VirtualBrainCloud

Personalized Recommendations for Neurodegenerative Disease

www.VirtualBrainCloud-2020.eu

Public deliverable report

D4.1 Paper Manuscript describing the initial characterization of the Studies, the equivalence of their variables and the summary statistics (mean, standard deviation, etc.) submitted to a relevant journal

Date	May 2020
Authors	Fraunhofer SCAI (Colin Birkenbihl, Holger Frölich, Martin Hofmann-Apitius)
Dissemination level	**public**
Website	www.VirtualBrainCloud-2020.eu

This project has received funding from the **European Union´s Horizon 2020** research and innovation programme under **grant agreement No 826421**

Evaluating the Alzheimer's Disease Data Landscape

Colin Birkenbihl[1,2], Yasamin Salimi[1,2], Daniel Domingo-Fernández[1,2], Simon Lovestone[3] on behalf of the AddNeuroMed consortium, Holger Fröhlich[1,2], Martin Hofmann-Apitius[1,2], and the Japanese Alzheimer's Disease Neuroimaging Initiative[*], and the Alzheimer's Disease Neuroimaging Initiative[†]

1. Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53754, Germany

2. Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany

3. Department of Psychiatry, University of Oxford, Oxford, UK

Key Words: Cohort Studies, Dementia, Datasets, Sample Bias, Comparison, Metadata, FAIR, Interoperability

*Data used in preparation of this article were obtained from the Japanese Alzheimer's Disease Neuroimaging Initiative (J-ADNI) database deposited in the National Bioscience Database Center Human Database, Japan (Research ID: hum0043.v1, 2016). As such, the investigators within J-ADNI contributed to the design and implementation of J-ADNI and/or provided data but did not participate in analysis or writing of this report.A complete listing of J-ADNI investigators can be found at: https://humandbs.biosciencedbc.jp/en/hum0043-j-adni-authors.

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

Abstract:

INTRODUCTION: Numerous studies have collected AD cohort datasets. To achieve reproducible, robust results in data-driven approaches, an evaluation of the present data landscape is vital.

METHODS: Previous efforts relied exclusively on metadata and literature. Here, we evaluate the data landscape by directly investigating nine patient-level datasets generated in major clinical cohort studies.

RESULTS: The investigated cohorts differ in key characteristics, such as demographics and distributions of AD biomarkers. Analyzing the ethnoracial diversity revealed a strong bias towards white/caucasian individuals. We describe the measured data modalities and compare them across cohorts. Finally, the available longitudinal data for important AD biomarkers is evaluated. All results are explorable at https://adata.scai.fraunhofer.de

DISCUSSION: Our evaluation exposed critical limitations in the AD data landscape that impede comparative approaches across multiple datasets. Comparing our results to those gained by metadata-based approaches highlights that thorough investigation of real patient-level data is inevitable to assess a data landscape.

1. Background:

In the field of Alzheimer's disease (AD) research, numerous cohort studies have been conducted and their collected data build the basis for a plethora of research projects. However, each of these studies only reflect patients of a particular subpopulation defined by inclusion and exclusion criteria. This is becoming

specifically relevant in the context of the increasing popularity of data-driven approaches and machine learning [1]: After analyzing a single cohort, it is mandatory to demonstrate that results are reproducible in independent, external data originating from distinct cohort studies. Furthermore, it is essential to conduct comparative analyses across datasets in order to assess whether the observed patterns are robust [2]. Such systematic data-driven approaches are, however, hampered by the fact that patient-level data is often difficult to access or entirely inaccessible. Moreover, we have limited knowledge about how the distinct cohort datasets available in our field compare to each other on a qualitative (e.g. overlap of measured variables) as well as quantitative level (e.g. values encountered in the data) [3,4]. Thus, in order to leverage the full potential of collected patient-level data, it is important to characterize the clinical AD data landscape in detail.

Evaluating a data landscape involves organizing and comparing datasets in order to: 1) qualitatively assess their collected data modalities and variables, and 2) quantitatively describe the demographics of the study population and distributions of measured variables. Such characterization provides a detailed overview of the data accessibility and supports the design of research projects and future cohort studies. Finally, evaluating a data landscape inherently exposes potential flaws with regard to interoperability between existing datasets and underrepresentation of important disease or population characteristics.

In the AD field, previous studies have attempted to establish a comprehensive view of the AD data landscape as well as demonstrate how cohort datasets relate to each other. For example, the European Medical Information Framework (EMIF), collected metadata of AD cohort studies by providing data owners with a questionnaire in which they could specify the variables contained in their datasets. The resulting

metadata are presented through the EMIF-Catalog [5]. Similarly, the ROADMAP project generated an overview of clinical outcomes and data modalities that were collected in several European AD cohort studies [6]. By analyzing metadata (partially originating from the EMIF-Catalog), ROADMAP created the ROADMAP Data Cube, a web application that shows the availability of AD related outcomes in a selected set of European dementia cohorts (https://datacube.roadmap-alzheimer.org). Lawrence *et al.*, on the other hand, opted for a literature-based approach to assess the AD data landscape. The authors reviewed publications corresponding to AD cohort datasets and gathered the contained information [7].

All of the above-mentioned undertakings attempted to evaluate the AD data landscape solely on the basis of metadata and literature, without investigating the underlying patient-level data. However, reviewing study protocols can only explain the original design of a given study and thereby neglects unforeseen changes in procedures or participant recruitment throughout study runtime. The alternative approach is a patient-level and data-driven evaluation of the AD data landscape, which is a tedious and time-consuming endeavor. The first hurdle of such an approach is gaining access to a sufficient number of cohort datasets. Data access typically requires completing an application procedure with numerous legal requirements and considerations. If access is granted, intensive manual curation and investigation of data follows. Although difficult to establish, a comprehensive data-driven view on the AD data landscape is crucial, since reliance exclusively on metadata assumes that these metadata correctly describe the underlying dataset and that this dataset is complete. In contrast, a patient-level and data-driven evaluation 1) is not subject to these assumptions, 2) allows for a quantitative investigation of important cohort statistics and 3) illustrates the amount and quality of

the data accessible to the field.

