Received: 16 December 2020 | Revised: 19 April 2021 | Accepted: 25 April 2021

DOI: 10.1002/alz.12387

Alzheimer's & Dementia®
THE JOURNAL OF THE ALZHEIMER'S ASSOCIATION

FEATURED ARTICLE

Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling

Colin Birkenbihl[1,2] | **Yasamin Salimi**[1,2] | **Holger Fröhlich**[1,2] | **for the Japanese Alzheimer's Disease Neuroimaging Initiative**[#] | **the Alzheimer's Disease Neuroimaging Initiative**[†]

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

[2] Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn, Germany

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

Colin Birkenbihl and Yasamin Salimi contributed equally to this work.

[#]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

Funding information
European Union's Horizon 2020, Grant/Award Number: 826421

Abstract

Introduction: Given study-specific inclusion and exclusion criteria, Alzheimer's disease (AD) cohort studies effectively sample from different statistical distributions. This heterogeneity can propagate into cohort-specific signals and subsequently bias data-driven investigations of disease progression patterns.

Methods: We built multi-state models for six independent AD cohort datasets to statistically compare disease progression patterns across them. Additionally, we propose a novel method for clustering cohorts with regard to their progression signals.

Results: We identified significant differences in progression patterns across cohorts. Models trained on cohort data learned cohort-specific effects that bias their estimations. We demonstrated how six cohorts relate to each other regarding their disease progression.

Discussion: Heterogeneity in cohort datasets impedes the reproducibility of data-driven results and validation of progression models generated on single cohorts. To ensure robust scientific insights, it is advisable to externally validate results in independent cohort datasets. The proposed clustering assesses the comparability of cohorts in an unbiased, data-driven manner.

KEYWORDS
Alzheimer's disease, cohort study, data mining, data-driven, disease modeling, machine learning, sampling bias, statistical learning, translational research

1 | BACKGROUND

In the last decade, understanding the progressive dynamics of Alzheimer's disease (AD) and AD clinical syndrome,[1] proved to be one of the fundamental challenges in our field.[2,3] Comprehensive knowledge on AD progression opens crucial opportunities for medical intervention to counteract or delay impediments to activities of daily living.[4] One path to facilitate this understanding manifests in the extraction of longitudinal progression signals from patient-level datasets collected in cohort studies. In this context, data mining and machine learning methods can be used to build mathematical models that elucidate and predict progression patterns hidden in the data. In the past, such progression models were used, for example, to approximate biomarker trajectories,[5] to identify distinct progression subtypes,[6] and to assess patient risk of progression toward more impaired disease stages.[7] However, to demonstrate that progression patterns identified in one cohort generalize beyond the discovery dataset itself, it is imperative to externally validate them in an independent dataset.[8] External validation data should originate from a separate cohort study independent from the training data used for building the model. Especially in the context of multifactorial and heterogenous diseases such as AD, external validation turns out to be a non-trivial undertaking.

The key limitation encountered in external validation manifests in the characteristics of clinical AD cohort data.[9] By nature of the disease, AD cohorts are very heterogeneous with respect to their exhibited progression,[10] for example, with respect to brain atrophy[11] and age of disease onset.[12] Furthermore, cohort study participants are recruited according to specific inclusion and exclusion criteria defined based on the goals of the study (e.g., selection of specific age ranges or risk factors). These specific sampling procedures shape potentially distinct statistical distributions from which each study's participants are recruited and, in turn, inevitably introduce cohort-specific statistical biases into the collected dataset itself.[13,14] These aspects potentially violate the fundamental assumption behind data mining and machine learning approaches that the participants of a validation dataset constitute a representative sample of the same population from which the original training data were drawn (Figure S1 in supporting information). Consequently, this indicates that training and validation data must be independently and identically distributed (i.i.d.) samples.[15] As such, a well-trained model should show similar performance on a validation dataset that was drawn from the identical statistical distribution as the training data, while an overfitted model would fail such validation. However, on a validation dataset that is violating the assumption of being sampled from the same statistical distribution as the training data even a well-trained model would fail, because the validation data falls outside the domain of the model (Figure S1). In conclusion, data-driven models trained on cohort datasets cannot be expected to generalize appropriately beyond the statistical distribution from which this cohort's participants were sampled.[16,17]

The heterogeneity found in AD cohort datasets, therefore, raises several important questions with respect to data-driven modeling of AD. First, it warrants an evaluation as to whether exhibited trends of disease progression are consistent across cohorts despite possible differences in their underlying populations. Further investigation should also determine whether progression models fitted on such datasets learn potential cohort-specific biases that could impede the generalizability of findings. Finally, as of now, there is no way to measure and express the general comparability between patient-level datasets on the level of disease progression. In the past, researchers mainly relied on comparing baseline study characteristics of their studied datasets.[7,18,19] However, for obvious reasons, evaluating variable distributions at a singular time point is a very limited comparison in the scope of disease progression. Deriving a quantitative measure to compare longitudinal progression patterns across multiple clinical studies could aid researchers to better understand the landscape of existing studies and to identify datasets that might fulfill the i.i.d. assumption. Furthermore, it could be used to investigate whether the cause of a significant drop in prediction performance lies in systematic differences between the training and validation datasets (i.e., a probable violation of the i.i.d. assumption) or simply in an overfitted model.

