0 represents a region becoming abnormal at the earliest stages of the disease and 1 represents late stage.

Table 2: **Characteristics of atrophy-based subtypes assigned by the trained Snowphlake and SuStaIn models, pooled across the validation datasets** (held-out validation dataset and external dataset).

Method: Snowphlake											
		Diffuse cortical atrophy		Parieto-temporal atrophy		Frontal atrophy		Subcortical atrophy		Outliers	
Characteristic per diagnostic group		Aβ+	All	Aβ+	All	Aβ+	All	Aβ+	All	Aβ+	All
Age	CU[#]	71.9 ± 9.2	60.4 ± 15.1	73.6 ± 3.8	62.7 ± 13.8	70.2 ± 10.0	66.0 ± 11.4	69.9 ± 6.0	62.0 ± 14.6	69.6 ± 9.0	61.0 ± 14.2
	MCI	71.1 ± 7.9	71.4 ± 8.8	68.8 ± 8.7	69.3 ± 9.4	71.0 ± 6.9	71.0 ± 8.7	71.3 ± 7.5	71.4 ± 8.8	71.3 ± 8.9	69.7 ± 9.6
	AD-D*[#]	67.7 ± 9.3	71.1 ± 9.3	61.2 ± 8.1	62.9 ± 8.8	67.6 ± 8.8	70.8 ± 8.9	68.3 ± 8.6	71.4 ± 8.9	66.5 ± 10.2	70.0 ± 10.6
N (%)	CU	90 (30.8)	533 (27.2)	5 (1.7)	49 (2.5)	53 (18.2)	293 (14.9)	73 (25.0)	421 (21.5)	71 (24.3)	665 (33.9)
	MCI	228 (31.1)	781 (27.9)	34 (4.6)	117 (4.2)	130 (17.7)	575 (20.5)	246 (33.5)	816 (29.1)	96 (13.1)	511 (18.3)
	AD-D	202 (21.7)	520 (22.4)	182 (19.6)	328 (14.1)	228 (24.5)	655 (28.1)	234 (25.2)	556 (23.9)	83 (8.9)	265 (11.4)
Sex (n_{male}/ n_{female})	CU	46/44	259/274	2/3	25/24	32/21	159/134	41/32	207/214	14/57	198/467
	MCI[#]	111/117	384/397	21/13	75/42	81/49	346/229	134/112	465/351	39/57	187/324
	AD-D*	76/126	217/303	88/94	145/183	113/115	298/357	121/113	276/280	28/55	96/169
MMSE	CU	28.5 ± 1.5	28.6 ± 1.5	28.4 ± 1.1	28.6 ± 1.2	28.7 ± 1.4	28.7 ± 1.5	28.5 ± 1.4	28.6 ± 1.5	28.8 ± 1.2	28.6 ± 1.6
	MCI[#]	26.7 ± 2.5	26.6 ± 2.7	25.2 ± 4.1	26.4 ± 3.2	26.6 ± 2.2	26.5 ± 2.7	26.3 ± 2.5	26.6 ± 2.6	27.1 ± 2.5	26.8 ± 2.8
	AD-D*[#]	21.8 ± 4.5	21.4 ± 4.7	19.1 ± 5.4	18.7 ± 5.7	21.0 ± 4.9	20.8 ± 5.3	22.1 ± 4.2	21.7 ± 4.5	20.3 ± 5.7	20.9 ± 5.4
APOE4 carriers (n_{APOE}/ n_{total})	CU	38/69	103/334	3/5	9/34	20/41	56/192	37/65	69/269	27/47	86/302
	MCI	125/198	265/594	21/28	43/88	82/120	200/451	131/193	283/598	49/74	122/342
	AD-D*[#]	137/185	240/388	103/173	140/275	135/216	256/529	159/217	267/428	54/80	93/194
Method: SuStaIn											
		Typical		Limbic-predominant		Hippocampal-sparing		Subcortical atrophy		Outliers	
Characteristic per diagnostic group		Aβ+	All	Aβ+	All	Aβ+	All	Aβ+	All	Aβ+	All
Age	CU[#]	71.5 ± 8.9	65.4 ± 13.4	67.6 ± 8.8	59.0 ± 16.5	75.8 ± 13.6	62.2 ± 15.8	76.0	68.2 ± 11.1	70.3 ± 8.2	61.0 ± 14.1
	MCI[#]	71.5 ± 7.5	72.1 ± 8.5	71.5 ± 8.2	71.4 ± 8.9	67.7 ± 9.6	68.1 ± 9.7	75.2 ± 4.4	73.3 ± 8.7	70.5 ± 7.5	69.7 ± 9.3
	AD-D*[#]	66.0 ± 9.3	69.5 ± 9.7	69.9 ± 8.2	72.2 ± 8.3	61.0 ± 7.1	63.0 ± 8.4	70.4 ± 9.3	71.8 ± 10.0	67.5 ± 10.6	72.3 ± 10.0
N (%)	CU	87 (29.8)	504 (25.7)	18 (6.2)	124 (6.3)	6 (2.1)	52 (2.7)	1 (0.3)	5 (0.3)	180 (61.6)	1276 (65.0)