In this work, we traced down, accessed, investigated, and compared nine of the major clinical cohort study datasets available in the AD field. Instead of solely relying on metadata and / or literature, we assessed the current AD data landscape by curating and investigating the accessible patient-level cohort datasets. Here, we comprehensively describe the acquired data and show which data modalities we found in the datasets as well as their overlap with other studies. Additionally, we assessed the longitudinal follow-up on biomarker-level and demonstrated to what extent current AD data is covering the progression of the disease. Furthermore, we compared the content we observed in these datasets with the reported findings of metadata-based approaches [5,7]. Finally, we made all results available through an interactive web-portal (https://adata.scai.fraunhofer.de), such that researchers can explore the AD data landscape generated on our investigated datasets.

2. Methods:

2.1 Investigated Cohorts

We aimed to acquire as many major AD cohort studies as possible to allow for a thorough investigation of the data landscape. We only considered datasets that were downloadable, hereby excluding data portals with restricted data access from our investigations. Most of the datasets we accessed were shared after completing an official data request process. We applied for access to 18 distinct AD cohort datasets. Until submitting this work for publication, we were granted access to nine (**Table 1**).

Table 1. The investigated AD cohorts and their references.

Cohort	Consortium	Reference
A4	Anti-Amyloid Treatment in Asymptomatic Alzheimer’s Disease	[8]
ADNI	The Alzheimer’s Disease Neuroimaging Initiative	[9]
ANMerge†	AddNeuroMed	[10]
AIBL	The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing	[11]
EMIF-1000	European Medical Information Framework	[12]
EPAD v1500	European Prevention of Alzheimer's Dementia	[13]
JADNI	Japanese Alzheimer’s Disease Neuroimaging Initiative	[14]
NACC	The National Alzheimer's Coordinating Center	[15]
ROSMAP	The Religious Orders Study and Memory and Aging Project	[16]

NOTE: †: ANMerge is a new version of the AddNeuroMed dataset pending publication. It includes the Maudsley BRC Dementia Case Registry at King’s Health Partners cohort and the Alzheimer’s Research Trust UK cohort. [17]

It is important to be aware that not all of these studies followed the same design nor goals. Each study enforced its own recruitment criteria and enrolled participants following distinct selection processes. While some aimed for a case-control setting and included a substantial amount of AD patients into their cohort, others deliberately excluded them to focus on early disease progression. Thereby, the cohort datasets are all subject to inherent biases.

2.2 Generating the Summary Statistics

To illustrate the content of the datasets, we characterized the demographics of each cohort and described the encountered statistical distributions of important AD biomarkers. The demographic variables we considered are: participant age, sex, and

completed years of education. Additionally, we assessed the diversity of ethnoracial groups in our acquired AD cohorts, since it is known that ethnoracial factors may impact AD and related findings [18]. The AD biomarkers we compared between cohorts are motivated in the **Supplementary Text**.

For numerical variables, we describe the encountered distributions using the 25%, 50%, and 75% quantiles of the raw measurements. For categorical ones, we describe the proportion of study participants falling into its respective categories. In some datasets, single variables were only reported numerically given they placed within a defined value range (e.g. 400-1700). If the measurement appeared to be outside of this range, the exact number was not reported but replaced with a cut-off (e.g. “>1700”). To allow for calculations, we considered these values to be equal to the mentioned cut-off (here, 1700).

2.3 Generating the Data Availability Map

While establishing a data landscape, it is of high interest to identify the data modalities that were measured in the underlying studies as well as to compare their overlaps. However, assessing the availability of data modalities in clinical cohort datasets is not straightforward. It involves intensive and meticulous manual curation of the acquired datasets and thereby, the definition of applicable curation criteria specifying under which circumstances each data modality is considered as “available”. Furthermore, it is often necessary to define a gradual categorization to represent the degree of availability. For example, exclusively measuring two specific single nucleotide polymorphisms (SNP), is not equal to conducting a genome-wide genotyping of individuals. Similarly, distributing only normalized brain volumes summed over both hemispheres holds less information than providing the underlying

raw MRI images. The latter would enable researchers to process the images according to their needs, while the former impedes interoperability to other datasets due to differences in employed image processing pipelines. This could hamper conducting certain analyses like systematic comparisons across cohorts or validation approaches.

To enable a meaningful comparable assessment of the availability of data modalities, we established criteria for categorizing the availability of each modality into three discrete stages **(Supplementary Table 1)**: stage 0) no data were available for the respective modality, stage 1) data were partially available, and stage 2) more complete data or unprocessed raw data were available.

2.4 Investigating Longitudinal Follow-Up Across Studies

To assess how far existing cohort datasets cover the important time dimension of AD, we conducted a thorough investigation of the longitudinal follow-up performed in the acquired studies. For each cohort, we evaluated how many participants were assessed at each follow-up visit and implicitly analyzed the subsequent drop-out over study runtime. Since not all measurements were performed at each visit and not every individual participated in all sample collections, we further focused on the follow-up and coverage of important AD biomarkers. Determining the amount of available longitudinal data per biomarker provides an estimate on how much information we can exploit in order to model, and ultimately understand patterns of disease progression.

3. Results:

3.1 Investigation of the AD Data Landscape

Altogether, we investigated data from nine studies comprising a total of 60004 assessed study participants. **Table 2** shows how these participants were distributed among the analyzed cohorts. With NACC being the exception (n = 40858), all studies recruited individuals in the low thousands (n = ~1200 to 3600). According to their diagnosis, participants could be separated into three groups: cognitively healthy controls, mild cognitive impaired (MCI) patients, and AD patients. To file such a diagnosis, most studies applied the NINCDS-ADRDA diagnostic criteria [19]. The fact that diagnosis criteria are aligned across most datasets significantly increases the interoperability between them, since AD follows the same semantic description in context of these studies. Depending on each study's goals, the recruitment process focused on enrolling more or less individuals falling into any of these diagnosis groups.

