Comparison of subtyping methods for neuroimaging studies in Alzheimer’s disease: a call for harmonization

Rosaleena Mohanty¹*, Gustav Mårtensson¹*, Konstantinos Poulakis¹, J-Sebastian Muehlboeck¹, Elena Rodriguez-Vieitez¹, Konstantinos Chiotis¹, Michel J. Grothe²,³, Agneta Nordberg¹,⁴, Daniel Ferreira¹# and Eric Westman¹,⁵#, for the Alzheimer’s Disease Neuroimaging Initiative⁶

Affiliations:

¹Division of Clinical Geriatrics, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden

²Unidad de Trastornos del Movimiento, Servicio de Neurología y Neurofisiología Clínica, Instituto de Biomedicina de Sevilla, Hospital Universitario Virgen del Rocío/CSIC/Universidad de Sevilla, Seville, Spain.

³Clinical Dementia Research Section, German Center for Neurodegenerative Diseases (DZNE), Rostock, Germany

⁴Theme Aging, Karolinska University Hospital, Stockholm, Sweden

⁵Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology and Neuroscience, King’s College London, London, UK

¹Data used in this study were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.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 the analysis or writing of this report. A complete listing of
ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

*Shared first authors

#Shared senior authors

**Corresponding Author:**

Rosaleena Mohanty, rosaleena.mohanty@ki.se

**Short running title:** Towards harmonization of subtyping methods

**Keywords:** Alzheimer's disease, subtypes, structural MRI, tau PET, method comparison
ABSTRACT

Biological subtypes in Alzheimer’s disease, originally identified on neuropathological data, have been translated to in vivo biomarkers such as structural magnetic resonance imaging (sMRI) and positron emission tomography (PET), to disentangle the heterogeneity within Alzheimer’s disease. Although there is methodological variability across studies, comparable characteristics of subtypes are reported at the group level. In this study, we investigated whether group-level similarities translate to individual-level agreement across subtyping methods, in a head-to-head context. We compared five previously published subtyping methods. Firstly, we validated the subtyping methods in 89 amyloid-beta positive Alzheimer’s disease dementia patients (reference group: 70 amyloid-beta negative healthy individuals) using sMRI. Secondly, we extended and applied the subtyping methods to 53 amyloid-beta positive prodromal Alzheimer’s disease and 30 amyloid-beta positive Alzheimer’s disease dementia patients (reference group: 200 amyloid-beta negative healthy individuals) using sMRI and tau...
PET. Subtyping methods were implemented as outlined in each original study. Group-level and individual-level comparisons across methods were performed. Each individual subtyping method was replicated, and the proof-of-concept was established. At the group level, all methods captured subtypes with similar patterns of demographic and clinical characteristics, and with similar cortical thinning and tau PET uptake patterns. However, at the individual level large disagreements were found in subtype assignments. Although characteristics of subtypes are comparable at the group level, there is a large disagreement at the individual level across subtyping methods. Therefore, there is an urgent need for consensus and harmonization across subtyping methods. We call for establishment of an open benchmarking framework to overcome this problem.
| Abbreviation | Full Form |
|--------------|-----------|
| Aβ           | amyloid-beta |
| AD           | Alzheimer’s disease |
| ADNI         | Alzheimer’s disease neuroimaging initiative |
| APOE         | apolipoprotein E |
| AVRA         | automatic visual ratings of atrophy |
| FDG          | fluorodeoxyglucose |
| HC           | healthy control |
| ICV          | intracranial volume |
| MMSE         | mini mental state examination |
| NFT          | neurofibrillary tangle |
| PET          | Positron emission tomography |
| PVC          | partial volume correction |
| RID          | participant identifier |
| sMRI         | structural magnetic resonance imaging |
| SUVR         | standardized uptake value |
INTRODUCTION

The study of biological subtypes has opened a great opportunity to unravel the heterogeneity within Alzheimer’s disease (AD). The topic was rekindled in 2011 by the seminal study from Murray et al. (Murray et al. 2011), and during the last five years it has exploded with numerous structural magnetic resonance imaging (sMRI) subtyping studies (see (Ferreira, Nordberg, and Westman 2020) for a review). In 2018, the first tau positron emission tomography (PET) subtyping study was published (Whitwell et al. 2018), and more are expected to come in the near future.

However, studies investigating AD subtypes differ considerably, with almost no methodological consensus. Murray et al. (Murray et al. 2011) based subtyping on postmortem tau neurofibrillary tangle (NFT) counts in the hippocampus and three cortical regions. All patients were at Braak’s stage V or VI (Braak and Braak 1995) and were classified into three subtypes according to the 25th and 75th percentiles in the hippocampus-to-cortex index:
typical AD, limbic-predominant AD, and hippocampal-sparing AD. Byun et al. (Byun et al. 2015) translated this subtyping method to sMRI data using volumes of the same brain regions as in Murray’s method, but defined abnormality as -1 standard deviation from age-, sex-, and intracranial volume (ICV)-adjusted normative data of healthy controls. This method identified a fourth subtype: minimal atrophy AD (Byun et al. 2015). In contrast, Risacher et al. (Risacher et al. 2017) followed the 25th and 75th percentiles procedure using the hippocampus-to-cortex index, but extended the three cortical regions used by Murray et al. (Murray et al. 2011) to seven regions. Risacher et al. (Risacher et al. 2017) also corrected for age, sex, and ICV, but they based this correction on a reference group of amyloid-beta negative (Aβ-) healthy controls, and used a different correction method additionally including the MRI field strength. Ferreira et al. and follow-up studies from our lab used visual rating scales of brain atrophy in medial temporal, frontal, and posterior cortices (Ferreira et al. 2017, 2018, 2019; Persson et al. 2017; Oppedal et al. 2019; Machado et al. 2020; Ekman, Ferreira, and Westman 2018), and determined clinical cut points for abnormality (Ferreira et al. 2015). We also
used unsupervised clustering in another cross-sectional study by Poulakis et al. (Poulakis et al. 2018), which was recently extended for subtyping on longitudinal data (Poulakis et al. 2019). Other groups used different unsupervised clustering methods (Noh et al. 2014; Na et al. 2016; Hwang et al. 2016; Dong et al. 2015, 2017; Varol et al. 2017; Zhang et al. 2016; Park et al. 2017), highlighting the methodological variability across studies. Additionally, Charil et al. (Charil et al. 2019) recently translated Murray’s method to tau PET while Whitwell et al. (Whitwell et al. 2018) applied a clustering method on tau PET data.

Despite this variability, all these studies tend to identify subtypes with similar characteristics, arguing for validation (see (Ferreira, Nordberg, and Westman 2020) for a review). However, this validation is reported at the group level. The ultimate goal of investigating heterogeneity in AD is to understand individual variability, hence, necessitating individual-level validation. Surprisingly, no head-to-head comparison of subtyping methods has been published so far. Such a comparison arises as an urgent and important step.
towards facilitating consistent progress in this field, especially with the current surge in subtyping studies using sMRI investigating subtype or disease progression (Poulakis et al. 2019; Marinescu et al. 2019; Young et al. 2018) and tau PET (Whitwell et al. 2018; Jeon et al. 2019; Charil et al. 2019). To illustrate this problem, in the present study, we applied different subtyping methods reported in five previous studies (Murray et al. 2011; Risacher et al. 2017; Byun et al. 2015; Ferreira et al. 2017; Poulakis et al. 2018; Charil et al. 2019) on sMRI and tau PET data from the same cohort. Thereby, we substantiated our claim for the need for harmonizing subtyping methods, which aims at achieving consensus at group- and individual-levels despite methodological differences. In our primary analyses, we performed a head-to-head comparison and report subtypes’ frequencies, characteristics, and cortical thickness and tau PET uptake maps from the different methods. In our secondary analyses, we investigated how methodological variations influence the performance of the different subtyping methods. We hypothesized that across subtyping studies, the comparability of subtypes at the group level may not translate to the individual level.
METHODS

Participants

All participants were selected from the Alzheimer’s Disease Neuroimaging Initiative (ADNI; http://adni.loni.usc.edu/). The goal of the ADNI (launched in 2003, principal investigator: Michael W. Weiner; (Mueller et al. 2005)) is to measure the progression of prodromal AD and early AD using MRI, PET, biomarkers, and clinical and neuropsychological assessments. We included two separate ADNI cohorts:

Firstly, since subtypes have been predominantly identified in AD dementia, we validated the previously published subtyping methods using sMRI in a cohort of 89 AD dementia patients (Aβ+) from ADNI-1. We also included a control
group of 70 Aβ- healthy individuals (HC). Amyloid status was determined by cerebrospinal fluid biomarkers (Aβ$_{1-42}$ cut-off = 192 pg/mL) (Shaw et al. 2009).

