Received: 10 September 2020 | Accepted: 21 September 2020 | Published online: 2 December 2020

DOI: 10.1002/trc2.12102

Alzheimer's & Dementia
Translational Research & Clinical Interventions

RESEARCH ARTICLE

Evaluating the Alzheimer's disease data landscape

Colin Birkenbihl[1,2] | **Yasamin Salimi**[1,2] | **Daniel Domingo-Fernández**[1,2] | **Simon Lovestone**[3] | **AddNeuroMed consortium** | **Holger Fröhlich**[1,2] | **Martin Hofmann-Apitius**[1,2] | **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, Germany

[2] Bioinformatics Group, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

[3] Department of Psychiatry, University of Oxford, Oxford, UK

Correspondence
Colin Birkenbihl, Fraunhofer-Institute for Algorithms and Scientific Computing (SCAI), Schloss Birlinghoven, D-53757 Sankt Augustin, Germany.
Email: colin.birkenbihl@scai.fraunhofer.de

[*] Japanese Alzheimer's Disease Neuroimaging Initiative: 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.

[†] Alzheimer's Disease Neuroimaging Initiative: 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

Funding information
European Union's Seventh Framework Programme, Grant/Award Number: FP7/2007-2013

Abstract

Introduction: Numerous studies have collected Alzheimer's disease (AD) cohort data sets. 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 data sets 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 toward White/Caucasian individuals. We described and compared the measured data modalities. Finally, the available longitudinal data for important AD biomarkers was evaluated. All results are explorable through our web application ADataViewer (https://adata.scai.fraunhofer.de).

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

KEYWORDS
Alzheimer's disease, biomarker, clinical study, cohort, cohort study, data, data access, data sharing, data viewer, data-driven, data set, dementia, disease modeling, FAIR data, magnetic resonance imaging, open-science, patient level data

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 reflects patients of a particular subpopulation defined by inclusion and exclusion criteria. This is becoming especially relevant with respect to the increasing popularity of data-driven approaches and machine learning.[1,2] 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 data sets to assess whether the observed patterns are robust.[3] Such systematic data-driven approaches are, however, hampered because patient-level data are often difficult to access or entirely inaccessible. Moreover, we have limited knowledge about how the distinct cohort data sets available in our field compare to each other on a qualitative (eg, overlap of measured variables) as well as quantitative level (eg, values encountered in the data).[4,5] Thus, to leverage the full potential of collected patient-level data, it is important to characterize the clinical AD data landscape in detail.

1.1 | Metadata-driven evaluations of the Alzheimer's disease data landscape

Evaluating a data landscape involves organizing and comparing data sets 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 data sets 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 to demonstrate how cohort data sets 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 data sets. The resulting metadata is presented through the EMIF-Catalog.[6] Similarly, the Real world Outcomes across the Alzheimer's Disease spectrum for better care: Multi-modal data Access Platform (ROADMAP) project generated an overview of clinical outcomes and data modalities that were collected in several European AD cohort studies.[7] 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 data sets and gathered the contained information.[7]

RESEARCH IN CONTEXT

1. Systematic review: The authors reviewed relevant literature through bibliographic search engines. Relevant cohort data sets have been discovered through data portals, data publications, and citations in the literature. Applications were filed for 18 cohort data collections of which 9 were successful.
2. Interpretation: The presented results illustrate the current state of the Alzheimer's disease (AD) data landscape from a patient-level data-centric perspective, whereas previous investigations relied solely on provided cohort metadata. This investigation exposes limitations in data availability and interoperability, and establishes a detailed overview on what resources current data sets provide for data-driven analyses.
3. Future directions: This work emphasizes the need for a common semantic framework for patient-level AD data to enable the community to work across cohort data sets and ultimately to generate robust scientific insights to advance AD research.

1.2 | Moving beyond metadata through data-level investigations

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 data sets. 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 follow. 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 data sets and that the data sets are 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.