While no data is shared through our web-portal, information on how to access these datasets can be found at https://adata.scai.fraunhofer.de/cohorts.

Table 2: Description of the investigated cohorts.

Cohorts	N	CTL	MCI	AD	N with 2+ visits	Follow-up Interval (months)	Location	Diagnostic criteria AD
A4	6943	6943	0	0	0*	~8	USA, Canada,	excluded AD†
ADNI	2241	516	1022	384	1978 (88%)	6	USA, Canada	NINCDS-ADRDA
AIBL	1378	803	134	181	1019 (74%)	18	Australia	NINCDS-ADRDA
ANM	1702	793	397	512	1254 (74%)	12	Europe	NINCDS-ADRDA
EMIF	1221	386	526	201	0	no follow-up	Europe	NINCDS-ADRDA
EPAD v1500	1500	1410	80	3	0*	6	Europe	NINCDS-ADRDA
JADNI	537	151	233	149	518	6	Japan	NINCDS-ADRDA
NACC	40858	15894	3649	11761	27657 (68%)	12	USA	UDS Form D1
ROSMAP	3627	2514	898	203	3335 (92%)	12	USA	NINCDS-ADRDA

NOTE: Number of diagnosed subjects does not always add up to N, since patients with different dementia diagnoses (e.g. Lewy-bodies or frontotemporal dementia) were excluded. **N:** Total number of participants. **CTL/MCI/AD:** Number of participants with the respective diagnosis at study baseline. **2+ visits:** Number of study participants for whom data for at least two time points is available. **Follow-up Interval:** Approximated regular time interval between participant visits. ***:** Longitudinal data has been collected but is not released yet. **†:** Recruited only cognitively healthy participants.

3.2 Characterization of the Cohorts

Investigation of the cohort demographics revealed considerable differences between key demographic characteristics of the acquired cohorts. EPAD, for example, recruited a comparably young and primarily non-symptomatic cohort, while participants of AddNeuroMed and ROSMAP were significantly older (**Table 3**). Across all cohorts, the age range spans roughly from 60 (lowest 25% quantile) to 85 years (highest 75% quantile). Theoretically, this opens the opportunity to construct a pseudo-continuum of 25 years of disease history. Furthermore, in most studies we observed the general tendency that more female than male participants enrolled into the studies. ROSMAP illustrates an extreme case, recruiting predominantly nuns from religious orders, explaining the high number of female study participants (72.8%). Overall, most individuals included in the AD cohort studies were highly educated (~ 14 years on average). As has been pointed out previously by Whitwell *et al.*, a high level of education can act as cognitive reserve possibly concealing a prodromal manifestation of AD. Numerous demographic differences found between studies may result from distinct recruitment criteria which, again, mirror the individual study goals. While differences in recruitment criteria lead to a broader sampling of the AD population, they reduce the direct comparability between datasets because they inevitably introduce bias into the data. One key example is recruitment specifically for participants with AD risk factors (e.g. APOE genotype). This could significantly bias the patterns exhibited in the data in comparison to another dataset with lower amount of APOE ε4 positive participants.

Table 3: Distribution of demographic variables and key AD biomarkers encountered in each cohort.

	Female %	Age	Education	APOE e4 %	MMSE	CDR	CDR-SB	Hippocampus	A-beta	tTau	pTau
A4	57.7	68, 71, 75	14, 16, 18	34.3	28, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	6, 7, 7			
ADNI	47	68, 73, 78	14, 16, 18	45.6	26, 28, 29	0.0, 0.5, 0.5	0.0, 1.0, 2.0	5948, 6864, 7651	596, 854, 1396	193, 258, 350	17, 24, 34
AIBL	57.9	67, 73, 79	10, 12, 15	36	26, 28, 30	0.0, 0.0, 0.5	0.0, 0.0, 1.0	3, 3, 3	445, 567, 802	238, 366, 516	43, 64, 81
ANM	59.3	71, 77, 81	8, 11, 14	38.8	24, 28, 29	0.0, 0.5, 0.5	0.0, 0.5, 4.0	5311, 6270, 7142			
EMIF	46.2	62, 68, 74	9, 12, 15	46.8	25, 28, 29	0.5, 0.5, 0.5		6357, 7223, 8004	385, 525, 739	160, 278, 504	37, 52, 74
EPAD	56.9	60, 66, 71	12, 15, 17	37.7	28, 29, 30	0.0, 0.0, 0.0	0.0, 0.0, 0.0	4413, 4808, 5182	899, 1319, 1700	162, 201, 252	13, 17, 22
JADNI	52.7	66, 72, 77	12, 12, 16	46.1	24, 26, 29	0.0, 0.5, 0.5	0.0, 1.5, 3.0	5260, 6133, 7132 (1, 7, 40)	254, 315, 454	67, 101, 138	36, 48, 73
NACC	57.2	65, 72, 79	12, 16, 18	40.6	23, 27, 29	0.0, 0.5, 0.5	0.0, 1.0, 4.0	43.5%*	46.5%*	43.9%*	43.9%*
ROSMAP	72.8	73, 79, 84	14, 16, 18	25.1	27, 29, 30						

NOTE: Shown are the 25%, 50% and 75% quantile of numerical variables at baseline. Categorical variables are given as proportion of participants falling into one respective category. **APOE e4 %:** Proportion of participants with at least one APOE e4 allele. **Hippocampus:** Hippocampal volume. **A-beta, tTau, pTau:** Collected from CSF samples. ***:** NACC values are given as proportion of “abnormal observations”.

To further highlight one potential bias in AD data, we analyzed the ethnoracial diversity encountered in the investigated AD cohorts (**Figure 1**). An aggregated analysis of all acquired datasets demonstrates that the vast majority of these recruited individuals come from a white/caucasian background (83.9%). The second largest group were black/African descendents with 11.6%, followed by participants of Asian heritage with 2.3%. Here, we would like to point out that these findings are heavily influenced by the study location and the number of enrolled participants per study. Since the majority of the studies have been conducted in the USA, their locally exhibited ethnoracial diversity overshadows signals from European cohorts. However, the analogous plots for each European cohort show not only a similar, but even more extreme tendency towards white/caucasian individuals (EPAD: 99% white; AddNeuroMed: 98,5% white; see https://adata.scai.fraunhofer.de/ethnicity).