Secondly, subtyping was applied to a cross-sectional cohort of 84 patients (54 Aβ+ prodromal AD patients, 30 Aβ+ AD dementia patients) subsampled from ADNI-2 and -3 using sMRI and tau PET. The control group was comprised of 200 Aβ- HC. Amyloid status was determined through amyloid PET (florbetapir SUVR cut-off = 1.11; (Joshi et al. 2012) or florbetaben SUVR cut-off = 1.08, following ADNI’s current recommendation, http://adni.loni.usc.edu/).

We will refer to these two cohorts as the sMRI cohort (ADNI-1, AD dementia patients) and the sMRI-tauPET cohort (ADNI-2 and -3, prodromal AD and AD dementia patients). We validated the previously published methods (Charil et al. 2019; Risacher et al. 2017; Byun et al. 2015; Ferreira et al. 2017; Poulakis et al. 2018) in the sMRI cohort and extended our analyses to the sMRI-tauPET cohort. The study protocol followed by all participating centers within...
the ADNI was approved by their respective institutional review board.

Informed and written consent was obtained from all the participants.

**MRI and PET imaging**

*MRI acquisition and processing*

3-D accelerated T1-weighted sequences were acquired with sagittal slices and voxel size $1.1 \times 1.1 \times 1.2$ mm$^3$. MRI data for the ADNI-1 were acquired on 1.5T scanners, and MRI data for ADNI-2 and -3 were acquired on 3.0T scanners.

For the sMRI cohort, processed data were already available from our previous studies (Ferreira et al. 2017; Poulakis et al. 2018). For methods from other labs (Risacher et al. 2017; Byun et al. 2015) and for all the methods in the sMRI-tauPET cohort, data were unavailable, so we processed the sMRI through TheHiveDB system (Muehlboeck, Westman, and Simmons 2014) with FreeSurfer 6.0.0 ([http://freesurfer.net/](http://freesurfer.net/)). Following the cross-sectional stream, quality control of the output from FreeSurfer was conducted visually. Automatic region of interest parcellation yielded volumetric measures for cortical and subcortical brain structures (Desikan et al. 2006; Destrieux et al. 2010; Fischl et al. 2002). For the subtyping method using visual rating scales (Ferreira et al. 2017), the rating scales were computed automatically using AVRA (Automatic Visual Ratings of Atrophy) v0.8
(https://github.com/gsmartensson/avra_public) (Mårtensson, Ferreira, Cavallin, et al. 2019), a deep learning model trained on over 3000 MRI scans rated by an expert neuroradiologist with excellent inter-rater agreement (Mårtensson, Ferreira, Granberg, et al. 2019).

**Tau PET acquisition and processing**

Tau PET scans were collected using PET/CT scanners. $^{[18]}$FAV-1451 was injected with a dosage of 370 MBq (10.0 mCi) ± 10% and scans were acquired between 75-105 min post-injection. Dynamic acquisition was 30 min long and comprised of $6 \times 5$ min frames. For each tau PET scan, a sMRI was available within 90 days (except in 3 AD dementia and 5 prodromal AD patients, >90 days).

For subtyping methods using tau PET (Murray et al. 2011; Risacher et al. 2017; Byun et al. 2015; Charil et al. 2019), processing was performed using the PetSurfer Toolbox (Greve et al. 2016) within FreeSurfer 6.0.0. AV-1451 images were co-registered onto the corresponding FreeSurfer-processed
sMRI. The regions (cortical and subcortical grey matter) estimated for each individual were consistent with those used for sMRI-based subtypes (Desikan et al. 2006). Partial volume correction (PVC) was applied using the symmetric geometric matrix method (ROUSSET and OG 1998). AV-1451 signal was quantified in each region as the standardized uptake value ratio (SUVR), computed with the cerebellum grey matter as the reference region with PVC.

Subtyping methods

Based on two recent systematic reviews (Ferreira, Nordberg, and Westman 2020; Habes et al. 2020), we identified four sources of methodological variation in subtyping studies:

(i) Type of method (hypothesis-driven vs. data-driven).

(ii) Definition of subtype (dependent on the sample of study vs. dependent on an external reference group).
(iii) *Modality* (postmortem NFT vs. sMRI vs. tau PET).

(iv) *Measure* (regional NFT count vs. automated regional volumes/SUVR values vs. gross visual ratings).

The method proposed by Murray et al. (Murray et al. 2011) is the only one based on postmortem NFT count and motivated subsequent neuroimaging studies. In this study, we focused on neuroimaging-based methods based on five subtyping studies, covering all these levels of methodological variation: Risacher et al. (Risacher et al. 2017), Byun et al. (Risacher et al. 2017), Ferreira et al. (Ferreira et al. 2017), Poulakis et al. (Poulakis et al. 2018), and Charil et al. (Charil et al. 2019). Each subtyping method was implemented to replicate the original method as closely as possible, as elaborated further in Table 1, Figure 1 and Supplementary Table 1. We also translated some sMRI-based methods to tau PET to test subtyping based on tau pathology (Risacher et al. 2017; Byun et al. 2015). For Byun's method on tau PET, we identified a minimal tau subtype that is not captured by Charil’s or Risacher’s methods.
Quantification of AV-1451 signal in the hippocampus, a key region for subtyping in many studies (Charil et al. 2019; Risacher et al. 2017; Byun et al. 2015), is contentious (Lemoine et al. 2018; Lee et al. 2018). Hence, we additionally applied subtyping using the entorhinal cortex instead of the hippocampus, also facilitating comparability with the study by Whitwell et al. (Whitwell et al. 2018).
Methodological variations

As a secondary objective, we implemented the following methodological variations, evaluating their potential impact on agreements among subtyping methods:

i. The effect of using three vs. seven cortical regions in Risacher’s method

Although Risacher et al. (Risacher et al. 2017) translated Murray’s method (Murray et al. 2011) to sMRI, Risacher’s method included seven cortical regions instead of the original three regions in Murray’s method. Here, we compared these two versions of Risacher’s method: with three vs. seven cortical regions.

ii. The effect of statistical corrections for ICV and age on sMRI methods

In our primary analysis, we evaluated the method by Risacher et al. (Risacher et al. 2017) (seven cortical regions) by adjusting for ICV and age using a single regression model for both covariates. Here,
we evaluated the impact of adjusting for ICV only, or adjusting for ICV and age using separate regression models for each covariate.

We also performed these comparisons for Risacher’s method using three cortical regions.

iii. The effect of statistical corrections for age on tau PET methods

In the primary analysis of tau PET-based subtyping (Charil et al. 2019; Risacher et al. 2017; Byun et al. 2015), potential covariates were not considered. Correction for ICV is not necessary unlike in sMRI methods, but age may potentially affect tau PET SUVR (Schöll et al. 2016). Here, we compared subtyping with age-corrected SUVR and uncorrected SUVR.

iv. The effect of PVC on tau PET-based subtyping methods

In the primary analysis, we used PVC for reliably quantifying tau PET SUVR, accounting for any off-target binding, especially in the hippocampus (Lowe et al. 2016; Ikonomovic et al. 2016). Here, we compared subtyping between PVC SUVR and non-PVC SUVR.
Statistical analysis

We compared subtyping methods at the group-level in terms of age, sex, Mini Mental State Exam (MMSE), education, and APOE ε4 status. Within each subtyping method, hypothesis testing was performed to compare the distribution of subtypes with the Kruskal-Wallis test. A \( p \)-value \( \leq 0.05 \) was deemed significant. Group-level cortical thickness and PVC tau PET uptake maps were generated by comparing each subtype with the healthy controls. In each hemisphere, data were smoothed onto the surface using a 10 mm Gaussian kernel with a full width at half maximum. A general linear model was fitted at each vertex. All maps were visualized at \( p \leq 0.01 \) (uncorrected).

Individual-level agreement among subtyping methods was quantified by Cohen’s kappa (\( \kappa < 0 \), no agreement; \( \kappa = 0–0.20 \), slight agreement; \( \kappa = 0.21–0.40 \), fair agreement; \( \kappa < 0.41–0.60 \), moderate agreement; \( \kappa = 0.61–0.80 \), substantial agreement; \( \kappa = 0.81–1.0 \), almost perfect agreement) (Landis and Koch 1977).
Data availability

Source data are available as a part of the ADNI. All data generated or analyzed during this study are included within this article and its supplementary information files.