In this work, we evaluated the heterogeneity of disease progression patterns encountered in six longitudinal clinical AD cohort studies. Relying on multi-state models (MSM),[20] a well-established data mining approach in the AD field,[7,21–24] we performed a systematic comparison of progression patterns extracted from these studies to assess whether discovered signals are robust. Furthermore, we investigated whether cohort-specific biases propagate into trained

RESEARCH IN CONTEXT

1. **Systematic review**: The authors reviewed relevant literature using standard bibliographic search engines. Accessible cohort datasets have been discovered through data portals and citations in literature (primarily https://adata.scai.fraunhofer.de/).
2. **Interpretation**: The presented results illustrate the comparability of Alzheimer's disease (AD) progression across six major AD cohorts. We identified evident differences in progression patterns between cohorts and, furthermore, observed that data-driven approaches learn cohort-specific effects from their training data. These findings can impede the generalization of results generated on single cohorts. We propose a novel clustering approach for cohort data that helps to better understand which cohorts are comparable with respect to their exhibited disease progression.
3. **Future directions**: This work emphasizes the need for thorough validation of data-driven results. To eventually support clinical decision-making using data-driven approaches, it might be more promising to build models specific for disease subtypes or use domain adaptation techniques to address the encountered heterogeneity in cohort datasets.

progression models. Finally, we propose a novel method for clustering cohorts based on their exhibited progression patterns. This approach reveals the similarity of cohort studies in a data-driven and unbiased manner. It allows researchers to adequately understand and characterize performances measured via external validation of statistical and machine learning models developed on another cohort. In conclusion, our approach allows for better understanding of statistical differences that have previously been reported between various AD studies.[13]

2 | METHODS

2.1 | Data selection

Six longitudinal datasets stemming from the Alzheimer's Disease Neuroimaging Initiative (ADNI),[25] AddNeuroMed (ANMerge),[26] Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing (AIBL),[27] Japanese Alzheimer's Disease Neuroimaging (J-ADNI),[28] National Alzheimer's Coordinating Center (NACC),[29] and the Religious Orders Study and Rush Memory and Aging Project (ROSMAP)[30] were used as training datasets for our progression models. All of these studies obtained ethical approval for human data collection and informed patient consent for data sharing. We excluded participants whose mild cognitive impairment (MCI) diagnoses were not attributed to AD. Information on the cohorts with respect to key variables, as well as the number of participants, can be found in Table S1 in supporting information.

2.2 | Progression models applied for statistical analysis

To extract disease progression patterns from the investigated datasets, we fitted one MSM per cohort using the msm R package.[20] The states in our models represent the three commonly assessed stages for AD progression: cognitively unimpaired (CU), MCI, and AD. Consequently, transitions between states illustrate conversions from one clinical diagnosis stage to another. We modeled AD as an absorbing state, that is, we assumed that patients were not able to recover once deterioration was advanced enough to receive an AD diagnosis. However, because the classification of patients into CU, MCI, and AD in all cohorts had been performed based on clinical assessments, reversions from AD were observed in the data. These reversions were modeled as misclassifications. A graphical representation of the model can be seen in Figure S3 in supporting information. Each transition rate was estimated based on a set of covariates to account for the individual compositions of the cohorts. For determining the most informative combination of covariates, we performed a rigorous model selection using the Akaike's information criterion (AIC). The choice of covariates was mainly limited by their availability across the cohorts (Figure S2 in supporting information). Ultimately, the selected covariates comprised participant's age, biological sex, completed years of education, apolipoprotein E (*APOE*) $\varepsilon 4$ status, and the Mini-Mental State Examination (MMSE). Likelihood-ratio tests comparing each MSM to a null model demonstrated that all models extracted progression signals from their training dataset ($P < .05$). To rule out potential overfitting of the models, we built 150 models on repeated bootstrap samples from each respective cohort and observed low variation in model estimates (Table S3 in supporting information). Application of interval censoring allowed for the inclusion of participants with missing intermediate visits while right censoring was used for individuals who did not receive an AD diagnosis during study runtime. More details on the methodology and model selection are presented in the supporting information.

2.3 | Comparison of data mined progression patterns across cohorts

To explore and assess the heterogeneity in disease progression trends across cohorts, we estimated several progression patterns using each cohort's MSM: the state transition probabilities, probability of staying AD diagnosis free over time, and sojourn times (i.e., the expected time a participant spends in a considered state). All patterns were separately investigated for the CU and MCI states. For estimation of a cohort's progression patterns starting in the CU state, we used the covariate values observed at the study baseline of each of the respective cohort's CU participants. Similarly, for estimating transitions from the MCI state, we relied on the covariate values of participants at their first MCI diagnosis. Where appropriate, uncertainty of estimates was quantified using 95% confidence intervals (CI). Differences between cohort-specific distributions of the aforementioned progression estimates were determined using Kruskal-Wallis and pairwise Mann-Whitney *U* tests employing a confidence level of 95%. *P*-values were corrected for multiple testing using the Bonferroni-Holm method.

2.4 | Evaluation of cohort biases in statistical models

The second set of analyses aimed at elucidating whether MSMs fitted to data from a single cohort would learn cohort-specific effects that reduce generalizability to other cohorts. Hazard ratios, for example, are covariate-specific parameters of a model that quantify the influence of covariates onto the transition risk between two states. Comparing these ratios, it becomes apparent whether models learned the same covariate influences from distinct cohorts. Furthermore, we used each cohort's previously trained MSM to estimate the progression patterns for the same, combined set of participants from all cohorts. By fixing the data to be estimated across models, all variability in the progression patterns stems from the cohort-specific effects learned by the model. To evaluate the existence of these cohort-specific biases, we performed Kruskal-Wallis tests and pairwise Mann-Whitney