1.3 | Novelty and impact of this work

In this work, we aimed at assessing the current AD data landscape through meticulous investigation and curation of accessible cohort

TABLE 1 The investigated AD cohorts and their references

Cohort	Consortium	Reference
A4	Anti-Amyloid Treatment in Asymptomatic Alzheimer's Disease	9
ADNI	The Alzheimer's Disease Neuroimaging Initiative	10
ANMerge	AddNeuroMed	11
AIBL	The Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing	12
EMIF-1000	European Medical Information Framework	13
EPAD v1500	European Prevention of Alzheimer's Dementia	14
JADNI	Japanese Alzheimer's Disease Neuroimaging Initiative	15
NACC	The National Alzheimer's Coordinating Center	16
ROSMAP	The Religious Orders Study and Memory and Aging Project	17

data sets on the data level rather than solely relying on metadata and/or literature. To accomplish this task, we traced down, accessed, investigated, and compared nine of the major clinical cohort study data sets available in the AD field. Here, we comprehensively describe the acquired data and show which data modalities we found in the data sets as well as their overlaps with other studies. In addition, we assessed the longitudinal follow-up on the biomarker level and demonstrated to what extent current AD data are covering the progression of the disease. Furthermore, we compared the content we observed in these data sets with the reported findings of metadata-based approaches.[6,8] Finally, we made all results available through ADataViewer (https://adata.scai.fraunhofer.de), an interactive web-portal that allows researchers to explore the AD data landscape generated based on the investigated data sets.

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 data sets that were downloadable, hereby excluding data portals with restricted data access from our investigations. Most of the data sets we accessed were shared after completing an official data request process. We applied for access to 18 distinct AD cohort data sets. Until submitting this work for publication, we were granted access to nine (Table 1). We discuss the reasons behind failed data access applications in the Supplementary Text. Notably, not all of the accessed data sets are observational cohort studies in the strict sense; for more information, please see the Supplementary Text.

It is important to be aware that not all of these studies followed the same design or goals. Each study enforced its own recruitment criteria and enrolled participants following distinct selection processes. Although some aimed for a case-control setting and included a substantial amount of AD patients in their cohort, others deliberately excluded them to focus on early disease progression. Therefore, the cohort data sets are all subject to inherent biases.

2.2 | Generating the summary statistics

To illustrate the content of the data sets, 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. The AD biomarkers we compared between cohorts are motivated in the Supplementary Text. In addition, 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.[19] More detailed definitions of the ethnoracial groups can be found 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 data sets, single variables were reported only numerically given that they were placed within a defined value range (eg, 400 to 1700). If the measurement appeared to be outside of this range, the exact number was not reported but replaced with a cutoff (eg, ">1700"). To allow for calculations, we considered these values to be equal to the mentioned cutoff (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 data sets is not straightforward. This process involves intensive and meticulous manual curation of the acquired data sets 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 (SNPs) is not equal to conducting a genome-wide genotyping of individuals. Similarly, distributing normalized brain volumes summed over both hemispheres is less informative than providing the underlying raw magnetic resonance (MR) images. The latter would enable researchers to process the images according to their needs, whereas the former impedes interoperability to other data sets due to differences in employed image-processing pipelines. This could hamper certain analyses such as 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

TABLE 2 Description of the investigated cohorts

Cohorts	N	Healthy	MCI	AD	N with 2+ visits	Follow-up Interval (months)	Location	Diagnostic criteria AD
A4	6943	6943	0	0	0:	≈8	US, Canada, Australia	AD patients excluded
ADNI	2249	813	1016	389	1978 (88%)	6	USA, Canada	NINCDS-ADRDA
AIBL	1378	803	134	181	1019 (74%)	18	Australia	NINCDS-ADRDA
ANMerge	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	US	UDS Form D1
ROSMAP	3627	2514	898	203	3335 (92%)	12	US	NINCDS-ADRDA

NOTE: The numbers of diagnosed subjects do not always add up to N, since patients with different dementia diagnoses (eg, Lewy body 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 are available; Follow-up Interval, Approximated regular time interval between participant visits; Longitudinal data have been collected but are not yet released.