RESULTS

Table 2 (a, b) shows the demographic and clinical characteristics for the sMRI cohort and sMRI-tauPET cohorts respectively.

Validation of subtyping methods in the sMRI cohort

The frequencies of the subtypes in the sMRI cohort were very similar to the frequencies reported in the original studies (Risacher et al. 2017; Byun et al.
2015; Ferreira et al. 2017; Poulakis et al. 2018; Charil et al. 2019), suggesting we could replicate the subtyping methods (Table 3).

Group-level comparison of subtyping methods in the sMRI and sMRI-tauPET cohorts

Figure 2 shows that, at the group-level, the subtyping methods captured similar demographic and clinical characteristics of the subtypes in both cohorts. Typical AD was always the most frequent subtype and showed a greater frequency of males and lower MMSE scores relative to the other subtypes. Limbic-predominant AD showed lower MMSE scores relative to hippocampal-sparing AD. Hippocampal-sparing AD was the subtype with the lowest frequency of APOE ε4 carriers. Minimal atrophy/minimal tau AD included younger individuals and showed higher MMSE scores.

Figures 3-4 and Supplementary Table 2 show that, at the group-level, the subtyping methods captured similar cortical thickness and PVC tau PET
uptake maps of the subtypes relative to healthy individuals. Cortical thinning and elevated tau PET uptake included widespread regions in typical AD; temporal and limbic regions in limbic-predominant AD; frontal or parietal regions in hippocampal-sparing AD; and relatively fewer regions in minimal atrophy/minimal tau AD, across all subtyping methods. Typical and limbic-predominant AD showed smaller hippocampal volume and greater hippocampal tau PET SUVR relative to hippocampal-sparing and minimal atrophy/minimal tau AD (boxplots in Figures 3-4). Figure 5 shows the group-level tau PET uptake maps for entorhinal-based subtyping instead of hippocampus-based subtyping. Compared to hippocampus-based subtyping, albeit similar maps, hippocampal-sparing AD in entorhinal-based subtyping showed no tau PET uptake in lateral temporal lobe regions. Greater tau SUVR in the entorhinal cortex was seen in typical and limbic-predominant AD than hippocampal-sparing and minimal tau AD (boxplots in Figure 5).

Head-to-head comparison of subtyping methods in the sMRI-tauPET cohort
Figure 6a and Figure 6c show the head-to-head comparison of individual-level subtype assignments. Agreement among methods was low, reflected by low values of $\kappa$. Agreement among the tau PET-based methods was relatively higher than that of the sMRI-based methods. Since not all methods identify the minimal atrophy/minimal tau AD, we excluded this subtype in follow-up analyses and observed increased $\kappa$ values in both cohorts and modalities (Figure 6b and Figure 6d). ADNI’s participant identifiers (RID) are listed in Supplementary Figure 2 and in Supplementary Data File.

Methodological variations in the sMRI-tauPET cohort

When supplementing our head-to-head comparisons with several methodological variations, we observed the following (Supplementary Data File):

i. The effect of using three vs. seven cortical regions in Risacher’s method
Results from Risacher’s method using three cortical regions were consistent with Risacher’s method using seven cortical regions (85% agreement).

ii. The effect of statistical corrections for ICV and age on sMRI methods

Relative to Risacher’s method (seven cortical regions and adjusted for ICV and age in a single model), 82% of the individuals were classified consistently when performing the ICV correction only, and 69% when performing the ICV and age correction with separate models. Relative to the variation in Risacher’s method using three cortical regions (and adjusted for ICV and age in a single model), 98% of the individuals were classified consistently when performing the ICV correction only, and 74% when performing the ICV and age correction with separate models. Overall, agreements were better in typical AD (79-88%) compared to the other subtypes (15-83%).

iii. The effect of statistical corrections for age on tau PET methods
Over 80% of the individuals were consistently classified with and without age-adjusted tau SUVR (agreement for: Charil’s method=89%; Risacher’s method=100%; Byun’s method=80%).

iv. The effect of PVC on tau PET-based subtyping methods

Over 80% of the individuals were consistently classified with PVC and non-PVC SUVR (agreement for: Charil’s method=87%; Risacher’s method=89%; Byun’s method=80%). Overall, agreements were better in typical AD (83-94%) compared to the other subtypes (56-78%).

DISCUSSION

The field of biological subtypes of AD has expanded rapidly in the last decade, with numerous recent publications on neuropathological, MRI, and PET data. However, the great methodological variability is complicating reaching a definitive understanding of the heterogeneity within AD. The current study is the first head-to-head comparison of several subtyping
methods in the same cohort. We found that different methods identify
subtypes that are largely comparable at the group level (similar frequencies,
demographic, clinical characteristics, cortical thinning and tau PET uptake).
However, strikingly, the individual-level agreement among subtyping methods
is very low when compared head-to-head. This result may have important
implications for advancing the implementation of precision medicine. Below
we discuss several factors that may explain this finding and ways to minimize
this problem in future studies.

Comparability across studies at the group level suggests a convergence of
results and initial consensus on the existence of three to four major subtypes:
typical, limbic-predominant, and hippocampal-sparing AD in all the studies,
and minimal atrophy AD in several studies. Minimal atrophy AD is only
identified when considering disease severity, while the other subtypes are
identified when considering typicality (Ferreira, Nordberg, and Westman
2020). The dimensions of severity and typicality were defined in a recent
conceptual framework for biological subtypes of AD (Ferreira, Nordberg, and
Typicality spans from limbic-predominant to hippocampal-sparing, with typical AD in-between. Severity differentiates minimal atrophy from typical AD, accounting for neurodegeneration.

The seminal study by Murray et al. (Murray et al. 2011) based subtyping on tau NFT in the hippocampus and three cortical regions. Importantly, all the patients had a pathological diagnosis of AD with Braak stage of V or VI (Braak and Braak 1995). This means that all patients had NFT in the hippocampus by definition, and the method focused on separating the subset of patients with NFT predominantly in the hippocampus (limbic-predominant AD) from those with NFT predominantly in the cortical regions (hippocampal-sparing AD). Remainder of the patients had a rather balanced NFT count in the hippocampus and cortical regions, and were classified as typical AD.

Murray’s method (Murray et al. 2011) motivated many subsequent sMRI studies (Ferreira, Nordberg, and Westman 2020; Habes et al. 2020). However, these studies rely on sMRI, a marker of unspecific
neurodegeneration. This raises several problems. Firstly, while sMRI can reliably track neuropathologically-defined subtypes (Whitwell et al. 2012), the actual distribution of NFT in sMRI subtypes remains largely unknown. Recent studies have provided interesting preliminary data on tau PET uptake in sMRI-based subtypes (Ossenkoppele et al. 2019; Jeon et al. 2019). Secondly, the published sMRI subtype studies quite likely included patients in Braak NFT stage IV or lower. Thirdly, most sMRI studies investigated cohorts including both amyloid-beta positive and negative AD dementia patients except a few (Risacher et al. 2017; ten Kate et al. 2018), while all the patients in Murray et al. (Murray et al. 2011) had a pathological diagnosis of AD. Fourthly, neurodegeneration is downstream to NFT pathology (Dubois et al. 2014), and there is a time gap until overt brain atrophy can be visually observed or captured by automatic methods for data analysis. Nonetheless, some data-driven methods may capture subtle differences in regional covariance in the absence of overt brain atrophy, mitigating this problem. Altogether, we still need a better understanding of the correspondence between neuropathologically-, sMRI-, and tau PET-defined subtypes. A major
contribution of our current study is that subtypes identified with sMRI and tau PET are not interchangeable at the individual level.

At the group level, findings for the demographic and clinical measures were in agreement with previously reported studies and a recent meta-analysis (Ferreira, Nordberg, and Westman 2020). Broadly, typical AD was the most frequent subtype; typical and limbic-predominant AD were older in comparison to the hippocampal-sparing and minimal atrophy AD; MMSE scores were mostly comparable across subtypes with minimal atrophy AD showing the highest scores; a lower proportion of APOE ε4 carriers belonged to hippocampal-sparing relative to typical and limbic-predominant AD; and hippocampal-sparing AD had the highest levels of education.

