BRAIN COMMUNICATIONS

Deep learning-based patient stratification for prognostic enrichment of clinical dementia trials

Colin Birkenbihl,[1,2,3] Johann de Jong,[4] Ilya Yalchyk[1,2] and Holger Fröhlich[1,2];
for the Alzheimer's Disease Neuroimaging Initiative*

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

Dementia probably due to Alzheimer's disease is a progressive condition that manifests in cognitive decline and impairs patients' daily life. Affected patients show great heterogeneity in their symptomatic progression, which hampers the identification of efficacious treatments in clinical trials. Using artificial intelligence approaches to enable clinical enrichment trials serves a promising avenue to identify treatments. In this work, we used a deep learning method to cluster the multivariate disease trajectories of 283 early dementia patients along cognitive and functional scores. Two distinct subgroups were identified that separated patients into 'slow' and 'fast' progressing individuals. These subgroups were externally validated and independently replicated in a dementia cohort comprising 2779 patients. We trained a machine learning model to predict the progression subgroup of a patient from cross-sectional data at their time of dementia diagnosis. The classifier achieved a prediction performance of 0.70 ± 0.01 area under the receiver operating characteristic curve in external validation. By emulating a hypothetical clinical trial conducting patient enrichment using the proposed classifier, we estimate its potential to decrease the required sample size. Furthermore, we balance the achieved enrichment of the trial cohort against the accompanied demand for increased patient screening. Our results show that enrichment trials targeting cognitive outcomes offer improved chances of trial success and are more than 13% cheaper compared with conventional clinical trials. The resources saved could be redirected to accelerate drug development and expand the search for remedies for cognitive impairment.

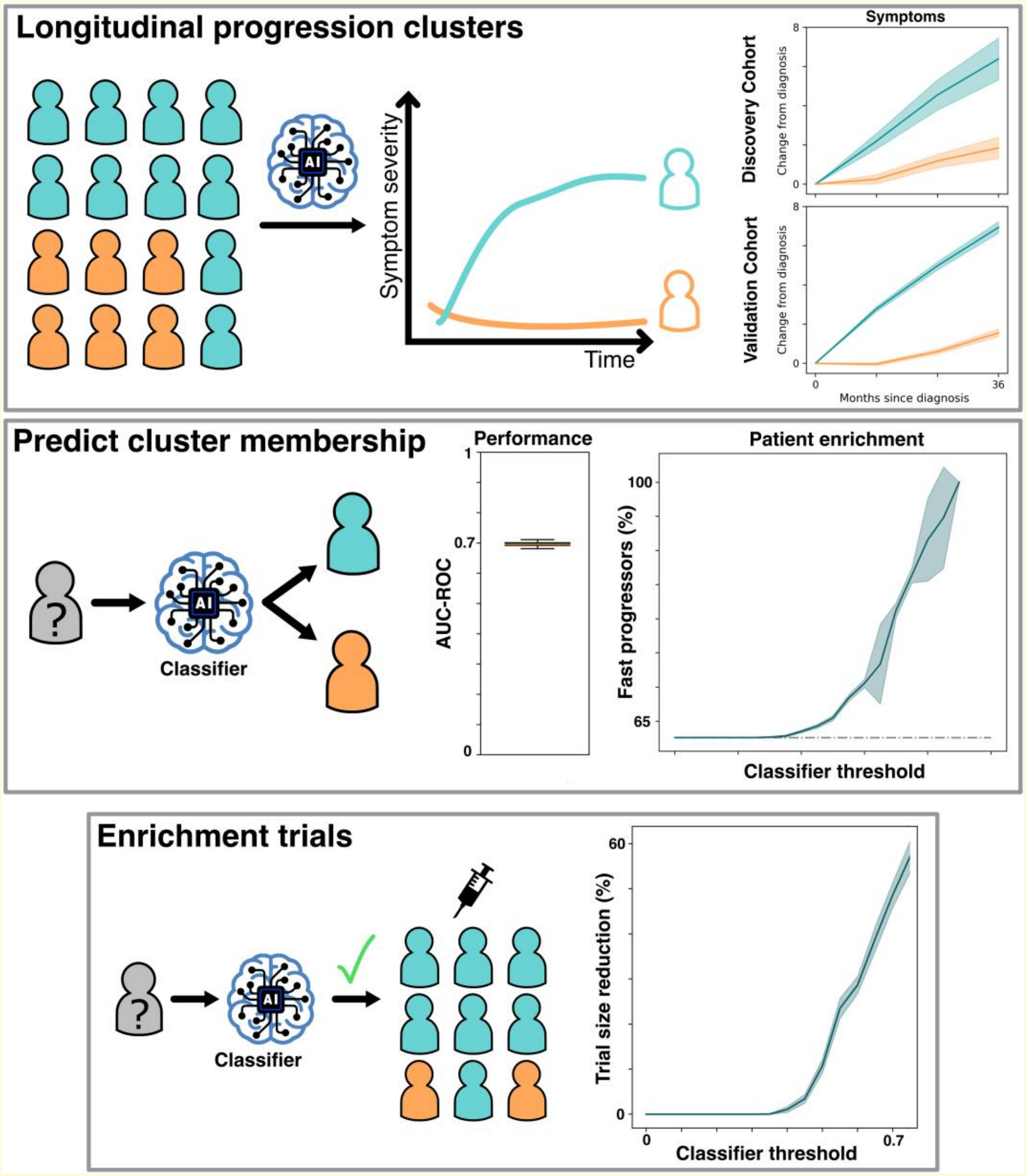
1 Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI), Sankt Augustin 53757, Germany
2 Bonn-Aachen International Center for IT, Rheinische Friedrich-Wilhelms-Universität Bonn, Bonn 53115, Germany
3 Department of Neurology, Massachusetts General Hospital, Harvard Medical School, Boston 02114, USA
4 Global Computational Biology and Digital Sciences, Boehringer Ingelheim Pharma GmbH & Co. KG, Ingelheim 55216, Germany