TABLE 3 Distribution of demographic variables and key AD biomarkers encountered in each cohort

	Female %	Age	Education	APOE ε4 %	MMSE	CDR	CDR-SB	Hippocampus	A-beta	t-Tau	p-Tau
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
ANMerge	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	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: We show the 25%, 50%, and 75% quantiles of numerical variables at baseline. Categorical variables are given as the proportion of participants falling into one respective category. *APOE* ε4%, Proportion of participants with at least one *APOE* ε4 allele; Hippocampus, A-beta, t-Tau, p-Tau, NACC values are given as the proportion of "abnormal observations".

of each modality into three discrete stages (Supplementary Table S1): 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 data sets cover the time dimension of AD, we conducted a thorough investigation of their respective longitudinal follow-up. For each cohort, we evaluated how many participants were assessed at each follow-up visit and implicitly analyzed the 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 insight on how much information we can exploit to model and ultimately understand patterns of AD progression. As of publication of this article, EPAD and NACC are still subject of ongoing data collection, while ADNI received funding to extend their study and continue participant recruitment.

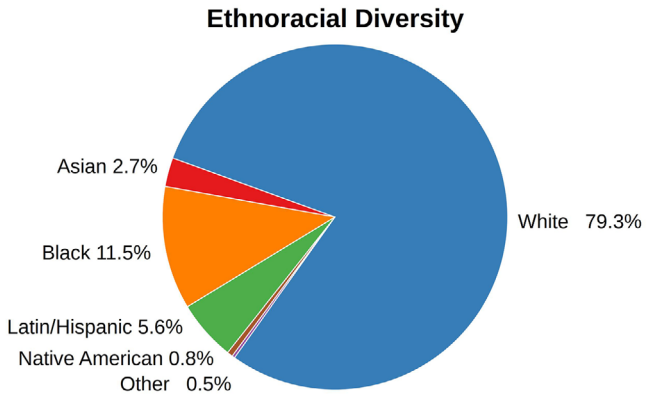

FIGURE 1 Combined ethnoracial diversity found across the investigated AD cohorts. Table S2 shows the individual compositions of each cohort

3 | RESULTS

3.1 | Investigation of the AD data landscape

Altogether, we investigated data from nine studies comprising a total of 60,004 assessed study participants. Table 2 shows how these participants were distributed among the analyzed cohorts. With NACC being the exception (n = 40,858), 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, patients with mild cognitive impairment (MCI), and patients with AD. Seven of the investigated studies based their diagnoses on the National Institute of Neurological and Communicative Disorders and Stroke-Alzheimer's Disease and Related Disorders Association (NINCDS-ADRDA) criteria[20] which significantly increases the interoperability between those data sets, since AD follows the same semantic description. Depending on each study's goals, the recruitment process focused on enrolling more or fewer individuals falling into specific diagnosis groups.

Although no data are shared through our web-portal, information on how to access the data sets can be found at https://adata.scai.fraunhofer.de/cohorts.

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, whereas participants of ANMerge 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. Overall, most individuals included in the AD cohort studies were highly educated (≈14 years on average). As previously mentioned by Whitwell et al., a high level of education can act as cognitive reserve, possibly concealing a prodromal manifestation of AD.[5] Numerous demographic differences found between studies may result from distinct recruitment criteria which, again, mirror the individual study goals. Although distinct recruitment criteria lead to a broader sampling of the AD population, they reduce the direct comparability between data sets because they inevitably introduce bias into the data. One key example is recruitment specifically for participants with AD risk factors (eg, *APOE* genotype). This could significantly bias the patterns exhibited in the data in comparison to another data set with a lower amount of *APOE* ε4–positive participants.

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 data sets demonstrates that most of these recruited individuals come from a White/Caucasian background (79.3%). The second largest group was Black/African descendants with 11.5%, followed by participants of Latin/Hispanic heritage with 5.6%. 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. Because the majority of the studies have been conducted in the United States, 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 toward White/Caucasian individuals (EPAD: 99% white; ANMerge: 98,5% white; see https://adata.scai.fraunhofer.de/ethnicity).