Overall, head-to-head comparisons revealed greater agreement of tau PET-based methods than the sMRI-based methods. This could be potentially attributed to: (i) lower resolution of tau PET and smaller proximity of the key regions involved in subtyping, (ii) comparison of merely three tau PET-based
subtyping methods with relatively less methodological variability, or (iii) a more consistent and direct emulation of postmortem NFT captured in Murray et al. (Murray et al. 2011) by tau PET compared to sMRI which can capture variance unrelated to subtyping. Future tau PET-based subtyping methods could shine light on this finding. Low levels of agreement across subtyping methods at the individual-level could be ascribed to a combination of one or more of the following factors: (i) variation in cut points of atrophy or tau uptake used to define the subtypes which may differ from dataset to dataset; (ii) accounting for (or lack thereof) the two dimensions of subtypes, namely typicality and severity, to different degrees by different methods; (iii) forced allocation of each individual into a subtype without an associated measure of (un)certainty; (iv) accounting for (or lack thereof) within-subtype variability (i.e. strong/weak resemblance) in the biological profiles of individuals assigned to the same subtype. We call for investigation of these factors as a promising avenue to increase the agreement between subtyping methods in the future.
Biologically, the head-to-head agreements are best understood by considering individual exemplars. A consistent scenario is RID 2239: across the sMRI methods, this individual was classified as hippocampal-sparing or minimal atrophy AD whereas, across the tau PET methods, the individual was classified as typical AD. The difference in sMRI-based subtyping could be attributed to differences in cut points for abnormality across methods. The fact that the corresponding tau PET-based subtype was typical AD (higher severity) could suggest greater tau pathology relative to structural atrophy. A more challenging case is RID 6377: across the sMRI methods, this individual was classified as typical, hippocampal-sparing, limbic-predominant, or minimal atrophy AD, whereas across the tau PET-based method, the individual was classified as limbic-predominant AD. Some differences in sMRI-based subtyping are relatively more plausible than others, considering the above-mentioned typicality and severity dimensions (Ferreira, Nordberg, and Westman 2020). To instantiate, it may be plausible that this individual demonstrated typical AD (with one method (Risacher et al. 2017)) and limbic-predominant AD (with another method (Ferreira et al. 2017)), as these two
subtypes are close to each other along the typicality dimension (Ferreira, Nordberg, and Westman 2020). However, classification as limbic-predominant AD (with one method (Ferreira et al. 2017)) and hippocampal-sparing AD (with another method (Byun et al. 2015)) seem incompatible, since these two subtypes correspond to the extremities of the typicality dimension. Therefore, a classification with all four subtypes for the same individual leaves the case biologically uninterpretable, calling for consensus across subtyping in the field as we aim for precision medicine.

Despite having several caveats, previous neuroimaging-based subtyping studies have made important contributions. Byun et al. (Byun et al. 2015) and Risacher et al. (Risacher et al. 2017) translated the NFT-based method by Murray et al. (Murray et al. 2011) to sMRI data, and Charil et al. (Charil et al. 2019) translated the method to tau PET. Our analyses of methodological variations showed that the age correction made a stronger impact on agreements among methods than the number of cortical regions or the PVC. This impact was more prominent for sMRI-based methods than for tau PET-
based methods; and for limbic-predominant and hippocampal-sparing subtypes than for typical AD. Contribution of aging to hippocampal atrophy may be at the basis of this finding. Lower disagreement in typical AD relative to the other subtypes is akin to the diagnostic challenge in the clinical setting.

An interesting result of our study is that the method of adjustment (single model for all covariates vs. separate models for each covariate) increased the disagreement. Future studies should take this finding into account when deciding on how to correct for potential confounders.

Ongoing research is moving the field forward by characterization of subtypes not only in AD dementia but also at earlier stages such as prodromal AD (Machulda et al. 2019; ten Kate et al. 2018; Zhang et al. 2016; Young et al. 2018). Preliminary data show that such characterization could be extended and evaluated at even the earliest stages of preclinical AD or individuals with subjective cognitive decline (Jung et al. 2016). In speculation, relative to full-blown dementia, atrophy levels are likely modest even if there exists overt tau pathology at pre-dementia stages. This could result in a greater dissociation
between atrophy and tau pathology, further leading to lower agreement across subtyping methods. In this scenario, group-level comparisons alone are insufficient. Individual-level agreement is thus warranted, and lack thereof will prevent or delay the use of subtyping in clinical routine, clinical trials, and research. Therefore, there is an urgent need for harmonization of the different subtyping methods.

To this end, we advocate for establishing a framework for benchmarking for future studies. A possibility could be selection of a well-characterized cohort (preferably with multimodal antemortem and postmortem data in a longitudinal setting). This could include preparing a dataset comprised of cognitively normal individuals (amyloid negative) and individuals on the AD continuum (preclinical AD, prodromal AD and AD dementia). Multiple longitudinal biomarkers during life, such as neuroimaging such as MRI (structural, diffusion, functional, etc.) and PET (FDG, amyloid, tau), cerebrospinal fluid, plasma and neuropsychological measures, could enable characterization of the subtypes in vivo while neuropathological assessments can provide a
ground truth for subtyping. Unimodal (based on a single image modality) as well as multimodal (based on combination of image modalities) subtypes should be differentiated and demonstrated within the same cohort. Additionally, establishment of clear evaluation metrics would allow for comparison of the performance of the subtyping methods and could include group-level characteristics, individual-level results, cut points for each measure used for subtyping, variability in cut points after accounting for potential covariates, certainty of assignment of subtype, variability in biomarker profiles within the same subtype, etc. Greater similarity across multiple evaluation metrics across methods would thus ensure harmonization across subtyping methods. The dataset should be standard so that it can be utilized by future subtyping methods to ensure individual-level consistency across methods. The dataset should also be open and accessible to all researchers in the field. Once validated, the subtyping method could obviously be extended to independent populations and data. As a preliminary step, we provide all the data used for subtyping in this study along with ADNI RIDs (Supplementary Data File).
This study has some limitations. The cohort was part of the ADNI, which has strict selection criteria and excludes individuals with non-amnestic presentations or cerebrovascular pathology. It is likely that agreement among subtyping methods is different in clinically oriented or more heterogeneous cohorts. The number of AD dementia patients was limited, and prodromal AD patients, in which the degree of atrophy may be smaller than in AD dementia, were overrepresented within the sMRI-tau PET cohort used to demonstrate the lack of consensus in subtyping. However, the additional and relatively large sMRI cohort of AD dementia patients strengthen and illustrate the case in point. Hypothesis-driven methods are well covered in our study (Murray et al. 2011; Risacher et al. 2017; Byun et al. 2015; Ferreira et al. 2017; Charil et al. 2019). However, previous subtyping studies have applied many different data-driven methods. Methods, especially involving clustering, can differ on if and/or how they account for critical aspects of certainty of subtype allocation and variability within each identified subtype. We selected Poulakis’ method (Poulakis et al. 2018), which resulted in notably distinct subtyping potentially
due to being most methodologically different from the rest, and our current study cannot provide direct insight on methods used by other groups (Ferreira, Nordberg, and Westman 2020; Habes et al. 2020). However, the selection of subtyping methods illustrates the case made in the current study. We based our analyses on cross-sectional tau PET and sMRI data. The next step should be to include longitudinal data. However, the availability of such a dataset is limited at present, particularly for tau PET. Longitudinal data will be relevant to investigate disease progression in the subtypes, disentangling the disagreement due to the temporal lag between NFT accumulation (tau PET) and brain atrophy (sMRI) from pure methodological noise. Finally, the tau PET tracer used in our study, $^{[18}\text{F}]$AV-1451, is a first generation tracer with known off target binding (Leuzy et al. 2019). Better agreement among the tau PET-based subtyping methods than their sMRI counterparts could indicate that they need to be further pursued. Second generation tau PET tracers (e.g., 18F-RO-948, 18F-MK-6240, 18F-PI-2620) would be relatively more sensitive and specific, especially at the preclinical and prodromal stages of AD,
although their longitudinal trajectories remain to be fully investigated and validated (Bischof et al. 2020).

The field of biological subtypes is expanding rapidly with the investigation of multiple modalities/biomarkers and extending to pre-dementia stages and other neurodegenerative diseases (Habes et al. 2020). We conclude that subtyping methods may appear comparable across studies, at the group-level. However, a major finding of the present study is the large disagreement among subtyping methods based on tau PET and especially sMRI at the individual level. Hence, there is an urgent need for consensus and harmonization across subtyping methods. To achieve this, we suggest establishment of an accessible and standard framework for benchmarking. A comprehensive dataset along with clear evaluation metrics will facilitate a fair comparison, ultimately ensuring better agreement among future subtyping methods.