As expected, the ethnoracial composition in the investigated cohorts heavily relies on the diversity of populations from which the participants have been recruited. Nonetheless, our results elucidate that there is a tremendous bias towards white/caucasian in AD datasets and a severe underrepresentation of other ethnoracial groups, which, in turn, could be problematic for developing personalized treatments.

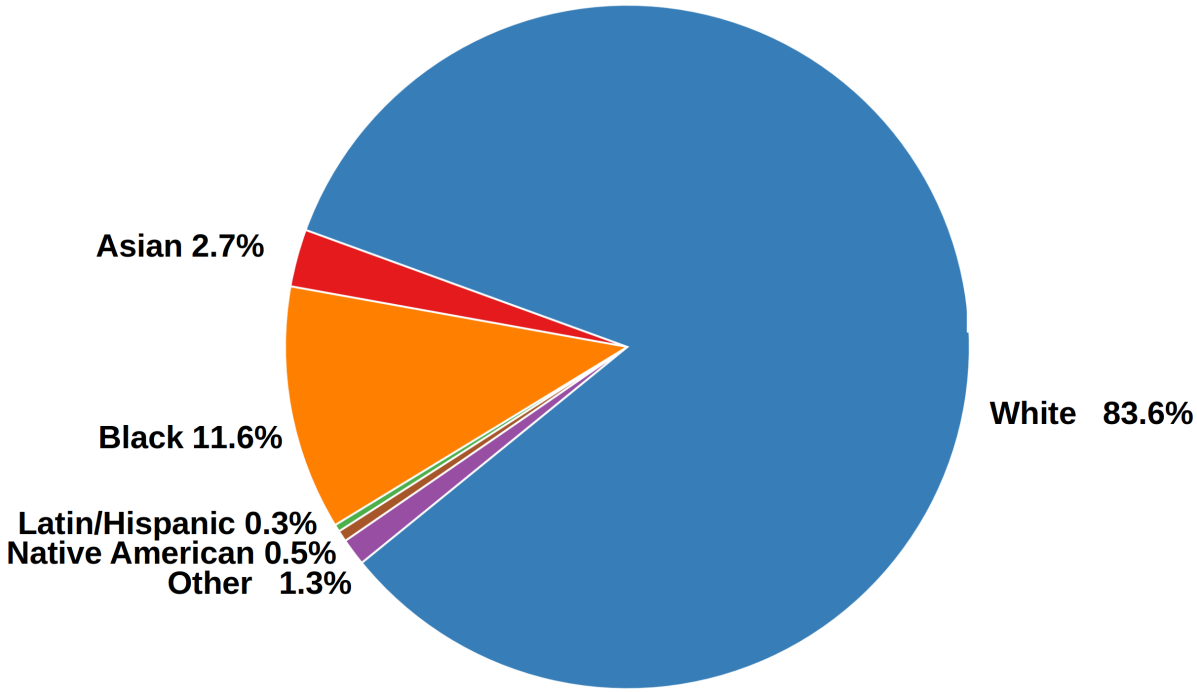

Figure 1: Combined ethnoracial diversity found across the investigated AD cohorts.

3.3 Availability of Data Modalities

The selection and amount of data modalities measured in a cohort study often depend on the study's aims and available funding. Thus, often only partially overlapping sets of data modalities are assessed in distinct cohort studies. To analyze, which data is available in our investigated cohorts and to explore the overlap between them, we assessed the grade of availability per data modality. Therefore, we manually curated each dataset according to our previously described criteria (**Supplementary Table 1**).

In **Figure 2,** we show an overview of the data modalities and their availability score in all acquired cohort datasets. Commonly assessed throughout all studies were demographic variables (e.g., participant age, sex, and education), as well as clinical assessments, (e.g., MMSE). In these two modalities, almost all studies were assigned the availability score 2. CSF biomarker measurements were found present in all datasets but AddNeuroMed, in which no CSF samples were taken. With regard

to post-mortem / autopsy data, ROSMAP contains a detailed collection of post-mortem data, ranging from simple measurements such as brain weight to comprehensive brain proteomics and transcriptomics. Although numerous studies conducted structural MRI, the data shared with us were sometimes limited to processed MRI features (e.g. brain volumes). In our case, only ADNI and AddNeuroMed granted access to the raw images. However, we would like to note that EPAD aims to provide raw images when the necessary infrastructure for image data distribution has been set up.

Although the purpose of this section is to provide a comprehensive overview on the availability of data per modality, we would like to emphasize that the presented results of this analysis are strongly dependent on our defined curation criteria, and different criteria could lead to different results. Additionally, all investigated datasets could potentially hold more information than we presented here. Due to our premise of exclusively looking into those patient-level data that have been shared with us, it is possible that we missed modalities or resources which were not shared in the first place (raw MRI images being an example). The results can be explored at https://adata.scai.fraunhofer.de/modality.