	MCI	368 (50.1)	1255 (44.8)	150 (20.4)	484 (17.3)	27 (3.7)	91 (3.3)	5 (0.7)	26 (0.9)	184 (25.1)	944 (33.7)
	AD-D	525 (56.5)	1292 (55.6)	221 (23.8)	542 (23.2)	118 (12.7)	224 (9.6)	5 (0.5)	18 (0.8)	60 (6.5)	248 (10.7)
Sex (n_{male}/ n_{female})	CU[#]	50/37	252/252	13/5	84/40	4/2	27/25	1/0	4/1	67/113	471/805
	MCI	207/161	706/549	80/70	274/210	17/10	55/36	4/1	18/8	78/106	404/540
	AD-D	242/283	567/725	118/103	281/261	49/69	98/126	3/2	11/7	14/46	75/173
MMSE	CU	28.2 ± 1.6	28.5 ± 1.6	28.5 ± 1.4	28.5 ± 1.4	27.5 ± 1.4	28.5 ± 1.3	29.0	29.0 ± 0.7	28.8 ± 1.2	28.7 ± 1.5
	MCI*	26.6 ± 2.3	26.5 ± 2.7	26.0 ± 2.4	26.2 ± 2.8	24.3 ± 4.9	26.1 ± 3.6	26.4 ± 1.8	27.0 ± 2.9	26.9 ± 2.6	27.0 ± 2.6
	AD-D*[#]	20.8 ± 5.2	20.6 ± 5.2	21.5 ± 4.1	21.3 ± 4.8	19.3 ± 5.6	19.2 ± 5.5	22.0 ± 5.5	21.6 ± 4.4	23.5 ± 3.1	22.8 ± 4.3
APOE4 carriers (n_{APOE}/ n_{total})	CU	37/73	101/315	11/16	22/92	1/3	7/31	0/1	0/5	76/134	193/688
	MCI	212/310	431/916	78/117	172/360	14/25	28/68	3/5	7/18	102/158	275/711
	AD-D*[#]	309/495	516/1003	153/202	256/435	80/111	125/193	3/5	8/14	43/58	91/169

* indicates the corresponding measure is significantly different ($p < 0.05$) between the different subtypes (excluding the outliers group) in Aβ+ validation dataset, using ANOVA test for Age and MMSE characteristics, and χ^2 contingency test for Sex and *APOE4* characteristics. [#] indicates the significant difference ($p < 0.05$) using similar tests in the clinical validation dataset. Abbreviations: CU: Cognitively unimpaired (Cognitively normal or subjective cognitive decline); MCI: Mild cognitive impairment; AD-D: Alzheimer's disease dementia;