The ethnoracial composition in the investigated cohorts relies on the diversity of populations from which the participants have been recruited. Nonetheless, our results elucidate that there is a substantial bias toward White/Caucasian in AD data sets and a severe underrepresentation of other ethnoracial groups, which, in turn, could be problematic for developing personalized treatments.

3.3 | Availability of data modalities

To analyze which modalities are available in our investigated cohorts and to explore the overlaps between them, we assigned a score of availability per data modality according to our previously described criteria (Table S1).

In Figure 2A, we show an overview of the data modalities and their availability score in all acquired cohort data sets. Commonly assessed modalities throughout all studies were demographic variables (eg, age, sex, and education) as well as clinical assessments (eg, Mini Mental State Examination [MMSE]). Regarding these two modalities, eight studies were assigned the availability score 2, with EMIF and AIBL being the only exceptions due to missing ethnoracial information. Cerebrospinal fluid (CSF) biomarker measurements were found to be present in all data sets but ANMerge. With regard to autopsy data, only ROSMAP contained a detailed collection, ranging from simple measurements such as brain weight to comprehensive brain proteomics

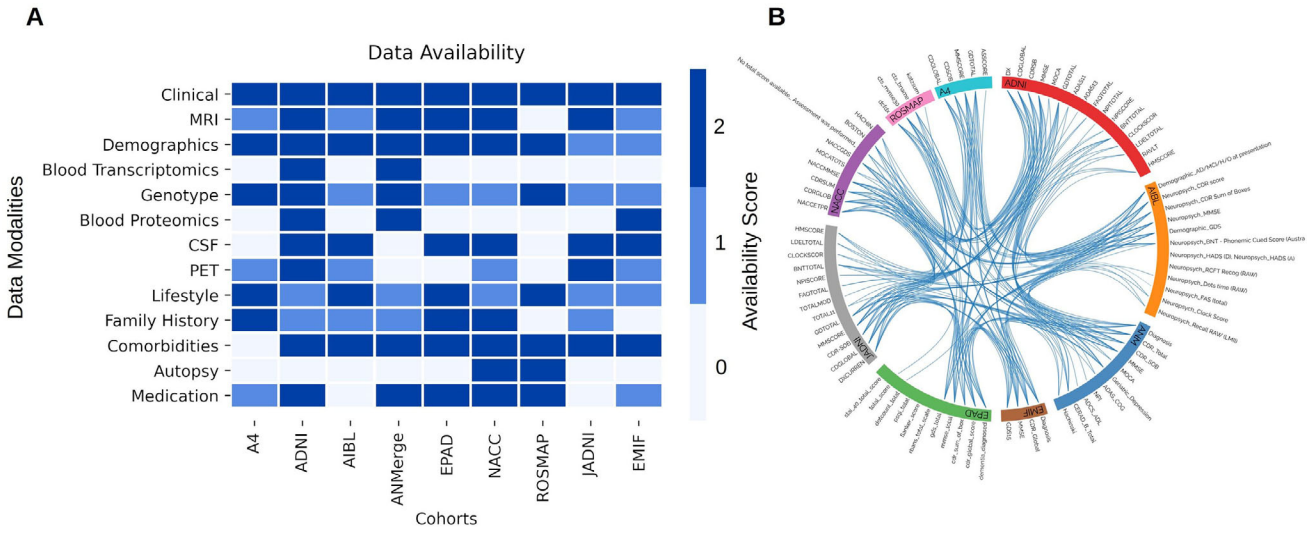

FIGURE 2 Interoperability of AD data sets. A, Availability of data modalities scored based on the defined criteria. The criteria are explained in Supplementary Table 1. B, Equivalence of clinical assessment variables across cohorts. PET = positron emission tomography

and transcriptomics. Although seven studies released some structural MRI data, three of those limited the shared data to processed MRI features (eg, brain volumes). In our case, only ADNI, NACC, JADNI, EPAD, and ANMerge granted access to the raw images.