Correspondence to: Holger Fröhlich
Department of Bioinformatics, Fraunhofer Institute for Algorithms and Scientific Computing (SCAI)
Schloss Birlinghoven
Sankt Augustin, North Rhine-Westphalia 53757, Germany
E-mail: holger.froehlich@scai.fraunhofer.de

Keywords: patient stratification; clinical trial; deep learning; machine learning; prognostic enrichment

Received February 13, 2024. Revised August 20, 2024. Accepted December 05, 2024. Advance access publication December 16, 2024

Graphical Abstract



Introduction

Dementia is a debilitating, progressive condition that is primarily described by cognitive decline. It can be caused by multiple neurological diseases, with Alzheimer's disease causing 75% of all cases.[1] With increasing symptom severity, patients become impaired in their daily life and require full-time care, which poses a great burden to patients, caregivers and society. With the recent trials of aducanumab,[2] donanemab[3] and lecanumab[4] being the only successes in the past 20 years, most clinical trials aiming to identify treatments against cognitive decline in Alzheimer's disease have failed.[5,6]

The low success rate of clinical trials aimed at improving cognitive outcomes can be attributed, in part, to the significant variability in how patients' symptoms progress, even during the earliest stages of the disease.[6,7] This variability poses a statistical challenge and can impede the identification of significant treatment effects. One potential solution to this issue is to increase the sample size of clinical trials to enhance statistical power. However, this approach is costly as it requires more treated patients.

Alternatively, clinical trials can aim for a targeted recruitment of patients that will likely exhibit a faster disease progression and change in cognitive outcomes.[8] This enrichment of patients from a subgroup experiencing a more homogeneous, fast symptomatic progression represents a so-called enrichment trial[9] and promises several advantages over traditional clinical trials: It can unmask treatments that are only efficacious in specific patient subgroups but fail in the average population,[10] and recruited trial cohorts can be smaller due to the reduced heterogeneity and increase in effect size.[11] The advantages of enrichment trials have also been recognized and promoted by the US Food and Drug Administration in 2019. To enable enrichment trials, however, robust patient subgroups with distinct symptom progression patterns must be identified and validated.[12] Furthermore, it must be possible to predict the subgroup membership of an individual from cross-sectional data already available during patient screening. Also, in the light of the recent approvals of three monoclonal antibodies targeting the Alzheimer's disease characteristic amyloid beta pathology,[2-4] the timely prognosis of patients' likely course of cognitive decline would help to optimize the correct timing and intensity of treatment. Prognostic models designed for this task would also be of immense value from a patient perspective, because they would allow them to better plan their future.