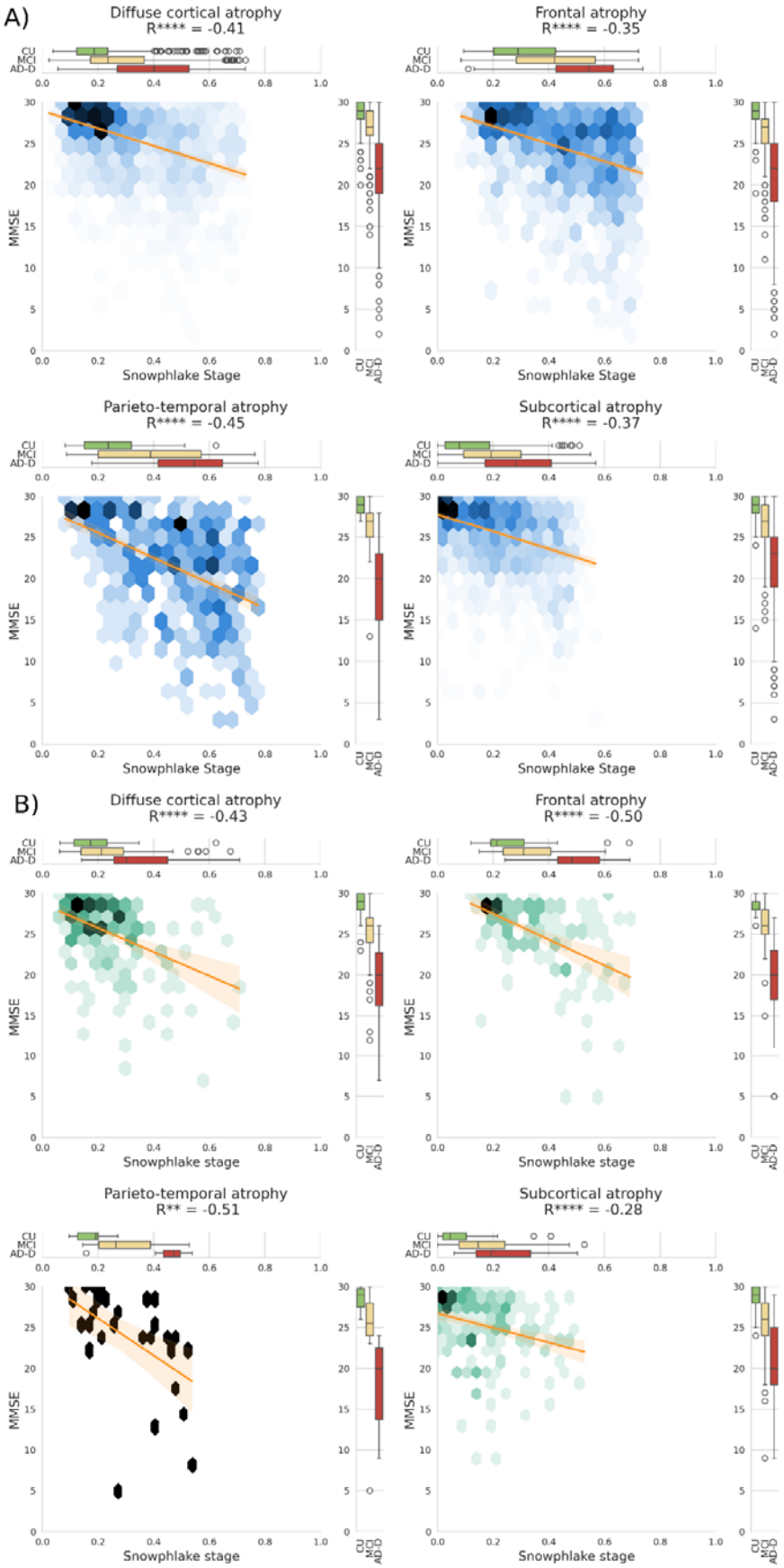

Figure 3: Experiment 1: Correlation of the estimated stage (measuring atrophy severity) using Snowphlake with MMSE in A) clinical-validation cohort B) external-validation cohort. Figures A) and B) both consists of 4 hex-plots, one for each subtype assigned by the trained Snowphlake model. The colour of a bin in the hex-plot denotes the relative proportion of the participants. The boxplot on top of each hex-plot shows the distribution of estimated Snowphlake stage for the participants in the different clinical groups. The boxplot at the right of each hex-plot shows the distribution of MMSE in the different clinical groups. The line overlaying on each hex-plot shows the regression line between MMSE and Snowphlake's stage. The text on top of each hex-plot shows the correlation coefficient (R) between estimated stage and MMSE. The asterisk (*) next to R denotes the significance level. * corresponds to $p < 0.05$; ** corresponds to $p < 0.01$; *** corresponds to $p < 0.001$; **** corresponds to $p < 0.0001$. Abbreviations: CU: Cognitively unimpaired (Cognitively normal or subjective cognitive decline); MCI: Mild cognitive impairment; AD-D: Alzheimer's disease dementia;