Although the purpose of this section is to provide a comprehensive overview about the availability of data modalities, we would like to emphasize that the presented results are strongly dependent on our defined curation criteria, and different criteria could lead to deviating results. In addition, all investigated data sets could hold more information than we presented here. Due to our premise of looking exclusively into those patient-level data that have indeed been shared with us, it is possible that we missed modalities or resources that are existent but were not shared (eg, MRI images). Our results can be explored at https://adata.scai.fraunhofer.de/modality.

3.4 | Metadata investigation versus data investigations

To establish how our observations of data availability differed from results gained by solely investigating metadata, we qualitatively compared our findings to the metadata presented in the EMIF catalog.[6*] Only four of our investigated studies were listed: ADNI, ANMerge, 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 data sets that are reported as absent in the catalog (eg, Global Deterioration Scale in ANMerge), or were not listed at all. Other variables and even modalities are reported to be present, yet could not be found in the respective data set. For instance, the catalog states that post-mortem brain autopsy was performed in ANMerge, for which we could not find any evidence.

[*] Accessed on February 2, 2020.

Similar observations were made when comparing our findings to the review by Lawrence et al.[8] Here, for example, the reported longitudinal follow-up of ANMerge is significantly shorter than what we observed in the data (reported: 12 months, data: 84 months). In addition, the reported number of participants with at least two visits does not equal our findings (reported: 378, data: 1254 participants).

3.5 | Availability of data modalities

The finding of common modalities across cohorts does not imply that the measured variables are interoperable or even comparable on a semantic level. By mapping a variety of variables across the data sets, we established an overview of their interoperability (Figure 2B). We would like to emphasize that the current version of these mappings is not complete but a proof of concept that a semantic integration of these data sets is, in theory, possible. However, this integration is cumbersome and time-consuming, as many data sets exhibit low interoperability and distinct variable naming conventions. An in-depth view of the preliminary mappings is given at https://adata.scai.fraunhofer.de/feature_comparison.

3.6 | 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 (ie, participants with an 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), ANMerge (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 Dementia Staging Instrument (CDR) Sum of Boxes (CDR-SOB) scores slightly shift the perspective. Here, ANMerge is the most affected cohort, with its 25%, 50%, and 75% quantiles of the CDR-SOB scores being 4, 6, and 9, respectively. AIBL patients scored 3.5, 5, and 7, which slightly contradicts the image painted by the MMSE scores. Again, ADNI shows the least cognitive symptoms with its CDR-SOB quantiles being 3, 4.5, and 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 data sets is an alignment of their underlying data models (ie, making data interoperable). In our analysis, we found that 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 (eg, combine "lh_hippo_volume" and "rh_hippo_volume" and map to "Hippocampus"). Another difficulty is that numerous data sets 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, whereas other studies provide numerical values that were capped at different thresholds between studies (eg, " >1700″). All these factors led to a severe lack of interoperability between data sets, which significantly limits comparative approaches and restricts them to more standardized variables like clinical assessment scores.

3.7 | 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 biomarker values were acquired for the same individual on all visits. Thus we additionally investigated the amount of study participants for which select 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 biomarker that we selectively investigated 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, taken only from a small fraction of participants consistently over time. Summed over all the investigated cohorts, only 273 participants (0.5%) 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). Although 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 limited mainly to cognitive assessments or suffer from small sample size.

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 data sets 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 overrepresentation of White/Caucasian individuals compared to other ethnoracial backgrounds. To appraise the 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 ADataViewer, an interactive web application that can help researchers to identify cohort data sets that are suitable for their research.

4.1 | Achieving data set interoperability through one common data model

Our analysis exposed major challenges that severely impede comparative approaches on AD cohort data. Although there has been work on standardizing data collection[21,22] as well as on guidelines defining an AD-specific data model,[23] we still experience a deficit in interoperability across AD data sets. The investigated cohort data sets neither followed a common naming system for variables nor represented values of the same measurement in an equal manner. On top of that, some studies shared only 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 data set interoperability could be a comprehensive, AD-specific common data model. Such a data model could support the alignment and mapping of variables by providing easy-to-follow guidelines and a dedicated interface for retrospective data harmonization.