COMPETING INTERESTS

The authors have no competing interests to declare.

FUNDING

This study was supported by the Swedish Foundation for Strategic Research (SSF); the Strategic Research Programme in Neuroscience at Karolinska...
Institutet (StratNeuro); the Swedish Research Council (VR, 2016-02282); the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet; Center for Innovative Medicine (CIMED); the Swedish Alzheimer Foundation; the Swedish Brain Foundation; the Åke Wiberg Foundation; Demensfonden; Stiftelsen Olle Engkvist Byggmästare; Birgitta och Sten Westerberg; Demensförbundet; Stiftelsen för Ålderssjukdomar vid Karolinska Institutet. Michel Grothe is supported by the "Miguel Servet" program [CP19/00031] of the Spanish Instituto de Salud Carlos III (ISCIIFEDER). The funding sources did not have any involvement in the study design; collection, analysis, and interpretation of data; writing of the report; and the decision to submit the article for publication.

ACKNOWLEDGEMENTS

Data collection and sharing for this study 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

https://mc.manuscriptcentral.com/braincom
ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: Alzheimer's Association; Alzheimer's Drug Discovery Foundation; BioClinica, Inc.; Biogen Idec Inc.; Bristol-Myers Squibb Company; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; GE Healthcare; Innogenetics, N.V.; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Medpace, Inc.; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Synarc Inc.; and Takeda Pharmaceutical Company. 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 Disease Cooperative Study at the University of
California, San Diego. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of California, Los Angeles.
REFERENCES

Bischof, Gerard N, Alessandra Dodich, Nicholas J Ashton, Marina Boccardi, Henryk Barthel, Maria C Carrillo, Konstantinos Chiotis, Julie Corre, Jean-François Démonet, and Anton F Gietl. 2020. “Alzheimer’s Disease Biomarker Roadmap 2020: Second-Generation Tau PET Tracers.” In 2020 Alzheimer’s Association International Conference. ALZ.

Braak, Heiko, and E V A Braak. 1995. “Staging of Alzheimer’s Disease-Related Neurofibrillary Changes.” Neurobiology of Aging 16 (3): 271–78.

Byun, Min Soo, Song E Kim, Jinsick Park, Dahyun Yi, Young Min Choe, Bo Kyung Sohn, Hyo Jung Choi, Hyewon Baek, Ji Young Han, and Jong Inn Woo. 2015. “Heterogeneity of Regional Brain Atrophy Patterns Associated with Distinct Progression Rates in Alzheimer’s Disease.” PLoS One 10 (11): e0142756.

Charil, Arnaud, Sergey Shcherbinin, Sudeepti Southekal, Michael D Devous, Mark Mintun, Melissa E Murray, Bradley B Miller, and Adam J Schwarz. 2019. “Tau Subtypes of Alzheimer’s Disease Determined in Vivo Using Flortaucipir PET Imaging.” Journal of Alzheimer’s Disease, no. Preprint: 1–12.

Desikan, Rahul S, Florent Ségonne, Bruce Fischl, Brian T Quinn, Bradford C Dickerson, Deborah Blacker, Randy L Buckner, Anders M Dale, R Paul Maguire, and Bradley T Hyman. 2006. “An Automated Labeling System for Subdividing the Human Cerebral Cortex on MRI Scans into Gyral Based Regions of Interest.” Neuroimage 31 (3): 968–80.

Destrieux, Christophe, Bruce Fischl, Anders Dale, and Eric Halgren. 2010. “Automatic Parcellation of Human Cortical Gyri and Sulci Using Standard Anatomical Nomenclature.” Neuroimage 53 (1): 1–15.

Dong, Aoyan, Nicolas Honnorat, Bilwaj Gaonkar, and Christos Davatzikos. 2015. “CHIMERA: Clustering of Heterogeneous Disease Effects via Distribution Matching of Imaging Patterns.” IEEE Transactions on Medical Imaging 35 (2): 612–21.
Dong, Aoyan, Jon B Toledo, Nicolas Honnorat, Jimit Doshi, Erdem Varol, Aristeidis Sotiras, David Wolk, John Q Trojanowski, Christos Davatzikos, and Alzheimer’s Disease Neuroimaging Initiative. 2017. “Heterogeneity of Neuroanatomical Patterns in Prodromal Alzheimer’s Disease: Links to Cognition, Progression and Biomarkers.” *Brain* 140 (3): 735–47.

Dubois, Bruno, Howard H Feldman, Claudia Jacova, Harald Hampel, José Luis Molinuevo, Kaj Blennow, Steven T DeKosky, Serge Gauthier, Dennis Selkoe, and Randall Bateman. 2014. “Advancing Research Diagnostic Criteria for Alzheimer’s Disease: The IWG-2 Criteria.” *The Lancet Neurology* 13 (6): 614–29.

Ekman, Urban, Daniel Ferreira, and Eric Westman. 2018. “The A/T/N Biomarker Scheme and Patterns of Brain Atrophy Assessed in Mild Cognitive Impairment.” *Scientific Reports* 8 (1): 1–10.

Ferreira, Daniel, Lena Cavallin, E-M Larsson, J-S Muehlboeck, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen, and Simon Lovestone. 2015. “Practical Cut-offs for Visual Rating Scales of Medial Temporal, Frontal and Posterior Atrophy in Alzheimer’s Disease and Mild Cognitive Impairment.” *Journal of Internal Medicine* 278 (3): 277–90.

Ferreira, Daniel, Agneta Nordberg, and Eric Westman. 2020. “Biological Subtypes of Alzheimer Disease: A Systematic Review and Meta-Analysis.” *Neurology* 94 (10): 436–48. https://doi.org/10.1212/WNL.000000000009058.

Ferreira, Daniel, Joana B Pereira, Giovanni Volpe, and Eric Westman. 2019. “Subtypes of Alzheimer’s Disease Display Distinct Network Abnormalities Extending Beyond Their Pattern of Brain Atrophy.” *Frontiers in Neurology* 10: 524.

Ferreira, Daniel, Sara Shams, Lena Cavallin, Matti Viitanen, Juha Martola, Tobias Granberg, Mana Shams, Peter Aspelin, Maria Kristoffersen-Wiber, and Agneta Nordberg. 2018. “The Contribution of Small Vessel Disease to Subtypes of Alzheimer’s Disease: A Study on Cerebrospinal Fluid and Imaging Biomarkers.” *Neurobiology of Aging* 70: 18–29.

https://mc.manuscriptcentral.com/braincom
Ferreira, Daniel, Chloë Verhagen, Juan Andrés Hernández-Cabrera, Lena Cavallin, Chun-Jie Guo, Urban Ekman, J-Sebastian Muehlboeck, Andrew Simmons, José Barroso, and Lars-Olof Wahlund. 2017. “Distinct Subtypes of Alzheimer’s Disease Based on Patterns of Brain Atrophy: Longitudinal Trajectories and Clinical Applications.” Scientific Reports 7: 46263.

Fischl, Bruce, David H Salat, Evelina Busa, Marilyn Albert, Megan Dieterich, Christian Haselgrove, Andre Van Der Kouwe, Ron Killiany, David Kennedy, and Shuna Klaveness. 2002. “Whole Brain Segmentation: Automated Labeling of Neuroanatomical Structures in the Human Brain.” Neuron 33 (3): 341–55.

Greve, Douglas N, David H Salat, Spencer L Bowen, David Izquierdo-Garcia, Aaron P Schultz, Ciprian Catana, J Alex Becker, Claus Svarer, Gitte M Knudsen, and Reisa A Sperling. 2016. “Different Partial Volume Correction Methods Lead to Different Conclusions: An 18F-FDG-PET Study of Aging.” Neuroimage 132: 334–43.

Habes, Mohamad, Michel J Grothe, Birkan Tunc, Corey McMillan, David A Wolk, and Christos Davatzikos. 2020. “Disentangling Heterogeneity in Alzheimer’s Disease and Related Dementias Using Data-Driven Methods.” Biological Psychiatry.

Hwang, Jihye, Chan Mi Kim, Seun Jeon, Jong Min Lee, Yun Jeong Hong, Jee Hoon Roh, Jae-Hong Lee, Jae-Young Koh, Duk L Na, and Alzheimer’s Disease Neuroimaging Initiative. 2016. “Prediction of Alzheimer’s Disease Pathophysiology Based on Cortical Thickness Patterns.” Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring 2: 58–67.