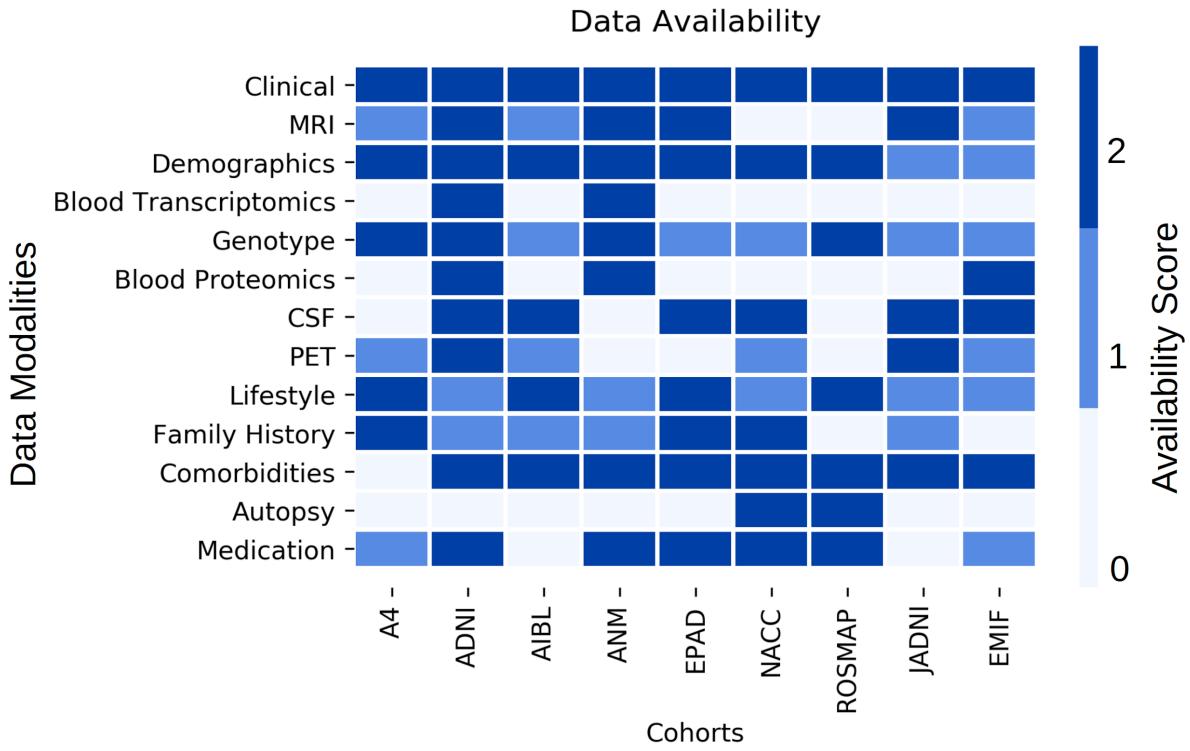

Figure 2: Availability of data modalities scored based on the defined criteria. The criteria are explained in **Supplementary Table 1**. **PET:** Positron emission tomography. **Blood Transcrip.:** Transcriptomic data gathered from blood samples.

To establish how our observations of data availability differed from results gained by relying solely on metadata, we compared our findings to the metadata presented in the EMIF catalog [5]. Only four of our investigated studies were listed[1]: ADNI, AddNeuroMed, EMIF, and EPAD. Although the majority of our findings are in concordance with the EMIF-catalog, deviations between metadata and the real data exist. We encountered variables in the datasets which are reported as absent in the catalog (e.g. Global Deterioration Scale in AddNeuroMed), or were not listed at all. Other variables and even modalities are reported to be present, yet could not be found in the respective dataset. For instance, the catalog states that post-mortem brain autopsy was performed in AddNeuroMed, which we could not find any evidence for.

Similar observations were made when comparing our findings with the review by

1 Accessed on 2th of February 2020

Lawrence *et al.* [7]. Here, for example, the reported longitudinal follow-up of AddNeuroMed is significantly shorter than we observed in the data (i.e., 12 months versus the 84 months in the data). Additionally, the number of participants with at least two visits is not concordant with the data we obtained (i.e., 378 versus the 1254 participants in the data).

These results show that there are two types of contradictions between the metadata assessments and our data-driven investigation: type 1 describes variables in the datasets which were reported to be missing according to the metadata sources. From this type of contradiction, we can conclude that approaches relying solely on metadata and literature potentially suffer in accuracy when estimating the real content available in cohort datasets. Contradiction type 2 is that metadata sources reported a variable to be present, while we were not able to find it in the underlying data. Type 2 contradictions do not lead to the same conclusion as type 1, since it may be possible that the respective variables have simply not been shared with us. However, it is arguable how practical correct metadata is if the data it describes is not available itself. We believe that the presented results and their conclusions highlight the importance of data access and curation when assessing a data landscape.

3.4 Disease Manifestation across Cohorts

To evaluate how severely patients from each cohort have been affected by AD, we compared the distributions of both cognitive outcomes and key biomarkers for the cognitively affected patient subgroups (i.e. participants with MCI or AD diagnosis). **Table 3** shows the distributions for each complete cohort including healthy controls, MCI and AD patients. Analogous tables per diagnosis subgroup can be found at

https://adata.scai.fraunhofer.de/cohorts.

According to the MMSE scores, AD patients from AIBL (Quantiles: 15, 20, 25), AddNeuroMed (Quantiles: 16, 21, 25) and NACC (Quantiles: 16, 21, 25) showed the worst cognitive performance. ADNI (Quantiles: 21, 23, 25) contained patients with fewer cognitive symptoms. The CDR sum of boxes scores (CDR-SB) slightly shifts the perspective. Here, AddNeuroMed is the most affected cohort with its 25%, 50% and 75% quantiles of the CDR-SB scores being 4, 6 and 9 respectively. AIBL patients scored 3.5, 5, 7, which slightly contradicts the image painted by the MMSE scores. Again, ADNI shows the least cognitive symptoms with its CDR-SB quantiles being 3, 4.5, 5.

A comparison of raw biomarker measurements between cohorts proved to be impossible, since encountered values are on different scales and may be subject to batch effects. Thus, we analyzed how much measurements diverged from their respective control population in each cohort (**Supplementary Text).**

The prerequisite for comparative approaches involving biomarker measurements across datasets is an alignment of their underlying data models (i.e. making data interoperable). In our analysis, each study had defined its own data model and variable names differed between them. This forced us to individually map variables to their corresponding counterparts in other studies to enable comparisons in the first place (e.g. combine “lh_hippo_volume” and “lh_hippo_volume” and map to “Hippocampus”). Another difficulty is that numerous datasets reported values of equivalent variables in different ways. For example, CSF biomarker measurements are reported to be either normal (0) or abnormal (1) in NACC, while other studies provide numerical values, which themselves were capped at different thresholds in

some of those studies (e.g. “>1700”). All these factors led to a severe lack of interoperability between datasets which significantly limits comparative approaches and restricts them to more standardized variables like clinical assessment scores.