4.2 | Data limitations hamper disease modeling

In the context of personalized medicine, training models on predominantly White/Caucasian participants can lead to biased models. It is known that exhibited patterns of biomarker measurements differ across AD patients from distinct ethnoracial groups.[25,26] Given that

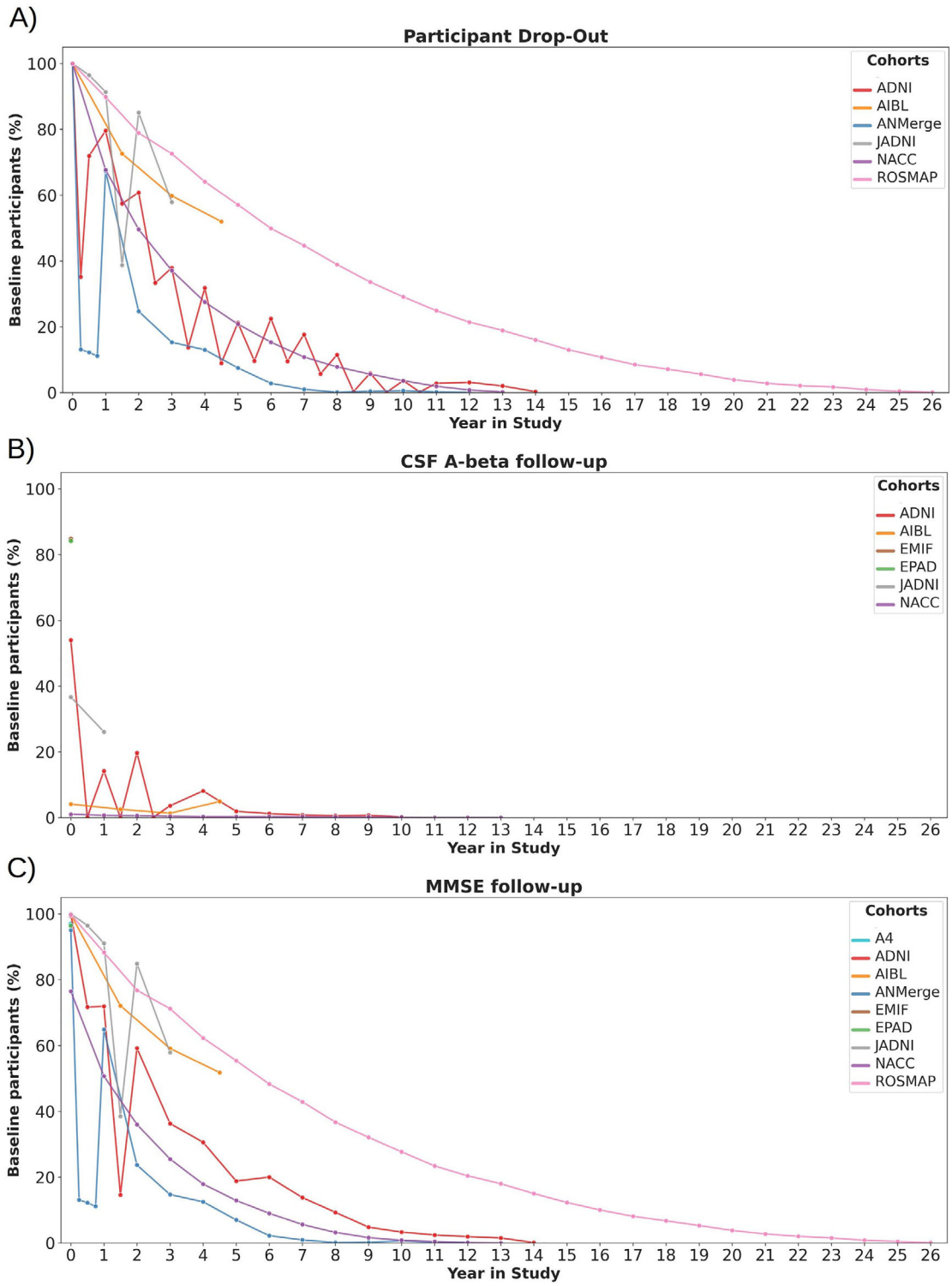

FIGURE 3 Longitudinal follow-up as the proportion of participants at study baseline (ie, participants were aligned based on their first visit). A, At least one variable measured. B, CSF amyloid beta. C, MMSE scores. CSF = cerebrospinal fluid. MMSE = Mini Mental State Examination

there are only limited data from non-White participants available, trained models could fail to learn such ethnoracial-specific signals, which, in turn, would result in poor performance for individuals of non-White background.