One family of approaches that allows for identifying patient subgroups based on multimodal patient-level cohort data is clustering methods. In the context of Alzheimer's disease, such approaches were predominantly applied to cross-sectional data of patients.[13-15] However, cross-sectional clusterings fall short in capturing the longitudinal dynamics of Alzheimer's disease dementia, and resulting subgroups can be biased by the disease stages in which patients resided at the time of data collection. Recent studies that take progressive signals into account focussed mainly on exploring the pathology of the disease in the form of imaging biomarkers[16] rather than the cognitive outcomes directly relevant for clinical trials.

In this work, we apply an artificial intelligence approach for clustering multivariate clinical disease trajectories of demented Alzheimer's disease patients. The resulting subgroups are then externally validated and independently replicated in another cohort study. We construct and validate a machine learning classifier to accurately predict the future progression type of individuals from cross-sectional data only. Finally, we demonstrate the value our classifier could provide to enrichment trials for treatments of cognitive decline by enabling cheaper trials with smaller cohort sizes.

Materials and methods

Cohort data sets and patient selection

Two independent cohort data sets were used in this study: the Alzheimer's Disease Neuroimaging Initiative (ADNI)[17] and the National Alzheimer's Coordinating Center (NACC).[18] ADNI is an observational cohort study with predominantly White, highly educated participants.[19] NACC aggregates data from Alzheimer's Disease Research Centers across the USA and is more diverse and heterogeneous than ADNI.[20] Both cohort studies adhered to the Declaration of Helsinki and got approval from their institutional review boards. We only included patients who developed dementia probably due to Alzheimer's disease during the runtime of their respective cohort study. Furthermore, patients must have had at least one follow-up assessment after their dementia diagnosis and visits prior to it were excluded. This led to 283 analysable patients for ADNI and 2779 for NACC, with a median of 2-year follow-up, respectively. Three years after diagnosis, 70 patients were available in ADNI and 1080 patients in NACC.

Multivariate patient trajectory clustering

To cluster patients into symptom progression subgroups, we used our previously published VaDER approach that was specifically designed with longitudinal clinical data in mind.[21] VaDER is a deep learning approach that clusters multivariate time-series data and imputes missing values implicitly during model training. Hyperparameters were optimized following the procedure described in de Jong *et al*.[21]: We evaluated several possible models using different hyperparameter configurations (including the number of sought-after subgroups) and selected the hyperparameters of the best-performing model. Model performance was measured by comparing the prediction strength of the clustering induced by the trained model against a random clustering of the same data.[22] To determine the optimal number of clusters, we selected the smallest number that showed a significant difference from random clustering (Supplementary Fig. 1). Selected hyperparameters are presented in Supplementary Table 1.

We clustered patients based on their trajectories of three major clinical scores measuring symptom progression: the Mini Mental State Examination (MMSE), Clinical Dementia Rating Sub of Boxes (CDRSB) and Functional Activities Questionnaire (FAQ). These measures were chosen due to their shared availability in the analysed data sets and relevance for clinical trials targeting early Alzheimer's disease (Supplementary Table 2). Patient trajectories were aligned on their dementia diagnosis to avoid biases introduced through patients being in different clinical disease stages at their study baseline. Considering the average length of currently ongoing Phase 3 trials for cognitive treatments enrolling early Alzheimer's disease patients (~24 months; Supplementary Table 2) and the longitudinal follow-up of ADNI and NACC, we clustered trajectories spanning up to 3 years. Each clustering was repeated 40 times, and the final subgroup assignments of patients were based on a consensus clustering across the repeats.

Cluster validation and replication in external data

To evaluate the robustness and validity of our identified patient subgroups, we conducted an external validation of

ADNI-derived subgroups in NACC and, additionally, performed a replication of the analysis starting with NACC data. For the external validation, we applied the ADNI-trained clustering model to NACC to determine whether the resulting subgroups of NACC patients resembled those identified in ADNI. Furthermore, we aimed to replicate the results by starting with a new, independent clustering of NACC to see if we would get an optimal clustering that was similar to the one achieved under the ADNI-trained model. We further externally validated this NACC-derived clustering in ADNI. Finally, we assessed the concordance between patient assignments within each data set under both clustering models.