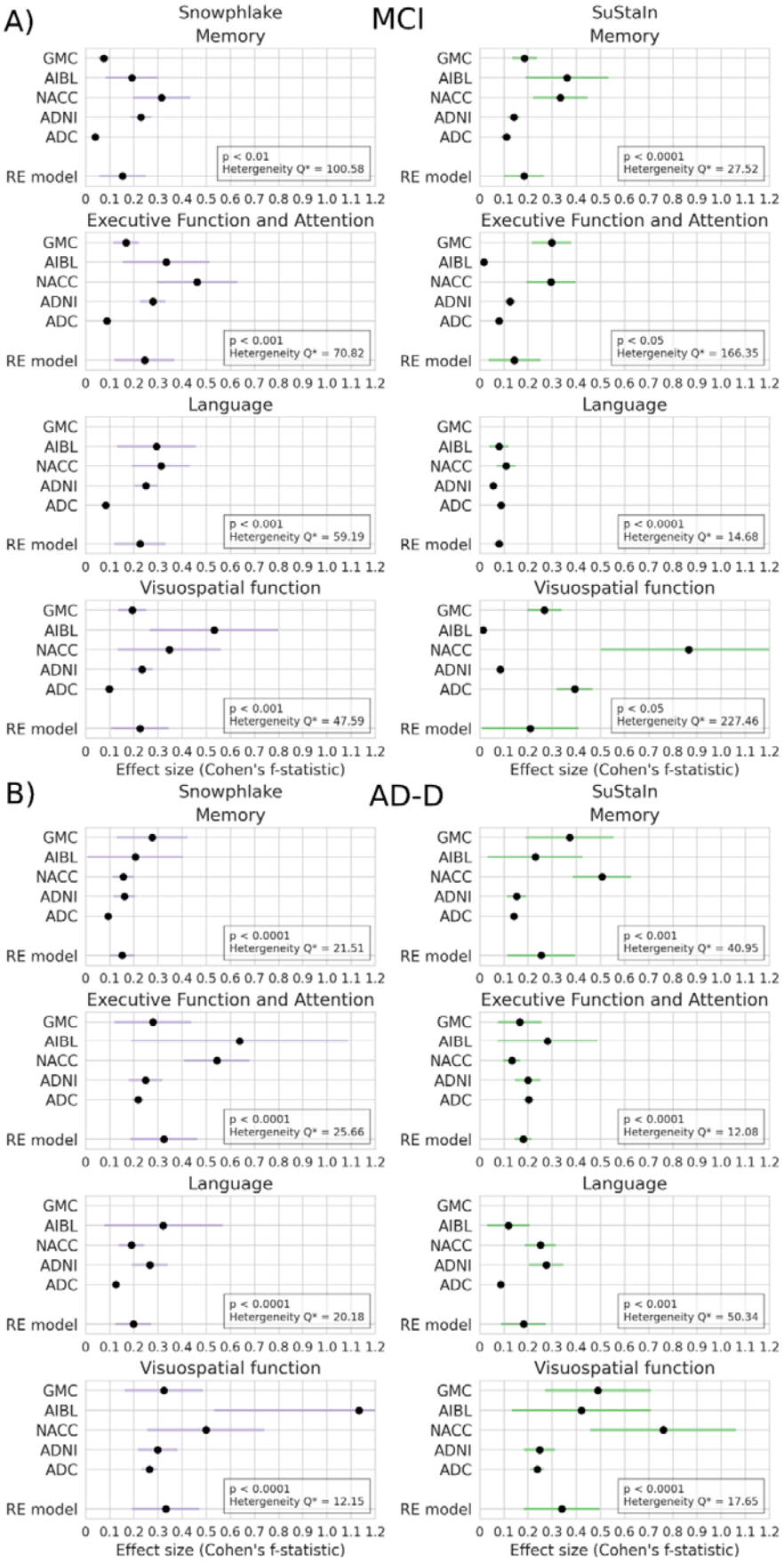

Figure 4: Experiment 2: Cognitive domain differences between subtypes assigned in the Aβ+ validation datasets. Cognitive domain differences are shown for subtypes assigned by Snowphlake (left) and SuStaIn (right) in **A)** MCI patients and **B)** AD-D patients. Each sub-plot shows the effect size (Cohen's f-statistic) and its confidence internal for a cognitive domain in 5 different cohorts within the Aβ+ validation datasets. The combined effect-size of the random effect (RE) model obtained via meta-analysis across the different cohorts, and the corresponding confidence internal is shown within each subplot as well. The p-value corresponding to the RE model and the Cochran's Q statistic measuring heterogeneity across cohorts is shown at the bottom right of each sub-plot. The Q* indicates that the shown Cochran's Q statistic is significant (< 0.0001).