Ikonomovic, Milos D, Eric E Abrahamson, Julie C Price, Chester A Mathis, and William E Klunk. 2016. “[F-18] AV-1451 Positron Emission Tomography Retention in Choroid Plexus: More than ‘Off-target’ Binding.” Annals of Neurology 80 (2): 307–8.

Jeon, Seun, Jae Myeong Kang, Seongho Seo, Hye Jin Jeong, Thomas Fabien Funck, Sang-Yoon Lee, Kee Hyung Park, Yeong-Bae Lee, Byeong Kil Yeon, and Tatsuo Ido. 2019. “Topographical Heterogeneity of Alzheimer’s Disease Based on MR Imaging, Tau PET,
and Amyloid PET.” *Frontiers in Aging Neuroscience* 11: 211.

Joshi, Abhinay D, Michael J Pontecorvo, Christopher M Clark, Alan P Carpenter, Danna L Jennings, Carl H Sadowsky, Lee P Adler, Karel D Kovnat, John P Seibyl, and Anupa Arora. 2012. “Performance Characteristics of Amyloid PET with Florbetapir F 18 in Patients with Alzheimer’s Disease and Cognitively Normal Subjects.” *Journal of Nuclear Medicine* 53 (3): 378–84.

Jung, Na-Yeon, Sang Won Seo, Heejin Yoo, Jin-Ju Yang, Seongbeom Park, Yeo Jin Kim, Juyoun Lee, Jin San Lee, Young Kyoung Jang, and Jong Min Lee. 2016. “Classifying Anatomical Subtypes of Subjective Memory Impairment.” *Neurobiology of Aging* 48: 53–60.

Kate, Mara ten, Ellen Dicks, Pieter Jelle Visser, Wiesje M van der Flier, Charlotte E Teunissen, Frederik Barkhof, Philip Scheltens, Betty M Tijms, and Alzheimer’s Disease Neuroimaging Initiative. 2018. “Atrophy Subtypes in Prodromal Alzheimer’s Disease Are Associated with Cognitive Decline.” *Brain* 141 (12): 3443–56.

Landis, J Richard, and Gary G Koch. 1977. “The Measurement of Observer Agreement for Categorical Data.” *Biometrics*, 159–74.

Lee, Christopher M, Heidi I L Jacobs, Marta Marquié, John A Becker, Nicolas V Andrea, David S Jin, Aaron P Schultz, Matthew P Frosch, Teresa Gomez-Isla, and Reisa A Sperling. 2018. “18F-Flortaucipir Binding in Choroid Plexus: Related to Race and Hippocampus Signal.” *Journal of Alzheimer’s Disease* 62 (4): 1691–1702.

Lemoine, Laetitia, Antoine Leuzy, Konstantinos Chiotis, Elena Rodriguez-Vieitez, and Agneta Nordberg. 2018. “Tau Positron Emission Tomography Imaging in Tauopathies: The Added Hurdle of off-Target Binding.” *Alzheimer’s & Dementia: Diagnosis, Assessment & Disease Monitoring* 10: 232–36.

Leuzy, Antoine, Konstantinos Chiotis, Laetitia Lemoine, Per-Göran Gillberg, Ove Almkvist, Elena Rodriguez-Vieitez, and Agneta Nordberg. 2019. “Tau PET Imaging in Neurodegenerative Tauopathies—still a Challenge.” *Molecular Psychiatry* 24 (8): 1112–
Lowe, Val J, Geoffrey Curran, Ping Fang, Amanda M Liesinger, Keith A Josephs, Joseph E Parisi, Kejal Kantarci, Bradley F Boeve, Mukesh K Pandey, and Tyler Bruinsma. 2016. “An Autoradiographic Evaluation of AV-1451 Tau PET in Dementia.” Acta Neuropathologica Communications 4 (1): 58.

Machado, Alejandra, Daniel Ferreira, Michel J Grothe, Helga Eyjolfsdottir, Per M Almqvist, Lena Cavallin, Goran Lind, Bengt Linderoth, Ake Seiger, and Stefan Teipel. 2020. “The Cholinergic System and Treatment Response in Subtypes of Alzheimer’s Disease.” MedRxiv.

Machulda, Mary M, Emily S Lundt, Sabrina M Albertson, Walter K Kremers, Michelle M Mielke, David S Knopman, Mark W Bondi, and Ronald C Petersen. 2019. “Neuropsychological Subtypes of Incident Mild Cognitive Impairment in the Mayo Clinic Study of Aging.” Alzheimer’s & Dementia 15 (7): 878–87.

Marinescu, Răzvan V, Arman Eshaghi, Marco Lorenzi, Alexandra L Young, Neil P Oxtoby, Sara Garbarino, Sebastian J Crutch, Daniel C Alexander, and Alzheimer’s Disease Neuroimaging Initiative. 2019. “DIVE: A Spatiotemporal Progression Model of Brain Pathology in Neurodegenerative Disorders.” NeuroImage 192: 166–77.

Mårtensson, Gustav, Daniel Ferreira, Lena Cavallin, J-Sebastian Muehlboeck, Lars-Olof Wahlund, Chunliang Wang, Eric Westman, and Alzheimer’s Disease Neuroimaging Initiative. 2019. “AVRA: Automatic Visual Ratings of Atrophy from MRI Images Using Recurrent Convolutional Neural Networks.” NeuroImage: Clinical 23: 101872.

Mårtensson, Gustav, Daniel Ferreira, Tobias Granberg, Lena Cavallin, Ketil Oppedal, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Matteo Pardini, and Milica Kramberger. 2019. “The Reliability of a Deep Learning Model in Clinical Out-of-Distribution MRI Data: A Multicohort Study.” ArXiv Preprint ArXiv:1911.00515.

Muehlboeck, J, Eric Westman, and Andrew Simmons. 2014. “TheHiveDB Image Data Management and Analysis Framework.” Frontiers in Neuroinformatics 7: 49.
Mueller, Susanne G, Michael W Weiner, Leon J Thal, Ronald C Petersen, Clifford Jack, William Jagust, John Q Trojanowski, Arthur W Toga, and Laurel Beckett. 2005. “The Alzheimer’s Disease Neuroimaging Initiative.” *Neuroimaging Clinics* 15 (4): 869–77.

Murray, Melissa E, Neill R Graff-Radford, Owen A Ross, Ronald C Petersen, Ranjan Duara, and Dennis W Dickson. 2011. “Neuropathologically Defined Subtypes of Alzheimer’s Disease with Distinct Clinical Characteristics: A Retrospective Study.” *The Lancet Neurology* 10 (9): 785–96.

Na, Han Kyu, Dae Ryong Kang, Sungsoo Kim, Sang Won Seo, Kenneth M Heilman, Young Noh, and Duk L Na. 2016. “Malignant Progression in Parietal-Dominant Atrophy Subtype of Alzheimer’s Disease Occurs Independent of Onset Age.” *Neurobiology of Aging* 47: 149–56.

Noh, Young, Seun Jeon, Jong Min Lee, Sang Won Seo, Geon Ha Kim, Hanna Cho, Byoung Seok Ye, Cindy W Yoon, Hee Jin Kim, and Juhee Chin. 2014. “Anatomical Heterogeneity of Alzheimer Disease: Based on Cortical Thickness on MRIs.” *Neurology* 83 (21): 1936–44.

Oppedal, Ketil, Daniel Ferreira, Lena Cavallin, Afina W Lemstra, Mara Ten Kate, Alessandro Padovani, Irena Rektorova, Laura Bonanni, Lars-Olof Wahlund, and Knut Engedal. 2019. “A Signature Pattern of Cortical Atrophy in Dementia with Lewy Bodies: A Study on 333 Patients from the European DLB Consortium.” *Alzheimer’s & Dementia* 15 (3): 400–409.

Ossenkoppele, Rik, Chul Hyoung Lyoo, Carole H Sudre, Danielle van Westen, Hanna Cho, Young Hoon Ryu, Jae Yong Choi, Ruben Smith, Olof Strandberg, and Sebastian Palmqvist. 2019. “Distinct Tau PET Patterns in Atrophy-Defined Subtypes of Alzheimer’s Disease.” *Alzheimer’s & Dementia*.

Park, Jong-Yun, Han Kyu Na, Sungsoo Kim, Hyunwook Kim, Hee Jin Kim, Sang Won Seo, Duk L Na, Cheol E Han, and Joon-Kyung Seong. 2017. “Robust Identification of Alzheimer’s Disease Subtypes Based on Cortical Atrophy Patterns.” *Scientific Reports* 7.
Persson, Karin, Rannevig Sakshaug Eldholm, Maria Lage Barca, Lena Cavallin, Daniel Ferreira, Anne Brita Knapskog, Geir Selbæk, et al. 2017. “MRI-Assessed Atrophy Subtypes in Alzheimer’s Disease and the Cognitive Reserve Hypothesis.” *PLoS ONE* 12(10). https://doi.org/10.1371/journal.pone.0186595.