Building a machine learning classifier for cluster prediction

We built machine learning classifiers to predict the progression cluster membership of individuals. Only cross-sectional data available at the time of each patient‘s dementia diagnosis were included as predictors. We used the XGBoost algorithm that employs an ensemble of decision trees and can handle missing values.[23] The classifiers were trained and evaluated in a nested 8-fold cross-validation, which we repeated 10 times to ensure robust results. The hyperparameter optimization was performed in the inner 8-fold cross-validation and model evaluation occurred in the outer one (details in the Supplementary material).

As more data modalities were available in ADNI than in NACC,[24] we built two separate classifiers: a multimodal classifier based on the ADNI data and another classifier using only features common between ADNI and NACC (in the latter referred to as ‘common predictor’ classifier). The ‘common predictor’ version of the classifier allowed for an external validation.

The multimodal classifier incorporated demographic information (7 features), clinical assessments and their subscores (11 assessments amounting to 66 features in total), biomarkers (62 magnetic resonance imaging-derived brain region volumes, 3 positron emission tomography and 4 cerebrospinal fluid], and genetic variables (*APOE* ϵ4 status, 75 disease pathway perturbation scores; see Supplementary material for details on pathway score calculation and Supplementary Data Set 1 for a list of all predictors). Individual MMSE questions were summed into subscores as described in the Supplementary material. Due to the increased requirements on the available data compared with the initial clustering, the sample size of ADNI was reduced to 230 patients for this analysis.

For the ‘common predictor’ version of the classifier, the number of available features decreased to 28, now comprising only *APOE* ϵ4 status, clinical and demographic features (1, 22 and 5 features, respectively). The clinical features are the Trail Making B score, Montreal Cognitive Assessment score, Digit Span score and summary scores and subscores of the MMSE, Clinical Dementia Rating and FAQ. A detailed list is provided in the Supplementary Data Set 1. This classifier was trained on the larger NACC data set and externally validated on ADNI.

Simulating the impact of patient enrichment on clinical trial design

We estimated a potential reduction in trial cohort sample size enabled through an enrichment of patients with ‘fast’ symptom progression while maintaining adequate statistical power. This analysis was performed by applying the NACC-trained ‘common predictor’ classifier to ADNI to mirror a scenario with classifier-independent data. Stratified patient recruitment was simulated by only including patients into our hypothetical trial cohort whose predicted probability of belonging to the ‘fast’ progressing cluster exceeded a threshold. The specifications of the hypothetical trial were adapted from recent clinical trials for early to mild Alzheimer’s disease dementia (Supplementary Table 2): As a primary outcome, we used the change from baseline in CDRSB since dementia diagnosis and considered a trial runtime of up to 24 months. The treatment arm was simulated using an effect size of 27%, a value that was observed for the recently approved lecanemab.[25] Effect size was calculated as Cohen’s *d*. Conservatively, we did not simulate the effect size dependent on the progression rate, but uniform over all patients. Effects were only emulated in patients who actually experienced a worsening of the outcome during the 24 months. Outcomes for patients showing improvement or no change remained unaltered. The theoretical control arm consisted of the same patients without simulated treatment effects. For a power analysis, we considered a two-sided *t*-test. Following the lecanemab trial, the required statistical power was considered 90% at an alpha level (Type I error) of 0.05.[4]

We approximated the impact of reducing the trial sample size through patient enrichment in terms of adverse events and monetary expenses using the ‘Clarity AD’ trial for lecanemab as guidance.[4] We ignored patient dropout in our estimations. Annual treatment costs of lecanemab amount to $26 500 per patient.[26] As exact information about the costs of patient screening in the lecanemab trial was missing, we assumed the same costs that were previously estimated for the aducanumab trial with $6957 per screened patient.[27]

Adverse events were simulated based on their frequencies of occurrence observed in the original lecanemab trial.[4] Amyloid-related imaging abnormality (ARIA) diagnosis and monitoring incurrences were assumed to involve an additional physician visit ($128) and monthly magnetic resonance imaging ($353 per scan)[27] for a mean duration of 4 months until resolvement.[4]

Statistical analysis

We assumed a significance level of 0.05. Where appropriate, 95% confidence intervals (CIs) were calculated to illustrate the uncertainty of estimates.