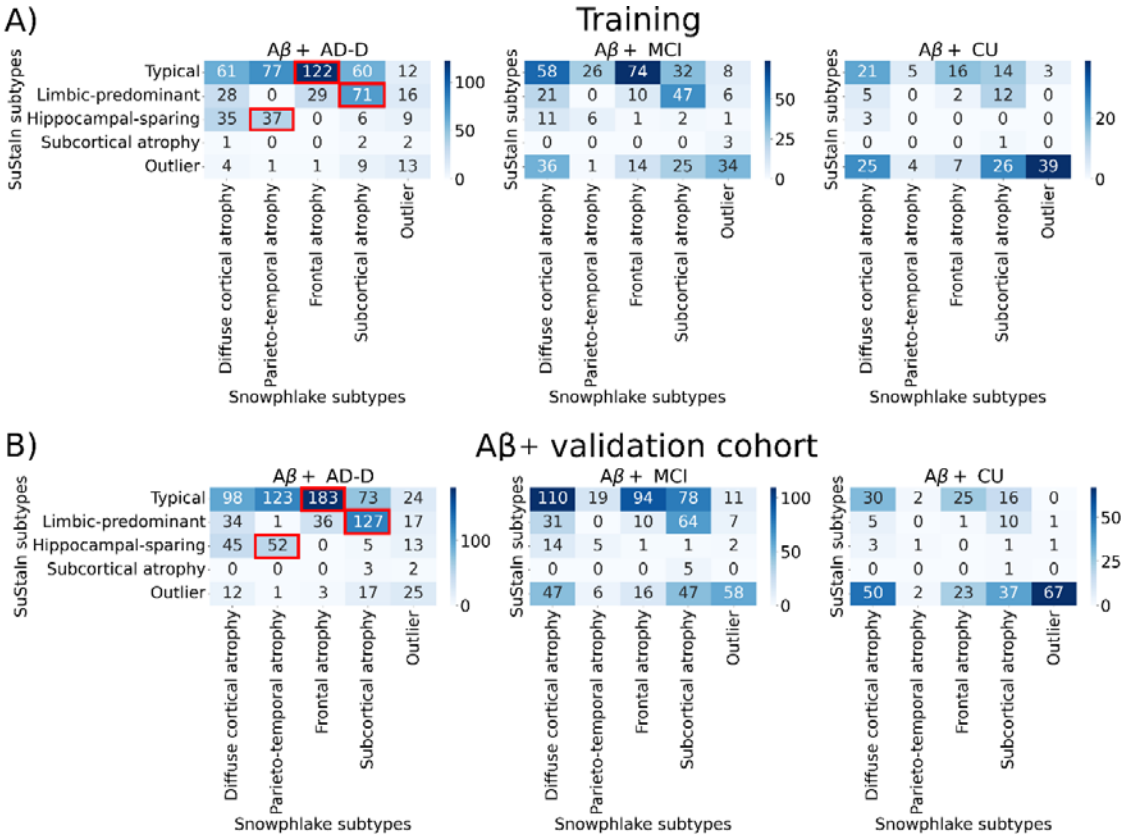

Figure 5: **Experiment 3**: **Concordance of Snowphlake and SuStaIn subtypes. A)** shows the contingency matrix of estimated atrophy-based subtypes using Snowphlake and SuStaIn for participants in the training dataset, in different clinical stages of the disease. **B)** shows a similar contingency matrices for participants in the Aβ+ validation dataset, in different clinical stages of the disease. The squares marked in red in the contingency matrix for AD-D patients correspond to the frequently co-occurring subtypes between SuStaIn and Snowphlake, also referred to as concordant subtypes. Abbreviations: CU: Cognitively unimpaired (Cognitively normal or subjective cognitive decline); MCI: Mild cognitive impairment; AD-D: Alzheimer's disease dementia;