3.5 Longitudinal Follow-up

The majority of the investigated studies have collected longitudinal data in the form of repeated measurements. The intervals of data collection differed across studies **(Table 2)**. **Figure 3A** displays the drop-out of study participants over time relative to the size of the cohort. In this analysis, participants were considered if at least one measurement was taken at the respective month. However, an individual’s participation in some assessments does not imply that all biomarkers values were acquired for the same individual on all visits. Thus, we additionally investigated the amount of study participants for which single important AD biomarkers were measured over time **(Figure 3)**. Plots for all of the investigated biomarkers can be found at https://adata.scai.fraunhofer.de/follow-up.

One example is CSF amyloid beta for which **Figure 3B** displays the longitudinal coverage**.** Comparing **Figure 3B** with **Figure 3A** demonstrates that CSF samples were, if at all, only taken from a small fraction of participants consistently over time. Summed over all the investigated cohorts, only 273 (0.5%) participants have undergone CSF sampling at baseline and again 3 years after. In contrast to CSF, cognitive assessments follow the drop-out curves quite closely (**Figure 3C**). While these findings are not surprising given the invasiveness of CSF sample collection, they raise severe concerns regarding the robustness of statistical analysis results obtained from CSF data. In turn, this again elucidates that comparative longitudinal approaches in the AD field are mainly limited to cognitive assessments or suffer from

small sample size.

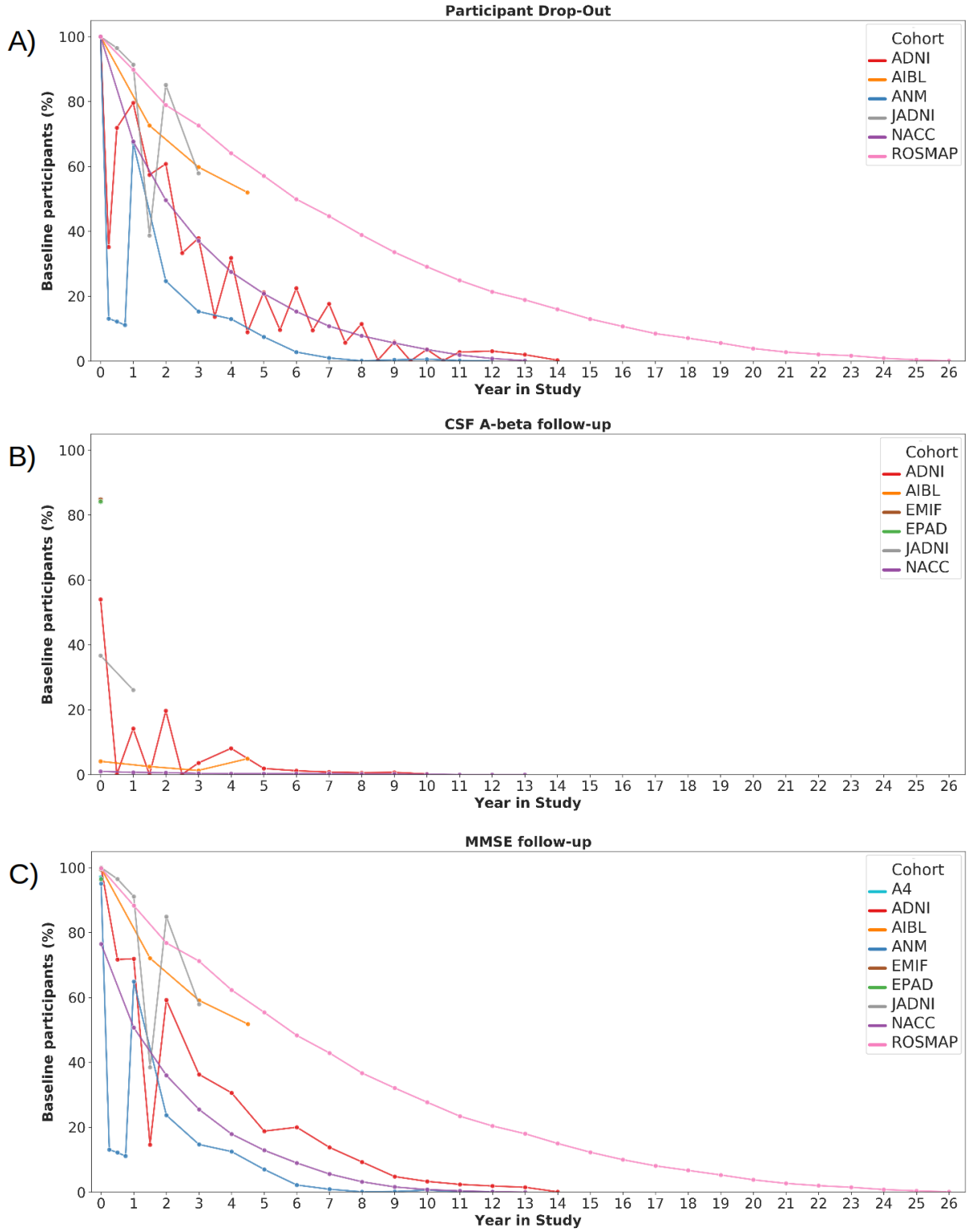

Figure 3: Longitudinal follow-up of **A)** at least one variable per participant, **B)** CSF amyloid beta, and **C)** MMSE scores.

4. Discussion:

In this work, we established an overview of the AD data landscape by investigating patient-level data from nine major clinical AD cohort studies. Our results demonstrate that the individual datasets vary with respect to key characteristics, such as number of enrolled participants per diagnosis, demographic composition, and distribution of important AD biomarkers. Assessing the ethnoracial diversity in the cohorts exposed a severe bias towards white/caucasian individuals since this group is predominantly overrepresented. To appraise availability of modalities in each study, we categorized each modality based on the relative presence of data in each cohort. Another important remark of our findings is the limited number of longitudinal follow-up measurements for important AD biomarkers like CSF amyloid beta. Finally, we made all results explorable through an interactive web application that can help researchers to identify cohort datasets suitable for their research.