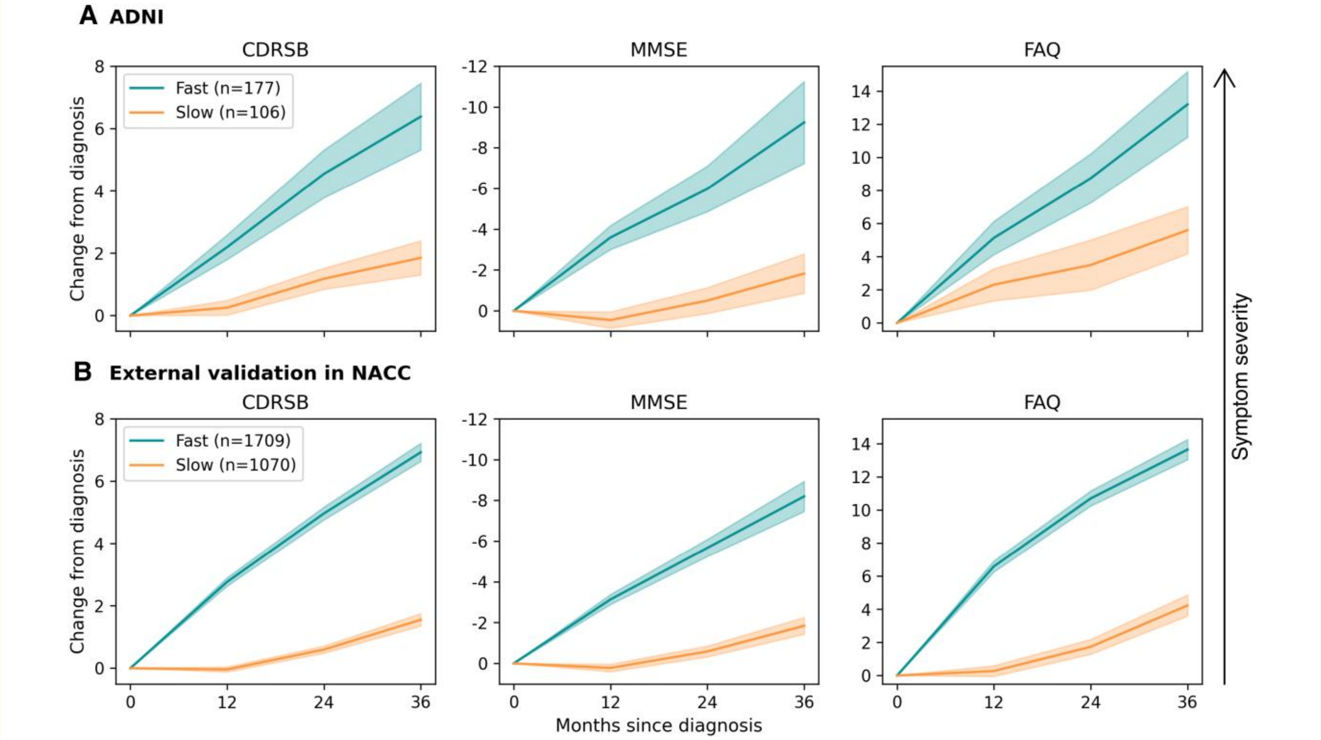


Figure 1 Symptom progression trajectories of identified subgroups. Average symptom progression trajectories of subgroups identified in ADNI (**A**) and NACC (**B**) under the ADNI-trained clustering model. For each clinical assessment, the severity of the symptom increases along the *y*-axis from bottom to top.