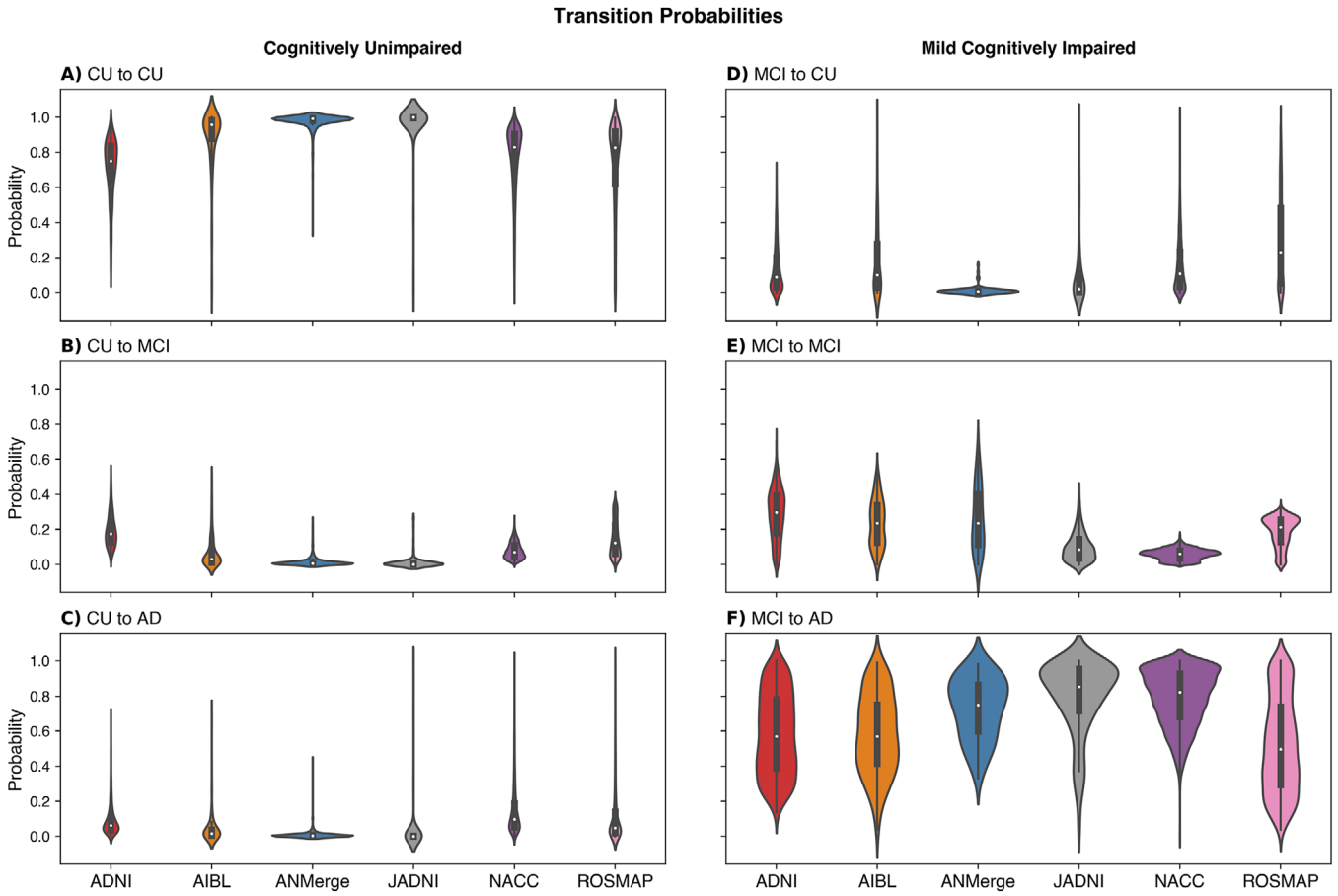


FIGURE 1 Probabilities to transition from one state to another are estimated for a 10-year period. Median probabilities are marked with white points. Statistical distributions are shown as box plots as well as superimposed kernel density estimates, resulting in violin plots. Because most deviations between depicted distributions were significant, we omit indication of significance for brevity. A-C, Transition probabilities starting from the cognitively unimpaired (CU) state. D-F, Transition probabilities starting from the mild cognitive impairment (MCI) state. AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

U tests, again correcting for multiple testing using Bonferroni-Holm and assuming a confidence level of 95%.

2.5 | Cohort similarity clustering

Whereas previous analyses focused on statistical differences between cohorts, we additionally developed an approach to cluster cohorts based on their global similarity across progression patterns. More specifically, each cohort's MSM was used to calculate the log-likelihood of observing the actual transitions of all the participants of each other cohort. These pairwise log-likelihoods were afterward averaged across the number of participants per cohort to eliminate biases toward cohort size. This resulted in a pairwise similarity matrix between cohorts which was subsequently transformed into a symmetric distance matrix. Mathematical details can be found in the supporting information. The resulting distance matrix was then used in an agglomerative hierarchical clustering approach using average linkage.

3 | RESULTS

3.1 | Progression patterns differ across cohorts

Transition probabilities estimated for a 10-year period varied significantly between cohorts (Figure 1). While we observed in all cohorts that participants in the CU state were most likely to remain CU over the next 10 years, the proportions of probabilities showed evident differences (Figure 1A-C). We discovered a range of 25% difference between the maximum and minimum observed median probability to remain CU (J-ADNI, > 99%; ADNI, 75%). All observed differences between pairwise combinations of cohorts were significant ($P < .001$), with the exception of ROSMAP–NACC for remaining in the CU state ($P = .3$).

When investigating the estimated transition probabilities from the MCI state (Figure 1D-F), all cohorts exhibited their most probable transition toward the AD state. J-ADNI showed the highest median probability across cohorts with 85%, while ROSMAP held the lowest median probability with 50%, exposing a difference of 35% between them.