As mentioned previously, the abundance of longitudinal CSF data was limited throughout all acquired data sets. One possible reason explaining participants' reluctance to provide CSF samples, especially repeatedly, is the invasiveness of its sampling procedure.[24] Although cross-sectional CSF biomarkers can support AD diagnosis, longitudinal measurements are fundamental to understand disease progression on a biomarker-level. Given the low CSF sample sizes currently available, it remains questionable whether longitudinal analyses of these data can

generate robust insights on conversions between normal and abnormal values of CSF biomarkers.

4.3 | Actionable knowledge through data-driven landscapes

The evident contradictions found between our data-driven investigation and the metadata-based approaches (Section 3.4) can be divided into two types. Type 1 describes that we found variables in the data sets that were reported as missing according to metadata resources. 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 data sets. Contradiction type 2, on the other hand, resembles cases in which 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 are not themselves available. We believe that our presented comparison highlights that, despite their significantly higher demand for time and effort, data-driven investigations should be preferred when assessing a data landscape.

4.4 | Future perspectives

The observed differences in demographic characteristics and disease risk factors across studies could severely hamper the comparison and validation of findings across disparate cohorts, since they can significantly influence the patterns and trends exhibited in the data.[2] Until now, only limited insight is available on how much the heterogeneous data landscape limits comparative approaches and cross-cohort disease modeling on AD data. Further systematic investigations are required to ensure that results generated on AD data sets are robust and reproducible across multiple cohorts. To support such endeavors, we aim to improve the ADataViewer to include more data sets, variable mappings, and the results of systematic data set comparisons in the future.

ACKNOWLEDGEMENTS

The authors want to thank Liu Shi and Alejo Nevado-Holgado for her help in accessing data sets.

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, the 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 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 at: 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 the 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).

The research leading to these results has received support from the Innovative Medicines Initiative Joint Undertaking under EPAD grant agreement no. 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, "TheVirtualBrain-Cloud."

Open access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors have nothing to declare.