Results

Identification, validation and replication of two Alzheimer's disease dementia progression subtypes

When clustering the ADNI patients' trajectories, we identified two distinct symptom progression subgroups that separated patients into 'slow' and 'fast' progressors (Fig. 1A; Supplementary Figs 1A and 3). One hundred seventy-seven of the 283 patients (63%) were assigned to the 'fast' progressing cluster and 106 (37%) to the 'slow' progressors. Over the 36-month period, 'fast' progressing patients experienced symptom worsening of 6.38 [95% CI: (5.32, 7.45)] for CDRSB, −9.24 [95% CI: (−7.23, −11.25)] for MMSE and 13.19 [95% CI: (11.22, 15.17)] for FAQ. In contrast, on average, 'slow' progressing patients showed significantly reduced worsening with 1.85 [95% CI: (1.30, 2.40)], 1.83 [95% CI: (0.87, 2.78)] and 5.59 [95% CI: (4.17, 7.01)], for CDRSB, MMSE and FAQ, respectively.

When externally validating the clustering achieved in ADNI by applying the ADNI-trained model to patient trajectories from NACC, we obtained two subgroups of NACC patients that were highly similar to those identified in ADNI (Fig. 1B). Matching the proportions in ADNI closely, ~61% (1709) of the NACC patients exhibited a 'faster' progression, while 39% (1070) experienced 'slower' symptom progression. Also, the observed empirical average trajectories of NACC subgroups were similar to those identified in ADNI. On average, the 'fast' progressors showed symptom worsening of 6.93 [95% CI: (6.63, 7.22)] for CDRSB, −8.20 [95% CI: (−7.46, −8.94)] for MMSE and 13.64 [95% CI: (13.04, 14.25)] for FAQ over 36 months. 'Slow' progressing patients symptoms increased by 1.55 [95% CI: (1.35, 1.75)], −1.84 [95% CI: (−1.43, −2.25)] and 4.22 [95% CI: (3.59, 4.86)], for CDRSB, MMSE and FAQ, respectively.

Beyond externally validating the ADNI clustering in NACC, we investigated whether the clustering of NACC under the ADNI-trained model would be concordant with an independent clustering achieved by training a new model on NACC. Indeed, a two-subgroup partition provided the best clustering solution for NACC, again splitting patients into 'fast' and 'slow' progressors (Supplementary Fig. 1B). Comparing the subgroup assignment of NACC patients into 'fast' or 'slow' progressors under the NACC-trained model and ADNI-trained model showed an agreement of 88%. We additionally applied the NACC-trained model on ADNI for external validation. Again, a highly similar clustering was found with 82% of the ADNI patients assigned to the same subgroup using the independent ADNI-trained and NACC-trained models, respectively.

Characterization of symptom progression subgroups

We compared demographic variables (age, education and biological sex) between 'slow' and 'fast' progressors and found no statistically or clinically significant differences at the time of dementia diagnosis in ADNI (Table 1). In NACC, a statistically significant difference was identified for patient age, however, of insignificant clinical relevance [1.01 years, 95% CI: (0.31, 1.70)]. Furthermore, in ADNI, we observed a statistically significant but small difference at patient's dementia diagnosis for the CDRSB score [0.49, 95% CI: (0.05, 0.94)]. In NACC, FAQ and MMSE scores at diagnosis differed statistically significantly across subgroups [−1.11 (−1.49, −0.73) and 1.86 (0.99, 2.72), respectively], but once again with small effect sizes. A further significant difference was found in the distribution of *APOE* ϵ4 carriers across NACC subgroups, with 9.33% (13.1%, 5.53%) more ϵ4 carriers being assigned to the 'fast' progressing group.

Comparison of cerebrospinal fluid biomarkers of Alzheimer's disease pathology, namely amyloid beta 42, phosphorylated tau and total tau, did not reveal any significant differences between the two clusters in neither ADNI nor NACC (Mann–Whitney *U* test $P > 0.05$ for all biomarkers in both cohorts). Also, regarding amyloid PET, no significant difference was identified ($P > 0.05$).

Predicting symptom progression subtype from cross-sectional data at time of diagnosis

The multimodal machine learning classifier trained on ADNI was able to differentiate between 'slow' and 'fast' progressing patients with an average area under the receiver operating characteristic (AUC) of 0.69 ± 0.02 estimated via a 10 times repeated cross-validation (Fig. 2) and an area under the precision-recall curve (AUC-PR) of 0.60 ± 0.03.