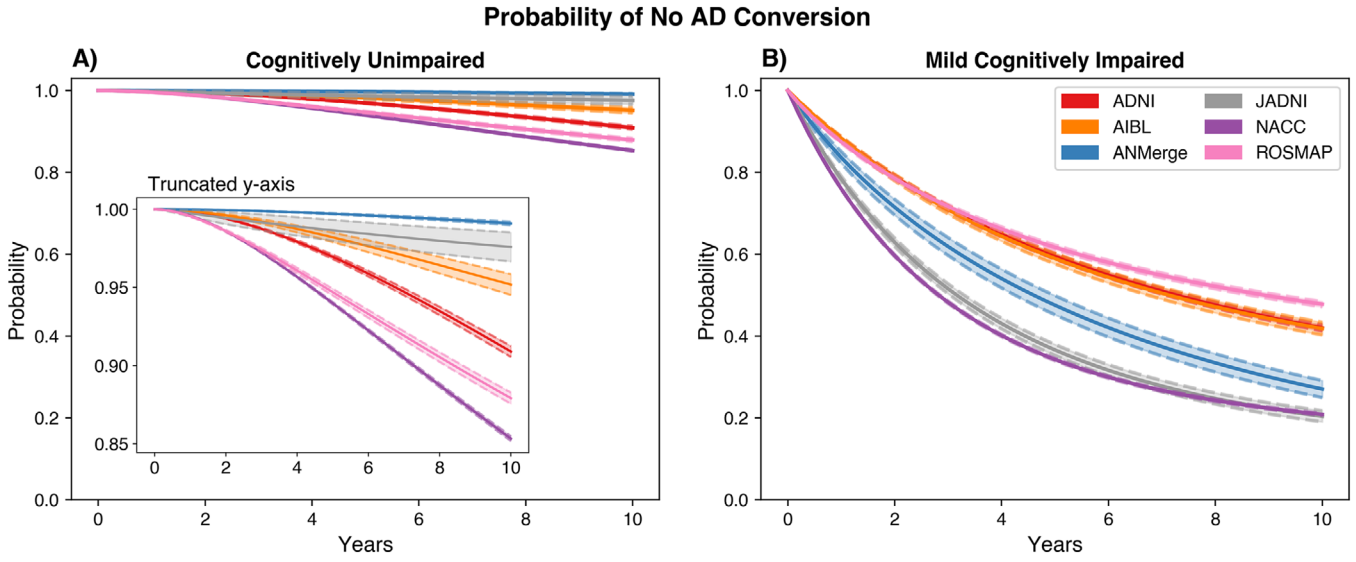


FIGURE 2 Average probability of staying AD diagnosis free over time for each cohort. Dashed lines indicate the standard errors of the estimates. A, Starting from cognitively unimpaired. B, Starting from mild cognitive impairment. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

Additionally, compared to the other cohorts, ROSMAP showed a considerably higher median probability to revert from MCI back to CU of 23%. All pairwise differences across cohorts proved to be significant ($P < .001$). Numerical values for the transition probabilities are presented in Table S4 in supporting information.

In concordance with the transition probabilities, the probability of staying AD diagnosis free over time differed substantially across cohorts. Starting in the CU state (Figure 2A), the trajectories of cohorts deviated significantly after approximately 4 years. NACC and ROSMAP exhibited the steepest decline (respectively, 85% and 87% after 10 years), while the probability for ANMerge stayed relatively stable (99%). Considering the MCI state as a starting point, the probability of remaining AD diagnosis free exhibited a steeper decline (Figure 2B). After 10 years, the most extreme estimates were made for ROSMAP (48%) and J-ADNI (20%), while no significant differences were observed between J-ADNI and NACC (both 20%), as well as between AIBL and ADNI (both 42%). Ultimately, we discovered a maximum deviation of 14% for the CU state and 28% for the MCI state.

All pairwise comparisons between the cohorts' sojourn time estimates turned out to be significant for the CU state ($P < .001$, with exception of ADNI–ROSMAP, $P < .05$; Figure 3A). Given their respective MSMs, ROSMAP displayed the shortest sojourn time with a median of 27.5 years, followed by ADNI (29.7 years), NACC (38.7 years), AIBL, ANMerge, and J-ADNI (all > 100 years). In the MCI state, again, most deviations were found to be significant ($P < .001$; Figure 3B). The only exception to this was ANMerge, which did not differ significantly from ADNI ($P = .9$) and AIBL ($P = .88$). The median sojourn time in the MCI state showed relatively lower values for J-ADNI (3.8 years) and NACC (3.1 years), while ADNI, AIBL, and ANMerge showed relatively higher values (7.7, 6.5, and 6.9 years, respectively). ROSMAP is placed in between with a median of 5 years. Detailed descriptions of the sojourn times distributions can be found in Table S5 in supporting information.

3.2 | Comparison of cohort-specific models

In the second set of analyses, we explored the cohort-specific biases learned by our MSMs from their respective training datasets. We observed that the cohort-specific models learned significantly different relationships between covariate values and the disease progression. Non-overlapping CIs indicated significant differences in hazard ratios for the transition from CU to MCI between ROSMAP (CI: 1.05 to 1.1), NACC (1.0 to 1.04), and ADNI (0.86 to 0.99) regarding education level. With respect to the MMSE, significant differences were found for ROSMAP, NACC, J-ADNI, and ADNI (CIs: 0.60 to 0.67, 0.76 to 0.81, 0.11 to 0.58, and 0.76 to 0.98, respectively; Figure 4A). The influence of education in J-ADNI (CI: 1.15 to 1.92) differed significantly from ADNI (0.93 to 1.12), NACC (0.94 to 1.04), and ROSMAP (0.93 to 1.03) with respect to reverting from MCI to CU (Figure 4B). Regarding the conversion from MCI to AD, significant differences were discovered in the hazard ratios for age between ROSMAP (1.02 to 1.05) and NACC (1.00 to 1.01), for *APOE* ε4 status between NACC (1.10 to 1.31) and ADNI (1.34 to 1.82), and for MMSE between NACC (0.83 to 0.87), ADNI (0.7 to 0.76), and ROSMAP (0.74 to 0.79; Figure 4C). In several cases, large CIs hampered the interpretation of the hazard ratios. The exact estimates of all hazard ratios are presented in Table S6 in supporting information.

When applying each MSM to the same set of data, the difference in the estimated progression patterns across models resembled the consequences of the learned cohort-specific biases (Figure 5). Numerical descriptions of the distributions in Figure 5 can be found in