REFERENCES

1. Kalra D. The importance of real-world data to precision medicine. *Per Med*. 2019;16(2):79-82.
2. Birkenbihl C, Emon MA, Vrooman H, et al. Differences in cohort study data affect external validation of artificial intelligence models for predictive diagnostics of dementia-lessons for translation into clinical practice. *EPMA J*. 2020;11(3):367-376.
3. Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med*. 2018;16(1):150.
4. Whitwell JL, Wiste HJ, Weigand SD, et al. Comparison of imaging biomarkers in the Alzheimer disease neuroimaging initiative and the Mayo Clinic Study of Aging. *Arch Neurol*. 2012;69(5):614-622.
5. Ferreira D, Hansson O, Barroso J, et al. The interactive effect of demographic and clinical factors on hippocampal volume: a multicohort study on 1958 cognitively normal individuals. *Hippocampus*. 2017;27(6):653-667.
6. Oliveira JL, Trifan A, Silva LAB. EMIF Catalogue: a collaborative platform for sharing and reusing biomedical data. *Int J Med Inf*. 2019;126:35-45.
7. Janssen O, Vos SJ, García-Negredo G, et al. Real-world evidence in Alzheimer's disease: the ROADMAP Data Cube. *Alzheimers Dement (N Y)*. 2020;16(3):461-471.
8. Lawrence E, Vegvari C, Ower A, Hadjichrysanthou C, De Wolf F, Anderson RM. A Systematic review of longitudinal studies which measure alzheimer's disease biomarkers. *J Alzheimers Dis*. 2017;59(4):1359-1379.
9. Sperling RA, Rentz DM, Johnson KA, et al. The A4 study: stopping AD before symptoms begin?. *Sci Transl Med*. 2014;6(228):228fs13-228fs13.
10. Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's Disease Neuroimaging Initiative (ADNI). *Alzheimers Dement (N Y)*. 2005;1(1):55-66.
11. Birkenbihl C, Westwood S, Shi L, et al. ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. *medRxiv*:2020.
12. Ellis KA, Bush AI, Darby D, et al. The Australian Imaging, Biomarkers and Lifestyle (AIBL) study of aging: methodology and baseline characteristics of 1112 individuals recruited for a longitudinal study of Alzheimer's disease. *Int Psychogeriatr*. 2009;21(4):672-687.
13. Bos I, Vos S, Vandenberghe R, et al. The EMIF-AD Multimodal Biomarker Discovery study: design, methods and cohort characteristics. *Alzheimers Res Ther*. 2018;10(1):64.
14. Solomon A, Kivipelto M, Molinuevo JL, Tom B, Ritchie CW. European prevention of Alzheimer's dementia longitudinal cohort study (EPAD LCS): study protocol. *BMJ Open*. 2018;8(12):e021017.
15. Iwatsubo T, Iwata A, Suzuki K, et al. Japanese and North American Alzheimer's Disease Neuroimaging Initiative studies: harmonization for international trials. *Alzheimers Dement (N Y)*. 2018;14(8):1077-1087.
16. Besser L, Kukull W, Knopman DS, et al. Version 3 of the National Alzheimer's coordinating center's uniform data set. *Alzheimer Dis Assoc Disord*. 2018;32(4):351.
17. Bennett DA, Buchman AS, Boyle PA, Barnes LL, Wilson RS, Schneider JA. Religious orders study and rush memory and aging project. *J Alzheimers Dis*. 2018;64(s1):S161-S189.
18. Hye A, Lynham S, Thambisetty M, et al. Proteome-based plasma biomarkers for Alzheimer's disease. *Brain*. 2006;129(11):3042-3050.
19. Babulal GM, Quiroz YT, Albensi BC, et al. Perspectives on ethnic and racial disparities in Alzheimer's disease and related dementias: update and areas of immediate need. *Alzheimers Dement (N Y)*. 2019;15(2):292-312.
20. McKhann G, Drachman D, Folstein M, Katzman R, Price D, Stadlan EM. 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*. 1984;34(7):939-939.
21. O'Bryant SE, Gupta V, Henriksen K, et al. Guidelines for the standardization of preanalytic variables for blood-based biomarker studies in Alzheimer's disease research. *Alzheimers Dement (N Y)*. 2015;11(5):549-560.
22. Weiner MW, Veitch DP, Aisen PS, et al. Impact of the Alzheimer's disease neuroimaging initiative, 2004 to 2014. *Alzheimers Dement (N Y)*. 2015;11(7):865-884.
23. Neville J, Kopko S, Romero K, et al. Accelerating drug development for Alzheimer's disease through the use of data standards. *Alzheimers Dement*. 2017;3(2):273-283.
24. Sand T, Stovner LJ, Dale L, Salvesen R. Side effects after diagnostic lumbar puncture and lumbar iohexol myelography. *Neuroradiology*. 1987;29(4):385-388.
25. Misiura MB, Howell JC, Wu J, et al. Race modifies default mode connectivity in Alzheimer's disease. *Transl Neurodegener*. 2020;9:8.
26. Howell JC, Watts KD, Parker MW, et al. Race modifies the relationship between cognition and Alzheimer's disease cerebrospinal

fluid biomarkers. *Alz Res Therapy*. 2017;9:88. https://doi.org/10.1186/s13195-017-0315-1.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of the article.

How to cite this article: Birkenbihl C, Salimi Y, Domingo-Fernández D, et al. Evaluating the Alzheimer's disease data landscape. *Alzheimer's Dement*. 2020;6:e12102. https://doi.org/10.1002/trc2.12102