To externally validate our classifier, we developed a second version that only incorporated features present in both ADNI and NACC. Since NACC holds a substantially larger sample size, we used this data set to train and internally validate the classifier and used ADNI for external validation. In internal validation on NACC, the classifier achieved 0.67 ± 0.003 AUC (Fig. 2) and an AUC-PR of 0.58 ± 0.003. External validation in ADNI demonstrated a performance of 0.70 ± 0.01 AUC and 0.56 ± 0.01 AUC-PR, which was similar to both the multimodal classifier's performance in ADNI and the model's internal validation scores on NACC (Fig. 2) and thereby indicated the model's generalizability. Feature importance is shown in Supplementary Fig. 2.

Table 1 Summary statistics describing the empirical distribution of clinical and demographic variables in symptom progression subtypes identified in ADNI and NACC at time of patient's dementia diagnosis

	N	MMSE	FAQ	CDRSB	Age (years)	Education (years)	APOE ϵ4 positive (%)	Female (%)
ADNI clusters								
Slow	106	24.21 ± 2.71	11.46 ± 5.9	3.98 ± 1.36	75.32 ± 6.95	15.91 ± 2.73	67.92	25.47
Fast	177	23.55 ± 3.33	12.79 ± 5.62	4.48 ± 1.81	75.53 ± 7.04	15.9 ± 2.88	76.84	36.16
Difference (95% CI)		−0.66 (−1.5, 0.17)	1.33 (−0.22, 2.88)	**0.49 (0.05, 0.94)**	0.2 (−1.68, 2.09)	0.0 (−0.76, 0.75)	−8.91 (−19.84, 1.65)	10.69 (−0.58, 21.04)
NACC clusters (predicted using the ADNI-trained model)								
Slow	1078	25.02 ± 4.12	14.57 ± 12.44	3.75 ± 3.19	77.68 ± 9.12	15.64 ± 3.22	48.42	51.58
Fast	1716	23.91 ± 4.2	16.43 ± 10.31	3.88 ± 2.29	78.68 ± 9.18	15.64 ± 3.14	57.75	52.39
Difference (95% CI)		**−1.11 (−1.49, −0.73)**	**1.86 (0.99, 2.72)**	0.14 (−0.07, 0.34)	**1.01 (0.31, 1.70)**	0.0 (−0.24, 0.24)	**−9.33 (−13.1, −5.53)**	0.81 (−2.99, 4.61)

Emboldened font indicates statistical significance. Numerical variables: mean ± standard deviation; the difference between subtypes is quantified as the difference in means and its 95% CI. Categorical variables: proportion of APOE ϵ4 carriers and female patients, respectively; differences across subtypes are quantified as the difference in proportions and its 95% CI.