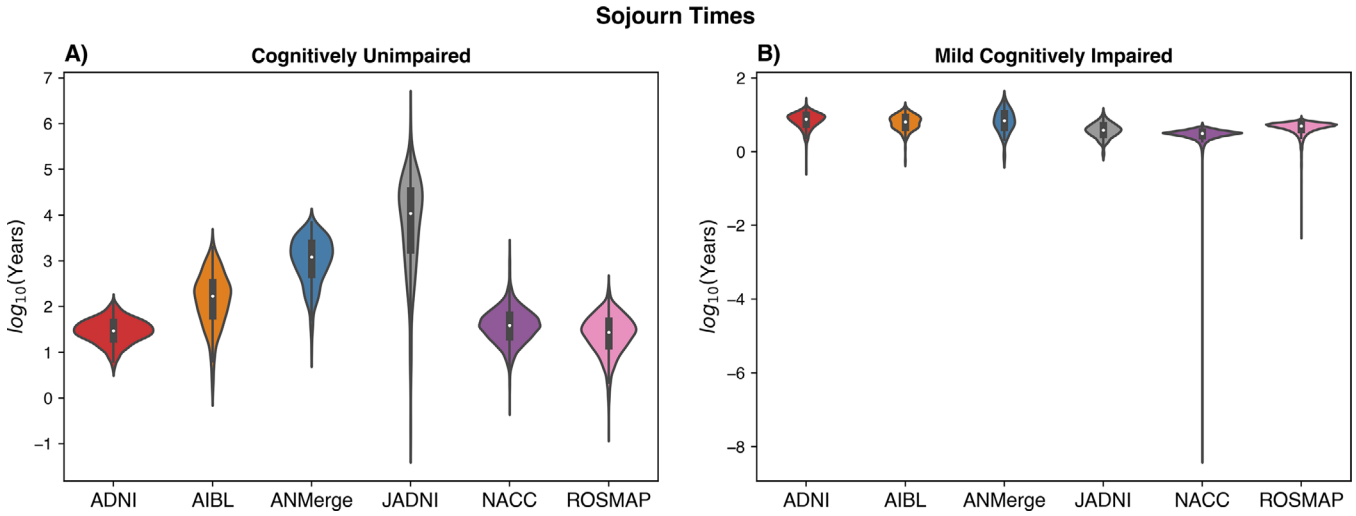


FIGURE 3 Sojourn times of cohort participants on a log10-scale. Because most deviations between depicted distributions were significant, we omit indication of significance for brevity and refer to the text. A, Occupying the cognitively unimpaired state. B, Occupying the mild cognitive impairment state. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

Tables S6 and S7 in supporting information. For all evaluated patterns (i.e. the transition probabilities, Figure 5A; sojourn times, Figure 5B; and estimated probability of staying AD diagnosis free, Figure 5C), significant Kruskal-Wallis tests underlined the presence of cohort-specific effects ($P < .001$). Additional pairwise comparisons using Mann-Whitney U tests are presented in the supporting information. We observed that naive pooling of datasets and training models on a combination of multiple, complete cohorts expectedly biases the estimates toward the cohort with the largest sample size (Figure S4 in supporting information).

We also found differences between cohorts when extracting progression patterns for a cohort's representative individual (Figure S5 in supporting information) and even when applying the same exemplary patients to each cohort's specific MSM (Figure S6 in supporting information).

3.3 | Clustering reveals overall similarity of studies

Figure 6 presents the results achieved by clustering the investigated cohorts based on the similarity of their progression patterns. ANMerge, AIBL, and NACC displayed close proximity indicating that their participants exhibited similar disease progression in combination with their trained MSMs. Furthermore, ADNI and J-ADNI formed a cluster that connected with the previously mentioned cluster in relatively high distance. ROSMAP was placed far from all other cohorts, constituting its own cluster.

4 | DISCUSSION

In this work, we explored the heterogeneity in AD progression across multiple, independent cohort datasets and the implications for data-driven approaches for progression modeling. Evident differences in mined progression patterns surfaced between six investigated cohorts. This finding raises concerns regarding the reliability of results discovered in single data resources and underlines the need for external validation. Furthermore, we demonstrated that models learn cohort-specific effects from their training dataset, which can impede model generalization. Last, we proposed a novel approach to identify similar cohort datasets that could help to find datasets that come closer to fulfilling the i.i.d. assumption. We demonstrated this approach by highlighting how six major AD cohorts relate to each other with regard to their exhibited disease progression.

4.1 | Progression trends differ across cohort datasets

Analyzing the characteristic progression trends extracted from the investigated cohorts revealed substantial differences among them. The observation of lower variability in estimates for the CU state compared to the MCI state can be explained by the fact that only a fraction of the CU participants will eventually develop cognitive symptoms. Thus, a substantial amount of CU participants are expected to show no signals of AD progression at all. Overall, the discovered heterogeneity could likely stem from differences in the recruitment processes of cohort studies. Compositional shifts across sampled populations pose a critical confounder comparing cohort datasets and model performance.[13] Here, statistical matching could potentially help to identify comparable subsets.

4.2 | Data-driven models learn systematic biases present in cohort datasets

Using all cohort-specific MSMs to estimate progression patterns for the same set of participants revealed the presence of strong