Our analysis exposed major challenges that severely impede comparative approaches on AD cohort data. While there has been work on standardizing data collection [20,21] as well as on guidelines for defining an AD related data model [22], we still experience a deficit in interoperability across AD datasets. The investigated cohort datasets neither followed a common naming system for variables, nor represented values of the same measurement in equal manner. On top of that, some studies only shared processed values instead of the underlying raw data. This further impedes interoperability since differences in applied processing pipelines inevitably introduce systematic biases into the data. One promising approach to increase dataset interoperability could be a comprehensive, AD-specific common data model that would facilitate the alignment and mapping of variables for acquired datasets.

As previously mentioned, the abundance of longitudinal CSF data was limited throughout all acquired datasets. It is possible that because CSF sample collection is an invasive procedure [23], a substantial number of participants did not provide CSF samples. Although CSF biomarkers support disease diagnosis, it remains questionable whether longitudinal analyses of CSF data can produce statistically robust results given the low sample sizes available. Thus, the development of less invasive approaches like blood biomarkers could pose a more promising alternative for longitudinal assessments.

There are multiple reasons that could have caused the observed differences in demographic characteristics and disease risk factors across studies, namely, the study goal, the employed recruitment criteria, or the distinct approaches for participant acquisition. Potentially, these observed differences could severely hamper the comparison and validation of findings across disparate cohorts since such systematic differences can significantly influence the patterns and trends exhibited in the data. Up to now, it remains unclear how far this limits comparative approaches on AD data in practice and further investigations are required to ensure that results generated on AD datasets are robust and reproducible across multiple cohorts.

Funding

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EPAD grant agreement n°117536, resources of which are composed of financial contribution from the European Union's Seventh Framework Programme (FP7/2007-2013) and EFPIA companies' in kind contribution.

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 826421.

Acknowledgements

The authors want to thank Liu Shi and Alejo Nevado-Holgado for their help to access datasets.

The authors would also like to thank Lauren DeLong for her helpful comments.

We thank the study participants and staff of the Rush Alzheimer's Disease Center. ROSMAP was supported by NIA grants P30AG010161, R01AG015819, and R01AG017917.

The A4 Study is a secondary prevention trial in preclinical Alzheimer's disease, aiming to slow cognitive decline associated with brain amyloid accumulation in clinically normal older individuals. The A4 Study is funded by a public-private-philanthropic partnership, including funding from the National Institutes of Health-National Institute on Aging, Eli Lilly and Company, Alzheimer's Association, Accelerating Medicines Partnership, GHR Foundation, an anonymous foundation and additional private donors, with in-kind support from Avid and Cogstate. The companion observational Longitudinal Evaluation of Amyloid Risk and Neurodegeneration (LEARN) Study is funded by the Alzheimer's Association and GHR Foundation. The A4 and LEARN Studies are led by Dr. Reisa Sperling at Brigham and Women's Hospital, Harvard Medical School and Dr. Paul Aisen at the Alzheimer's Therapeutic Research Institute (ATRI), University of Southern California. The A4 and LEARN Studies are coordinated by ATRI at the University of Southern California, and the data are made available through the Laboratory for Neuro Imaging at the University of Southern California. The participants screening for the A4 Study provided permission to share their de-identified data in order to advance the quest to find a successful treatment for Alzheimer's disease. We would like to acknowledge the dedication of all the participants, the site personnel, and all of the partnership team members who continue to make the A4 and LEARN Studies possible. The complete A4 Study Team list is available on: a4study.org/a4-study-team.

J-ADNI was supported by the following grants: Translational Research Promotion Project from the New Energy and Industrial Technology Development Organization of Japan; Research on Dementia, Health Labor Sciences Research Grant; Life Science Database Integration Project of Japan Science and Technology Agency; Research Association of Biotechnology (contributed by Astellas Pharma Inc., Bristol-Myers Squibb, Daiichi-Sankyo, Eisai, Eli Lilly and Company, Merck-Banyu, Mitsubishi Tanabe Pharma, Pfizer Inc., Shionogi & Co., Ltd., Sumitomo Dainippon, and Takeda Pharmaceutical Company), Japan, and a grant from an anonymous Foundation.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer’s Association; Alzheimer’s Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (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 NACC database is funded by NIA/NIH Grant U01 AG016976. NACC data are contributed by the NIA-funded ADCs: P30 AG019610 (PI Eric Reiman, MD), P30 AG013846 (PI Neil Kowall, MD), P30 AG062428-01 (PI James Leverenz, MD) P50 AG008702 (PI Scott Small, MD), P50 AG025688 (PI Allan Levey, MD, PhD), P50 AG047266 (PI Todd Golde, MD, PhD), P30 AG010133 (PI Andrew Saykin, PsyD), P50 AG005146 (PI Marilyn Albert, PhD), P30 AG062421-01 (PI Bradley Hyman, MD, PhD), P30 AG062422-01 (PI Ronald Petersen, MD, PhD), P50 AG005138 (PI Mary Sano, PhD), P30 AG008051 (PI Thomas Wisniewski, MD), P30 AG013854 (PI Robert Vassar, PhD), P30 AG008017 (PI Jeffrey Kaye, MD), P30 AG010161 (PI David Bennett, MD), P50 AG047366 (PI Victor Henderson, MD, MS), P30 AG010129 (PI Charles DeCarli, MD), P50 AG016573 (PI Frank LaFerla, PhD), P30 AG062429-01(PI James Brewer, MD, PhD), P50 AG023501 (PI Bruce Miller, MD), P30

AG035982 (PI Russell Swerdlow, MD), P30 AG028383 (PI Linda Van Eldik, PhD), P30 AG053760 (PI Henry Paulson, MD, PhD), P30 AG010124 (PI John Trojanowski, MD, PhD), P50 AG005133 (PI Oscar Lopez, MD), P50 AG005142 (PI Helena Chui, MD), P30 AG012300 (PI Roger Rosenberg, MD), P30 AG049638 (PI Suzanne Craft, PhD), P50 AG005136 (PI Thomas Grabowski, MD), P30 AG062715-01 (PI Sanjay Asthana, MD, FRCP), P50 AG005681 (PI John Morris, MD), P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Competing interests

The authors have nothing to declare.