Poulakis, Konstantinos, Daniel Ferreira, Joana Braga Pereira, Orjan Smedby, Prashanthi Vemuri, and Eric Westman. 2019. “Fully Bayesian Longitudinal Unsupervised Learning for the Assessment and Visualization of AD Heterogeneity and Progression.” *BioRxiv*, 854356.

Poulakis, Konstantinos, Joana B Pereira, Patrizia Mecocci, Bruno Vellas, Magda Tsolaki, Iwona Kłoszewska, Hilkka Soininen, Simon Lovestone, Andrew Simmons, and Lars-Olof Wahlund. 2018. “Heterogeneous Patterns of Brain Atrophy in Alzheimer's Disease.” *Neurobiology of Aging* 65: 98–108.

Risacher, Shannon L, Wesley H Anderson, Arnaud Charil, Peter F Castelluccio, Sergey Shcherbinin, Andrew J Saykin, Adam J Schwarz, and Alzheimer's Disease Neuroimaging Initiative. 2017. “Alzheimer Disease Brain Atrophy Subtypes Are Associated with Cognition and Rate of Decline.” *Neurology* 89 (21): 2176–86.

ROUSSET, and OG. 1998. “Correction for Partial Volume Effects in PET: Principle and Validation.” *J Nucl Med.* 39: 904–11. http://ci.nii.ac.jp/naid/10025136344/en/.

Schöll, Michael, Samuel N Lockhart, Daniel R Schonhaut, James P O’Neil, Mustafa Janabi, Rik Ossenkoppele, Suzanne L Baker, Jacob W Vogel, Jamie Faria, and Henry D Schwimmer. 2016. “PET Imaging of Tau Deposition in the Aging Human Brain.” *Neuron* 89 (5): 971–82.

Shaw, Leslie M, Hugo Vanderstichele, Malgorzata Knapik-Czajka, Christopher M Clark, Paul S Aisen, Ronald C Petersen, Kaj Blennow, Holly Soares, Adam Simon, and Plotr Lewczuk. 2009. “Cerebrospinal Fluid Biomarker Signature in Alzheimer’s Disease Neuroimaging Initiative Subjects.” *Annals of Neurology* 65 (4): 403–13.
Varol, Erdem, Aristeidis Sotiras, Christos Davatzikos, and Alzheimer’s Disease Neuroimaging Initiative. 2017. “HYDRA: Revealing Heterogeneity of Imaging and Genetic Patterns through a Multiple Max-Margin Discriminative Analysis Framework.” *Neuroimage* 145: 346–64.

Whitwell, Jennifer L, Dennis W Dickson, Melissa E Murray, Stephen D Weigand, Nirubol Tosakulwong, Matthew L Senjem, David S Knopman, Bradley F Boeve, Joseph E Parisi, and Ronald C Petersen. 2012. “Neuroimaging Correlates of Pathologically Defined Subtypes of Alzheimer’s Disease: A Case-Control Study.” *The Lancet Neurology* 11 (10): 868–77.

Whitwell, Jennifer L, Jonathan Graff-Radford, Nirubol Tosakulwong, Stephen D Weigand, Mary Machulda, Matthew L Senjem, Christopher G Schwarz, Anthony J Spychalla, David T Jones, and Daniel A Drubach. 2018. “[18F] AV-1451 Clustering of Entorhinal and Cortical Uptake in Alzheimer’s Disease.” *Annals of Neurology* 83 (2): 248–57.

Young, Alexandra L, Razvan V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C Firth, David M Cash, David L Thomas, Katrina M Dick, and Jorge Cardoso. 2018. “Uncovering the Heterogeneity and Temporal Complexity of Neurodegenerative Diseases with Subtype and Stage Inference.” *Nature Communications* 9 (1): 1–16.

Zhang, Xiuming, Elizabeth C Mormino, Nanbo Sun, Reisa A Sperling, Mert R Sabuncu, B T Thomas Yeo, and Alzheimer’s Disease Neuroimaging Initiative. 2016. “Bayesian Model Reveals Latent Atrophy Factors with Dissociable Cognitive Trajectories in Alzheimer’s Disease.” *Proceedings of the National Academy of Sciences* 113 (42): E6535–44.
FIGURE LEGENDS

Figure 1. Graphical representation of the subtyping methods implemented in this study
aThis figure corresponds to the sMRI-based method. For the tau PET-based method, volume measures are replaced with SUVR and classification of LP and HS is reversed; bZH = z-score for hippocampus; Zf = z-score for frontal regions; Zp = z-score for parietal regions; Zt = z-score for temporal regions. This figure corresponds to the sMRI-based method and z-scores are computed for volumes. For the tau PET-based method, volume measures are replaced with SUVR and abnormal tau levels have z-scores \( \geq 1 \). Key: TAD = typical AD; HS = hippocampal-sparing; LP = limbic-predominant; MA = minimal atrophy; SUVR = standardized uptake value ratio; SD = standard deviation.

Figure 2. Demographic and clinical characteristics captured by the different subtyping methods
The bar plots for sex and APOE \( \varepsilon 4 \) show what percentage of patients in each subtype were females and APOE \( \varepsilon 4 \) carriers, respectively. Key: sMRI = structural magnetic resonance imaging; PET = positron emission tomography; F = female; APOE = apolipoprotein; MMSE = mini mental state exam; MA = minimal atrophy; MT = minimal tau; LP = limbic-predominant; HS = hippocampal-sparing; TAD = typical AD. Kruskal-Wallis hypothesis testing was conducted comparing the subtypes within each method; *p \( < 0.05 \) within the subtyping method.

Figure 3. Group-level cortical thickness maps across subtyping methods in the sMRI-tauPET cohort
For simplicity, only left lateral and medial views are presented since very similar results were obtained for the right lateral and medial views. Differences in cortical thickness maps are shown in each subtype relative to HC, generated by fitting general linear model at each vertex. Yellow-red regions reflect thinner cortex in AD subtypes relative to HC. All brain maps are uncorrected for multiple comparisons at \( p < 0.01 \). Risacher et al. identified three subtypes only and hence, there are no cortical maps corresponding to MA subtype. Poulakis et al., identified all four subtypes. However, the HS subtype (1 individual) had to be excluded from the study due to invalid tau...
PET data. **Key:** TAD=typical AD; HS=hippocampal-sparing; LP=limbic-predominant; MA= minimal atrophy; HC=healthy control.

**Figure 4. Group-level PVC tau PET uptake maps across subtyping methods using the hippocampus in the sMRI-tauPET cohort**

For simplicity, only left lateral and medial views are presented since very similar results were obtained for the right lateral and medial views. Differences in tau PET uptake maps are shown in each subtype relative to HC, generated by fitting general linear model at each vertex. Cyan regions reflect greater PVC tau PET uptake in AD subtypes relative to HC. All brain maps are uncorrected for multiple comparisons at $p < 0.01$. **Key:** TAD=typical AD; HS=hippocampal-sparing; LP=limbic-predominant; MT=minimal tau; HC=healthy control.

**Figure 5. Group-level PVC tau PET uptake maps across subtyping methods using the entorhinal cortex in the sMRI-tauPET cohort**

For simplicity, only left lateral and medial views are presented since very similar results were obtained for the right lateral and medial views. Differences in tau PET uptake maps are shown in each subtype relative to HC, generated by fitting general linear model at each vertex. Blue-cyan regions reflect PVC greater tau PET uptake in AD subtypes relative to HC. All brain maps are uncorrected for multiple comparisons at $p < 0.01$. **Key:** TAD=typical AD; HS=hippocampal-sparing; LP=limbic-predominant; MT=minimal tau; HC=healthy control.