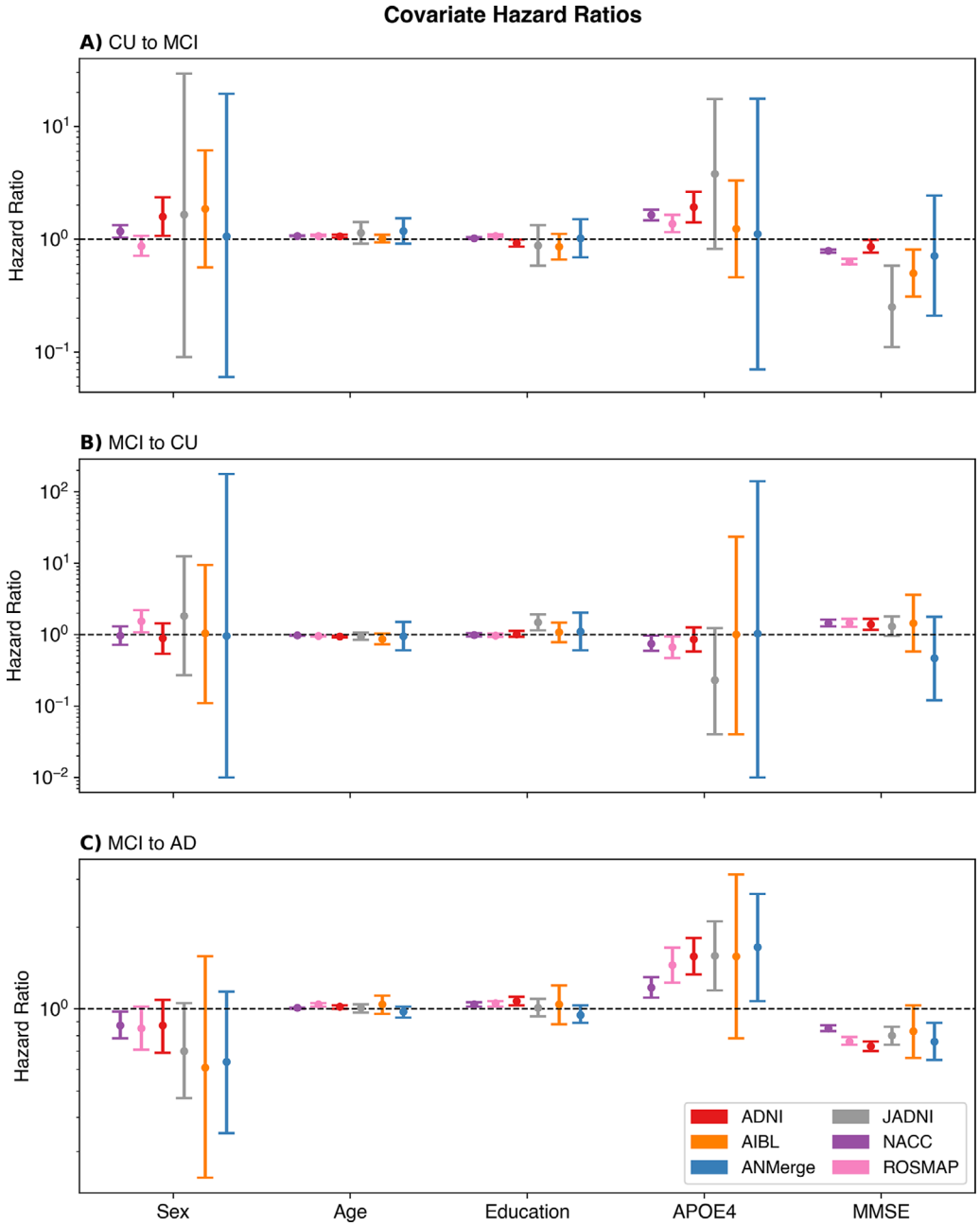
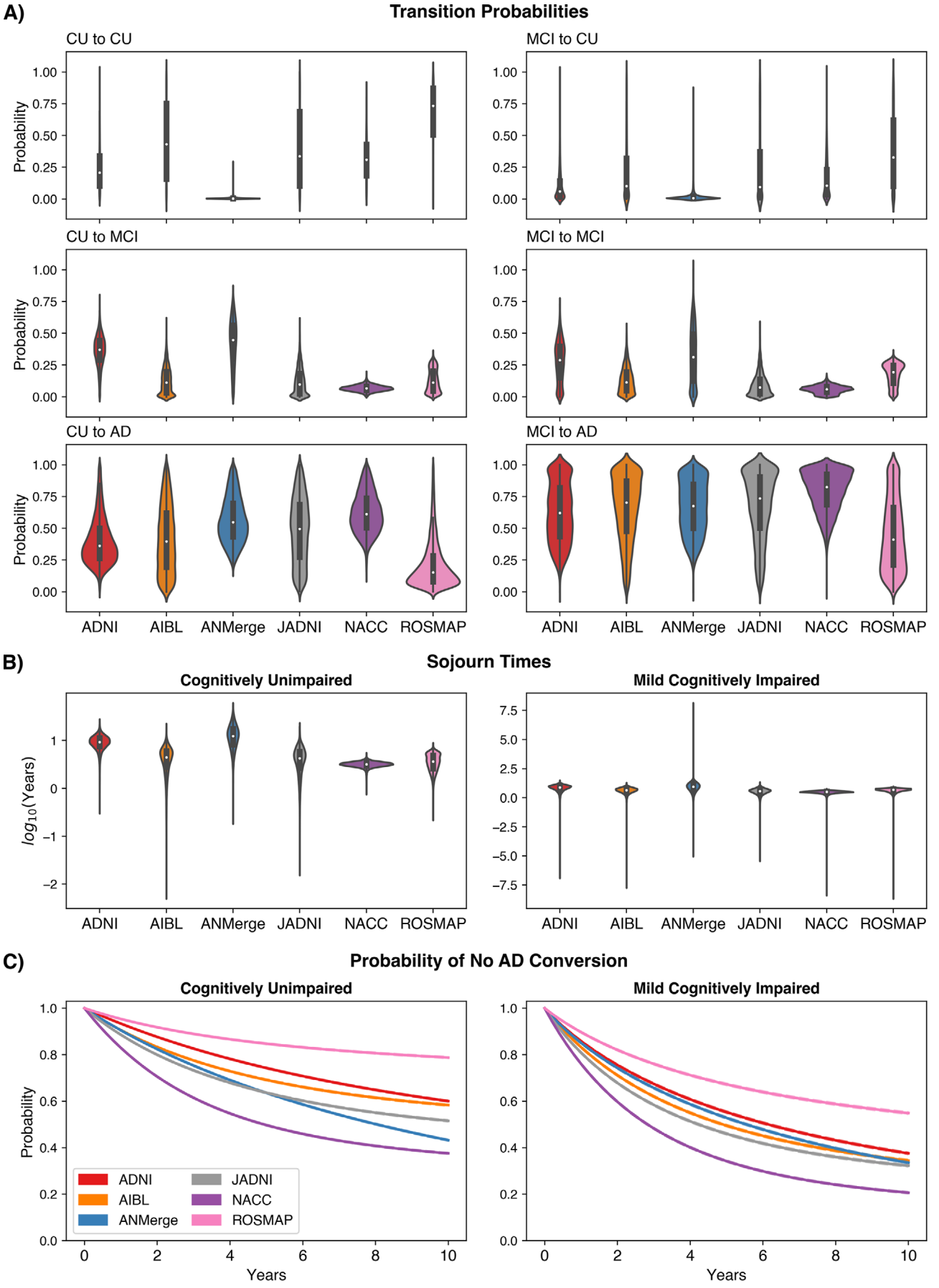


FIGURE 4 Covariate hazard ratios learned per cohort-specific multi-state models. For readability, significant deviations are not indicated visually. Instead, we refer to the text for the corresponding evaluations. A, B, C, Impact on transition from cognitively unimpaired (CU) to mild cognitive impairment (MCI), MCI to CU, and MCI to Alzheimer's disease (AD), respectively. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

cohort-specific effects that the models learned from their training datasets. The estimated covariate hazard ratios are an integral component of the cohort-specific progression signals and while we could observe commonalities in the directional influence of covariates, partially described by previous studies as well,[7,21] the magnitude of these influences exposed several significant differences. With regard to education, even contradicting influences were found. Differences in such fundamental parameters of a model propagate into, and thereby bias, their estimates; this became apparent in the subsequently estimated progression patterns.