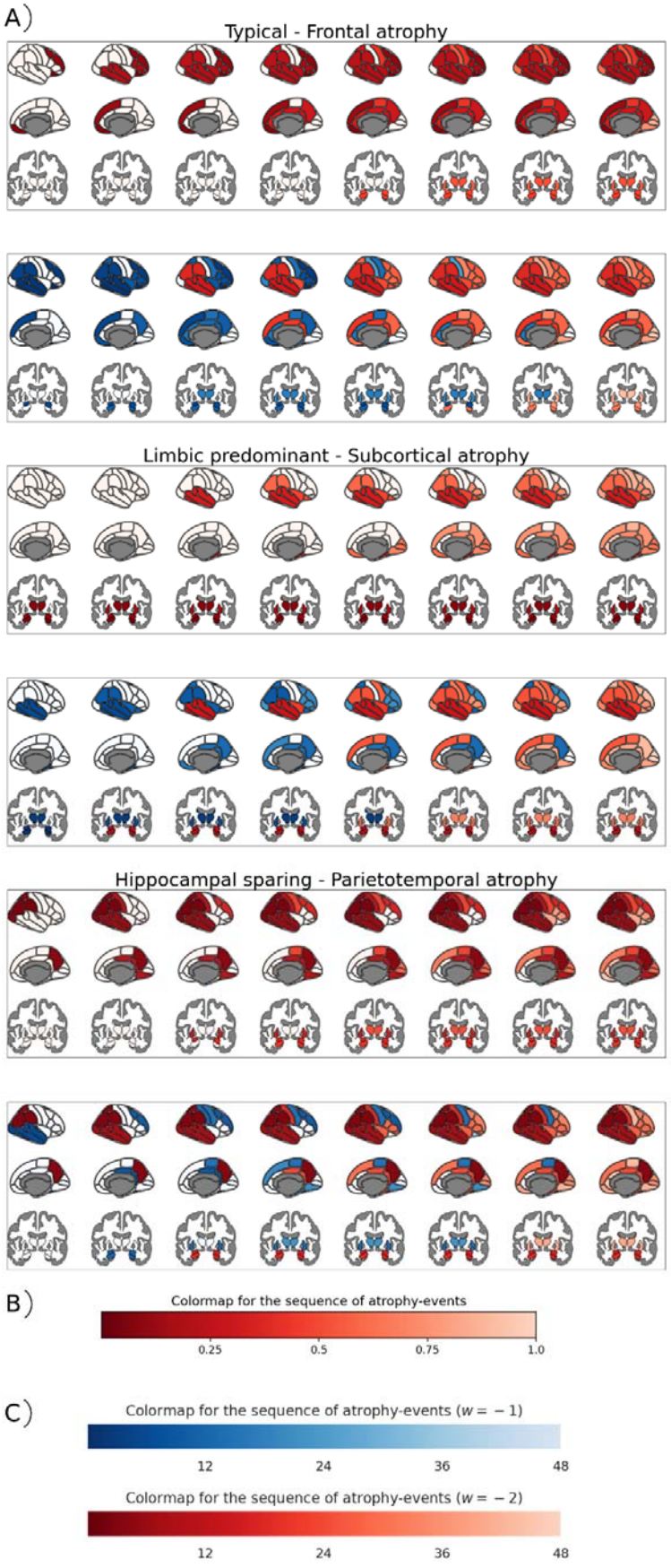

Figure 6: **Experiment 3: Snowphlake and SuStaIn modelling of the Aβ+ participants in the three identified concordant subtypes. A)** For each concordant subtype, the top row depicts the sequence of atrophy-events obtained using DEBM, the methodological equivalent of Snowphlake with 1-subtype. The bottom row depicts the sequence of atrophy-events obtained using w-score EBM the methodological equivalent of SuStaIn with 1-subtype. Within each subtype, the x-axis corresponds to the stage of the disease. Each column shows the brain in its lateral, medial, and subcortical views, with the regions that is expected to be abnormal at this stage. **B)** shows the scale of the colour map used for DEBM plots goes from 0 to 1, where 0 represents a region becoming abnormal at the earliest stages of the disease and 1 represents late stage. **C)** shows the scale of the colour map used for w-score EBM plots, in which regions that are expected to be mildly affected ($w = -1$) are shown in shades of blue, and severely affected ($w = -2$) in shades of red, and unaffected regions in white. The scale for the color map goes from 1 to 48, where 1 represents a region getting affected at the earliest stages of the disease and 48 represents late stage.