References:

1. Kalra, D. (2019). The importance of real-world data to precision medicine.

2. Fröhlich, H., Balling, R., Beerenwinkel, N., Kohlbacher, O., Kumar, S., Lengauer, T., *et al.* (2018). From hype to reality: data science enabling personalized medicine. *BMC medicine*, *16*(1), 150.

3. Whitwell, J. L., Wiste, H. J., Weigand, S. D., Rocca, W. A., Knopman, D. S., Roberts, R. O. *et al.* (2012). Comparison of imaging biomarkers in the Alzheimer disease neuroimaging initiative and the Mayo Clinic Study of Aging. *Archives of neurology*, *69*(5), 614-622.

4. Ferreira, D., Hansson, O., Barroso, J., Molina, Y., Machado, A., Hernández-Cabrera, J. A. *et al.* (2017). The interactive effect of demographic and clinical factors on hippocampal volume: A multicohort study on 1958 cognitively normal individuals. *Hippocampus*, *27*(6), 653-667.

5. Oliveira, J. L., Trifan, A., Silva, L. A. B. (2019). EMIF Catalogue: A collaborative platform for sharing and reusing biomedical data. *International journal of medical informatics*, 126, 35-45. doi: https://doi.org/10.1016/j.ijmedinf.2019.02.006

6. Janssen, O., Vos, S. J., García-Negredo, G., Tochel, C., Gustavsson, A., Smith, M., *et al.* (2019). Real-world evidence in Alzheimer's disease: The ROADMAP Data Cube. *Alzheimer's & Dementia*.

7. Lawrence, E., Vegvari, C., Ower, A., Hadjichrysanthou, C., De Wolf, F., Anderson, R. M. (2017). A Systematic review of longitudinal studies which measure alzheimer's disease biomarkers. *Journal of Alzheimer's Disease*, *59*(4), 1359-1379.

8. Mueller, S. G., Weiner, M. W., Thal, L. J., Petersen, R. C., Jack, C. R., Jagust, W., *et al.* (2005). Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). Alzheimer's & Dementia,

1(1), 55-66.

9. Lovestone, S., Francis, P., Kloszewska, I., Mecocci, P., Simmons, A., Soininen, H., *et al.* (2009). AddNeuroMed—the European collaboration for the discovery of novel biomarkers for Alzheimer's disease. Annals of the New York Academy of Sciences, 1180(1), 36-46.

10. Ellis, K. A., Bush, A. I., Darby, D., De Fazio, D., Foster, J., Hudson, P., *et al.* (2009). 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. *International Psychogeriatrics*, *21*(4), 672-687.

11. Bos, I., Vos, S., Vandenberghe, R., Scheltens, P., Engelborghs, S., Frisoni, G., *et al.* (2018). The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. *Alzheimer's research & therapy*, *10*(1), 64.

12. Solomon, A., Kivipelto, M., Molinuevo, J. L., Tom, B., Ritchie, C. W. (2018). European prevention of Alzheimer’s dementia longitudinal cohort study (EPAD LCS): study protocol. *BMJ open*, *8*(12), e021017.

13. Iwatsubo, T., Iwata, A., Suzuki, K., Ihara, R., Arai, H., Ishii, K., *et al.* (2018). Japanese and North American Alzheimer's Disease Neuroimaging Initiative studies: harmonization for international trials. *Alzheimer's & Dementia*, *14*(8), 1077-1087.

14. Besser, L., Kukull, W., Knopman, D. S., Chui, H., Galasko, D., Weintraub, S., *et al.* (2018). Version 3 of the National Alzheimer’s coordinating center’s uniform data set. *Alzheimer disease and associated disorders*, *32*(4), 351.

15. Bennett, D. A., Buchman, A. S., Boyle, P. A., Barnes, L. L., Wilson, R. S., Schneider, J. A. (2018). Religious orders study and rush memory and aging project. *Journal of Alzheimer's Disease*, *64*(s1), S161-S189.

16. Hye, A., Lynham, S., Thambisetty, M., Causevic, M., Campbell, J., Byers, H. L., *et al.* (2006). Proteome-based plasma biomarkers for Alzheimer's disease. *Brain*, *129*(11), 3042-3050.

17. Babulal, G. M., Quiroz, Y. T., Albensi, B. C., Arenaza-Urquijo, E., Astell, A. J., Babiloni, C., *et al.* (2019). Perspectives on ethnic and racial disparities in Alzheimer's disease and related dementias: Update and areas of immediate need. *Alzheimer's & Dementia*, *15*(2), 292-312.

18. McKhann, G., Drachman, D., Folstein, M., Katzman, R., Price, D., Stadlan, E. M. (1984). Clinical diagnosis of Alzheimer's disease: Report of the NINCDS-ADRDA Work Group* under the auspices of Department of Health and Human Services Task Force on Alzheimer's Disease. *Neurology*, *34*(7), 939-939.

19. O'Bryant, S. E., Gupta, V., Henriksen, K., Edwards, M., Jeromin, A., Lista, S., *et al.* (2015). Guidelines for the standardization of preanalytic variables for blood-based biomarker studies in Alzheimer's disease research. *Alzheimer's*

& Dementia, *11*(5), 549-560.

20. Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Cedarbaum, J., *et al.* (2015). Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimer's & Dementia*, *11*(7), 865-884.

21. Neville, J., Kopko, S., Romero, K., Corrigan, B., Stafford, B., LeRoy, E., *et al.* (2017). Accelerating drug development for Alzheimer's disease through the use of data standards. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, *3*(2), 273-283.

22. Sand, T., Stovner, L. J., Dale, L., & Salvesen, R. (1987). Side effects after diagnostic lumbar puncture and lumbar iohexol myelography. *Neuroradiology*, *29*(4), 385-388.