Naive pooling of data from several cohorts does not necessarily pose a solution for addressing the biases but leads to an overshadowing of signals in smaller cohorts by larger ones. Instead, more considerate methods must be applied, such as sampling the same number of participants from each cohort, weighting of subjects to favor smaller datasets, or ensemble techniques that combine


A)
Transition Probabilities
CU to CU
MCI to CU
CU to MCI
MCI to MCI
CU to AD
MCI to AD
Probability
1.00
0.75
0.50
0.25
0.00
ADNI
AIBL
ANMerge
JADNI
NACC
ROSMAP
B)
Sojourn Times
Cognitively Unimpaired
Mild Cognitively Impaired
log10(Years)
1
0
−1
−2
7.5
5.0
2.5
0.0
−2.5
−5.0
−7.5
C)
Probability of No AD Conversion
Cognitively Unimpaired
Mild Cognitively Impaired
Probability
1.0
0.8
0.6
0.4
0.2
0.0
0
2
4
6
8
10
Years
ADNI
AIBL
ANMerge
JADNI
NACC
ROSMAP

dataset-specific models. Future work should explore these options in more detail.

4.3 | Clustering allows assessment of cohort similarities

Our proposed approach to measure cohort similarity with regard to their global disease progression trends (informed by neuropsychological tests, biological sex, completed years of education, *APOE* ε4 status) elicited commonalities across cohorts that mirror the design of these studies. Finding ADNI and J-ADNI in one cluster together is reassuring as J-ADNI was designed as a complementary cohort to ADNI, and similar trends have been observed in both cohorts.[28] Their use of equal eligibility criteria for participant recruitment counteracts the risk of sampling from two distinct populations. The distance we observe between them could be explained partially due to differences in ethnoracial composition[31] and lifestyle.[32] ROSMAP, on the other hand, is a special case in the landscape of AD cohorts. Its participants are exclusively recruited from religious orders, are considerably older, and hold a higher proportion of female participants compared to the other cohorts.[13,30]

Our proposed method enables a quantitative description of differences across cohorts and, subsequently, an evaluation of cohort similarity based not only on cross-sectional values of covariates but on their general progression. Consequently, it could help researchers to better understand and characterize performance measures obtained during the external validation of machine learning models. More specifically, our cohort clustering can be used post hoc to indicate whether failed validation was likely caused by overfitting or systematic biases between discovery and validation cohort originating from, for example, sampling of distinct statistical distributions.

4.4 | Limitations

It is unknown how many of the CU participants per cohort would have eventually developed cognitive symptoms during their lifetime. While the models account for this factor using censoring, estimates based on the CU participants could be biased depending on the size of the participant fraction with prodromal AD.

One limitation of MSMs is the assumption that disease progression depends only on the current state of a participant. While this is a necessary and widely accepted assumption in the literature,[7,21–24] there is no universal way to prove that it always holds true for all possible state transitions.

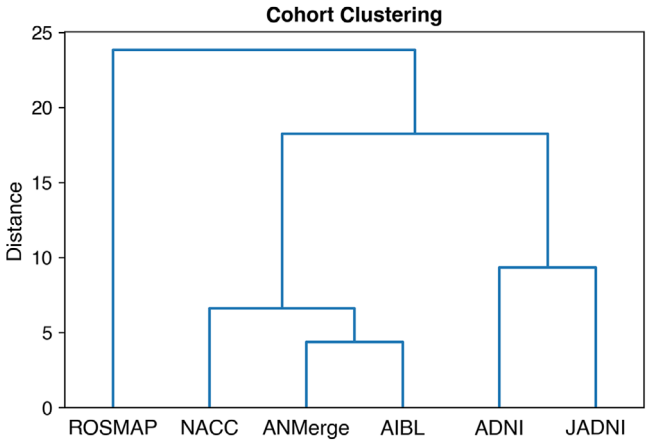


FIGURE 6 Cohort dendrogram resulting from the clustering of pairwise log-likelihoods. ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

In recent years, AD is more considered a biological entity[1] and while we aimed to account for as many clinically relevant covariates as possible, we were unable to include emerging biomarkers in our MSMs. Given the limited number of individuals participating in longitudinal biomarker collection, the inclusion of biomarkers would have led to underpowered models and reduced the number of cohorts available for analysis. However, using this limited set of covariates, our model selection showed that all chosen covariates added meaningful information to the models and that progression signals could successfully be learned.

5 | CONCLUSION

Applying machine learning and statistical modeling to single data resources can bias results and might render the generalizability of the models used infeasible. Ideally, it would be imperative that we go beyond single data resources and instead investigate and validate findings across the landscape of AD data we have at our disposal. In practice, however, external validation of data-driven machine learning models is often limited by the availability of semantically and statistically comparable datasets.[13] For some investigations only single cohorts might be suitable. While results originating from such single-cohort investigations hold value as initial indications, they should be (1) regarded as cohort-specific findings pending external validation, and

FIGURE 5 Consequences of learned cohort-specific biases onto estimated progression patterns. The same set of participants was considered under each cohort's trained multi-state models (i.e., variability in estimates stems from the models, not the data). Deviations between estimates illustrate the learned biases. Because most deviations between depicted distributions were significant, we omit indication of significance for brevity. AD, Alzheimer's disease; ADNI, Alzheimer's Disease Neuroimaging Initiative; AIBL, Australian Imaging, Biomarker & Lifestyle Flagship Study of Ageing; ANMerge, AddNeuroMed; CU, cognitively unimpaired; J-ADNI, Japanese Alzheimer's Disease Neuroimaging Initiative; MCI, mild cognitive impairment; NACC, National Alzheimer's Coordinating Center; ROSMAP, Religious Orders Study and Rush Memory and Aging Project

(2) meticulously validated internally. Here, resampling techniques and cross-validation can help to increase the robustness of single cohort studies.[8]

Dealing with such heterogeneous data as is encountered in our field, building a single model that serves all predictive purposes and is applicable to the general AD population seems inconceivable. Instead, the more promising alternative to support clinical decision-making using data-driven approaches for AD and dementia could be to build subpopulation-specific models that embrace the specifics of their target group. Here, the stratification of the AD population into specific progression subtypes could guide which model is applicable to which patient. Alternatively, artificial intelligence methods from the field of domain adaptation (e.g., transfer learning) might help to manage the heterogeneous signals when applying models across cohorts.