**Figure 6. Individual-level agreement among subtyping methods as illustrated by Cohen’s kappa values**

**Key:** sMRI=structural magnetic resonance imaging; PET=positron emission tomography; MA=minimal atrophy AD; MT=minimal tau.
Table 1. Overview of the subtyping methods implemented in this study

| Method       | Type of method | Definition of subtypes      | Modality     | Measure   | Subtypes | Graphical Representation |
|--------------|----------------|------------------------------|--------------|-----------|----------|--------------------------|
| Charil (25)  | Hypothesis-driven | Within-sample dependent    | tau PET    | SUVR      | TAD, LP, HS | Figure 1A               |
| Risacher (6) | Hypothesis-driven | Within-sample dependent    | sMRI and tau PET | Automated volumes and SUVR | TAD, LP, HS | Figure 1B               |
| Byun (5)     | Hypothesis-driven | External reference group   | sMRI and tau PET | Automated volumes and SUVR | TAD, LP, HS, MA* | Figure 1C               |
| Ferreira (7) | Hypothesis-driven | External reference group   | sMRI        | Visual ratings | TAD, LP, HS, MA | Figure 1D               |
| Poulakis (15)| Data-driven     | Within-sample dependent    | sMRI        | Automated volumes | TAD, LP, HS, MA | Figure 1E               |

Key: TAD=typical AD; HS=hippocampal-sparing; LP=limbic-predominant; MA=minimal atrophy; SUVR=standardized uptake value ratio; SD=standard deviation. *MA corresponds to the subtype identified by the sMRI-based method. For the tau PET-based method, the corresponding subtype would be minimal tau. †The two clusters reflecting typical AD patterns in the original publication by Poulakis et al., (15) were combined into a single typical AD subtype to allow comparisons across subtyping methods.
## Table 2. Demographic and clinical characteristics of the cohorts

(a) Validation of subtyping methods in AD dementia patients (sMRI cohort)

|                      | HC (Aβ−) | AD dementia (Aβ+) | p-value |
|----------------------|-----------|-------------------|---------|
| N                    | 70        | 89                | -       |
| Sex (F,%)            | 51        | 39                | 0.139   |
| Age (years)          | 75.15 ± 5.22 [62, 89] | 74.73 ± 7.72 [57, 88] | 0.757   |
| Education (years)    | 15.66 ± 2.65 [8, 20] | 15.16 ± 3.24 [4, 20] | 0.402   |
| APOE ε4 carriers (%) | 10        | 74                | <.0001  |
| MMSE                 | 29.04 ± 1.10 [25, 30] | 23.48 ± 1.87 [20, 26] | <.0001  |
| Word recall task     | 2.86 ± 1.17 [0, 5.67] | 6.24 ± 1.34 [3.33, 9.33] | <0.0001 |
| Naming objects and fingers | 0.08 ± 0.28 [0, 1] | 0.43 ± 0.56 [0, 2] | <0.0001 |
| Following commands   | 0.05 ± 0.23 [0, 1] | 0.40 ± 0.63 [0, 3] | <0.0001 |
| Constructional praxis| 0.41 ± 0.49 [0, 1] | 0.91 ± 0.65 [0, 3] | <0.0001 |
| Ideational praxis    | 0.05 ± 0.23 [0, 1] | 0.33 ± 0.72 [0, 5] | 0.0007  |
| Orientation          | 0.12 ± 0.37 [0, 2] | 2.06 ± 1.61 [0, 7] | <0.0001 |
| MMSE                 | 2.38 ± 2.24 [0, 12] | 6.51 ± 2.88 [1, 2] | <0.0001 |
| Recall of test instructions | 0 ± 0 [0, 0] | 0.31 ± 0.88 [0, 5] | 0.0005  |
| Spoken language      | 0.02 ± 0.16 [0, 1] | 0.30 ± 0.64 [0, 3] | 0.0006  |
| Word finding difficulty | 0.04 ± 0.20 [0, 1] | 0.59 ± 0.95 [0, 4] | <0.0001 |
| ADAS Total Score     | 6.06 ± 2.79 [1,67, 14.33] | 18.38 ± 6.26 [8.67, 42.67] | <0.0001 |

(b) Subtyping methods in Prodromal AD and AD dementia patients (sMRI-tauPET cohort)

|                      | HC (Aβ−) | Prodromal AD (Aβ+) | AD dementia (Aβ+) | p-value |
|----------------------|-----------|-------------------|-------------------|---------|
| N                    | 200       | 54                | 30                | -       |
| Sex (F,%)            | 59        | 48                | 50                | 0.285   |
| Age (years)          | 70.45 ± 5.65 [55.8, 89] | 74.09 ± 7.34 [59.4, 90.1] | 77.46 ± 8.27 [55.9, 91.2] | <.0001  |
| Education (years)    | 16.90 ± 2.31 [11, 20] | 15.76 ± 2.66 [12, 20] | 15.77 ± 2.57 [12, 20] | 0.002   |
| APOE ε4 carriers (%) | 22        | 61                | 53                | <.0001  |
| MMSE                 | 29.24 ± 1.05 [23, 30] | 27.48 ± 2.30 [19, 30] | 22.13 ± 4.23 [9, 30] | <.0001  |
| Word recall task     | 2.36 ± 1.81 [0] | 4.34 ± 1.49 [1, 7] | 6.09 ± 1.61 [3, 10] | <0.0001 |
| Activity                        | Mean ± SD [Min, Max] | p value |
|--------------------------------|-----------------------|---------|
| Naming objects and fingers     | 0.03 ± 0.37 [0, 3]    | 0.56 ± 0.89 [0, 3] | <0.0001 |
| Following commands             | 0.06 ± 0.24 [0, 1]    | 0.40 ± 0.91 [0, 3] | 0.0011 |
| Constructional praxis          | 0.33 ± 0.55 [0, 3]    | 0.72 ± 0.79 [0, 3] | 0.0084 |
| Ideational praxis              | 0.05 ± 0.38 [0, 5]    | 0.28 ± 0.54 [0, 2] | <0.0001 |
| Orientation                    | 0.09 ± 0.29 [0, 1]    | 2.8 ± 2.1 [0, 7]  | <0.0001 |
| Word recognition task          | 4.92 ± 1.91 [0, 10]   | 5.9 ± 2.77 [0, 12] | <0.0001 |
| Recall of test instructions    | 0.005 ± 0.07 [0, 1]   | 0.95 ± 1.39 [0, 5] | <0.0001 |
| Spoken language                | 0.005 ± 0.07 [0, 1]   | 0.44 ± 1.00 [0, 4] | <0.0001 |
| Word finding difficulty        | 0.04 ± 0.27 [0, 3]    | 0.88 ± 1.05 [0, 3] | <0.0001 |
| ADAS Total Score               | 8.08 ± 2.86 [1, 19]   | 11.98 ± 4.66 [3, 24] | 21.68 ± 7.48 [7, 37] | <0.0001 |

Data are reported as mean ± standard deviation [minimum, maximum]; Hypothesis testing was performed using the Kruskal–Wallis test for the continuous variables and χ² test for the nominal variables. Additionally, the Kruskal–Wallis test was performed pairwise between groups in the sMRI-tauPET cohort. Key: HC=healthy control; AD=Alzheimer’s disease; sMRI=structural MRI; PET=positron emission tomography; Aβ=amyloid-beta; F=female; MMSE=Mini-Mental State Examination; APOE=apolipoprotein; ADAS=Alzheimer’s Disease Assessment Scale-Cognitive Subscale. *significantly different from each of the other two groups; **significantly different from HC group only.
Table 3. Frequencies of the subtypes compared with previously published studies in the sMRI cohort

| Subtype              | Risacher Pub. | Byun Pub. | Ferreira Pub. | Poulakis Pub.* |
|----------------------|---------------|-----------|---------------|----------------|
|                      | This Study    | This Study| This Study    | This Study     |
| Typical AD           | 69            | 69        | 59            | 55             | 51 | 52 | 69 | 66 |
| Hippocampal-sparing  | 17            | 19        | 12            | 17             | 17 | 19 | 7  | 10 |
| Limbic-predominant   | 14            | 12        | 19            | 21             | 17 | 18 | 4  | 1  |
| Minimal atrophy      | -             | -         | 10            | 7              | 15 | 11 | 19 | 23 |

Data are reported as % and rounded to the nearest integer for readability. **Key:** AD=Alzheimer’s Disease; Pub.=published study. *Frequencies of subtypes based on the ADNI cohort only, since the original study by Poulakis et al. (15) also includes another cohort.
Figure 1
Figure 2

304x457mm (300 x 300 DPI)
Figure 3
Figure 4
Figure 5
Figure 6
121x92mm (300 x 300 DPI)
Graphical Abstract
ABBREVIATED SUMMARY

Neuroimaging-based subtyping methods in Alzheimer’s disease seem largely comparable at the group-level. However, Mohanty et al. found large disagreements across subtyping methods at the individual level. They call for consensus and harmonization across subtyping methods and propose using an open benchmarking framework to facilitate better agreement among future subtyping methods.