ACKNOWLEDGMENTS

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

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 were 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). This project has received funding from the European Union's Horizon 2020 research and innovation program under grant agreement No. 826421, "TheVirtualBrain-Cloud."

CONFLICTS OF INTEREST

The authors have nothing to declare.

AUTHOR CONTRIBUTIONS

Colin Birkenbihl and Holger Fröhlich designed the study. Yasamin Salimi and Colin Birkenbihl implemented the methods and ran the experiments. Colin Birkenbihl wrote the manuscript. Holger Fröhlich and Yasamin Salimi revised the manuscript. Holger Fröhlich supervised the project.

REFERENCES

1. Jr Jack RC, Bennett DA, Blennow K, et al. NIA-AA research framework: toward a biological definition of Alzheimer's disease. *Alzheimers Dement*. 2018;14(4):535-562.
2. Jr Jack RC, Knopman DS, Jagust WJ. et al. Tracking pathophysiological processes in Alzheimer's disease: an updated hypothetical model of dynamic biomarkers. *Lancet Neurol*. 2013;12(2):207-216.
3. Winblad B, Amouyel P, Andrieu S, et al. Defeating Alzheimer's disease and other dementias: a priority for European science and society. *Lancet Neurol*. 2016;15(5):455-532.
4. Sperling RA, Jack CR, Aisen PS. Testing the right target and right drug at the right stage. *Sci Transl Med*. 2011;3(111):111cm33-111cm33.
5. Hadjichrysanthou C, Evans S, Bajaj S, et al. The dynamics of biomarkers across the clinical spectrum of Alzheimer's disease. *Alzheimers Res Ther*. 2020;12(1):1-16.
6. de Jong J, Emon MA, Wu P, et al. Deep learning for clustering of multivariate clinical patient trajectories with missing values. *Gigascience*. 2019;8(11):giz134.

7. Vermunt L, Sikkes SA, Van Den Hout A, et al. Duration of preclinical, prodromal, and dementia stages of Alzheimer's disease in relation to age, sex, and APOE genotype. *Alzheimers Dement*. 2019;15(7):888-898.
8. Fröhlich H, Balling R, Beerenwinkel N, et al. From hype to reality: data science enabling personalized medicine. *BMC Med*. 2018;16(1):150.
9. Golriz Khatami S, Robinson C, Birkenbihl C, Domingo-Fernández D, Hoyt CT, Hofmann-Apitius M. Challenges of integrative disease modeling in Alzheimer's disease. *Front Mol Biosci*. 2020;6:158.
10. Ryan J, Fransquet P, Wrigglesworth J, Lacaze P. Phenotypic heterogeneity in dementia: a challenge for epidemiology and biomarker studies. *Front Public Health*. 2018;6:181.
11. Habes M, Grothe MJ, Tunc B, McMillan C, Wolk DA, Davatzikos C. Disentangling heterogeneity in Alzheimer's disease and related dementias using data-driven methods. *Biol Psychiatry*. 2020.
12. Jacobs D, Sano M, Marder K, et al. Age at onset of Alzheimer's disease: relation to pattern of cognitive dysfunction and rate of decline. *Neurology*. 1994;44(7):1215-1215.
13. Birkenbihl C, Salimi Y, Domingo-Fernándéz D, et al. Evaluating the Alzheimer's disease data landscape. *Alzheimers Dement*. 2020;6(1):e12102.
14. 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.
15. Vapnik V. *Statistical Learning Theory*. New York: Wiley; 1998:624.
16. Ben-David S, Blitzer J, Crammer K, Pereira F, (2007). Analysis of representations for domain adaptation. *In advances in neural information processing systems* (pp. 137-144). MIT press.
17. Sun B, Feng J, Saenko K, (2016). Return of frustratingly easy domain adaptation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence* (pp. 2058-2065). AAAI Press.
18. 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.
19. 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.
20. Jackson CH. Multi-state models for panel data: the msm package for R. *J stat softw*. 2011;38(8):1-29.
21. Robitaille A, van den Hout A, Machado RJ, et al. Transitions across cognitive states and death among older adults in relation to education: a multistate survival model using data from six longitudinal studies. *Alzheimers Dement*. 2018;14(4):462-472.
22. Zhang L, Lim CY, Maiti T, et al. Analysis of conversion of Alzheimer's disease using a multi-state markov model. *Stat Methods Med Res*. 2019;28(9):2801-2819.
23. Brookmeyer R, Abdalla N. Estimation of lifetime risks of Alzheimer's disease dementia using biomarkers for preclinical disease. *Alzheimers Dement*. 2018;14(8):981-988.
24. Jr Jack CR, Therneau TM, Wiste HJ, et al. Rates of transition between amyloid and neurodegeneration biomarker states and to dementia among non-demented individuals: a population-based cohort study. *Lancet Neurol*. 2016;15(1):56.
25. Mueller SG, Weiner MW, Thal LJ, et al. Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). *AlzheimersDement*. 2005;1(1):55-66.
26. Birkenbihl C, Westwood S, Shi L, et al. ANMerge: a comprehensive and accessible Alzheimer's disease patient-level dataset. *J Alzheimers Dis*. 2021;79(1):423-431.
27. 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.
28. Iwatsubo T, Iwata A, Suzuki K, et al. Japanese and North American Alzheimer's disease neuroimaging initiative studies: harmonization for international trials. *Alzheimers Dement*. 2018;14(8):1077-1087.
29. 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.
30. 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.
31. 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*. 2019;15(2):292-312.
32. Xu W, Tan L, Wang HF, et al. Meta-analysis of modifiable risk factors for Alzheimer's disease. *J Neurol, Neurosurg Psychiatry*. 2015;86(12):1299-1306.

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, Fröhlich H. Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimer's Dement*. 2022;18:251–261. https://doi.org/10.1002/alz.12387