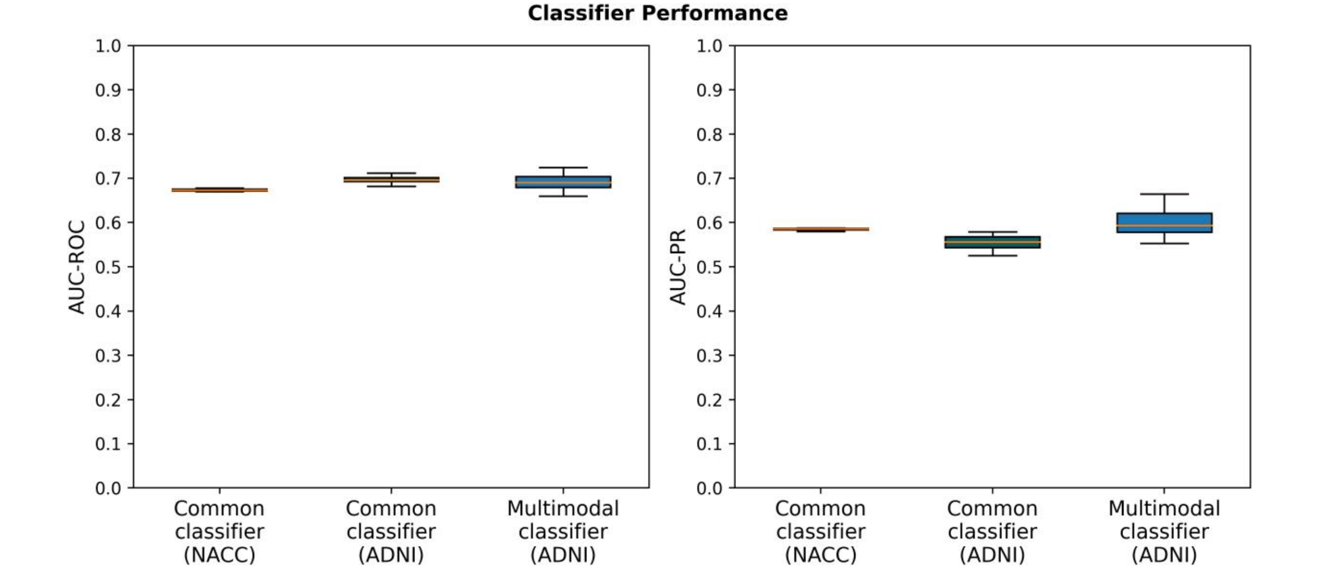


Figure 2 Predicting progression clusters from cross-sectional data. Performance of machine learning classifiers differentiating between 'fast' and 'slow' progressors at their time of dementia diagnosis, averaged across 10 repeats. The data set on which the respective performance was evaluated is shown in parenthesis on the *x*-axis. The application of the common classifier to ADNI represents an external validation. AUC-ROC, area under the receiver operator characteristic curve; AUC-PR, area under the precision-recall curve.

Artificial intelligence-based stratification to enrich cohorts with specific symptom progression subtypes

We emulated an enrichment of a hypothetical clinical trial cohort with patients experiencing 'fast' symptom progression by applying our 'common predictor' classifier to ADNI. The predicted probability for each patient to belong to the 'fast' progressing subgroup was used as an exclusion criterion and patients with predictions below a selected threshold were excluded from the trial.

Without stratification, at their time of dementia diagnosis, 144 of the 230 (62.6%) analysed ADNI patients belonged to the 'fast' progressing subgroup. Expectedly, increasing the classifier threshold required for patient inclusion caused a decrease in the number of patients remaining in the hypothetical trial cohort (Fig. 3B). Simultaneously, however, the proportion of 'fast' progressors among the remaining patients rose consistently (Fig. 3A). After stratifying ADNI using the classifier at a threshold of 0.65, 51.7% of patients [95% CI: (49.9%, 53.6%)] remained in the cohort. The resulting stratified cohort contained 73.4% 'fast' progressors [95% CI: (67.5, 79.2)].

Reducing the required trial cohort sample size through patient enrichment

To estimate a possible reduction in trial cohort sample size achieved through patient enrichment with our 'common predictor' classifier, we performed a statistical power analysis. The parameters for this analysis were taken from recent trials with cognitive endpoints targeting early Alzheimer's disease (Supplementary Table 2), primarily the 'Clarity AD' trial evaluating lecanemab, which identified an effect size of 27%.[4]

As previously discussed, increasing the required prediction threshold for patient inclusion led to a more homogeneous, faster-progressing trial cohort on average (Fig. 3A). This can lead to measuring greater effect sizes, which opens the opportunity to reduce the cohort sample size while maintaining appropriate statistical power (here, 90%). The relationship between the classifier prediction required for patient enrolment and the resulting potential for sample size reduction is presented in Fig. 4. Assuming a threshold of 0.65, for example, the classifier enabled a sample size reduction of 36.8% [CI: (34.0%, 39.5%)].

Estimating the impact of patient enrichment on economical expenses and patient harm

We approximated the economical impact of patient enrichment with our proposed classifier on trials by counterbalancing the possible sample size reduction with the additional expenses of increased patient screening (Table 2). We assumed a hypothetical clinical trial similar to the successful 'Clarity AD' trial for lecanemab[4] with 24-month runtime and CDRSB as primary outcome. The externally validated 'common predictor' classifier was used for patient enrichment. As a classifier prediction, threshold for patient recruitment 0.65 was selected.

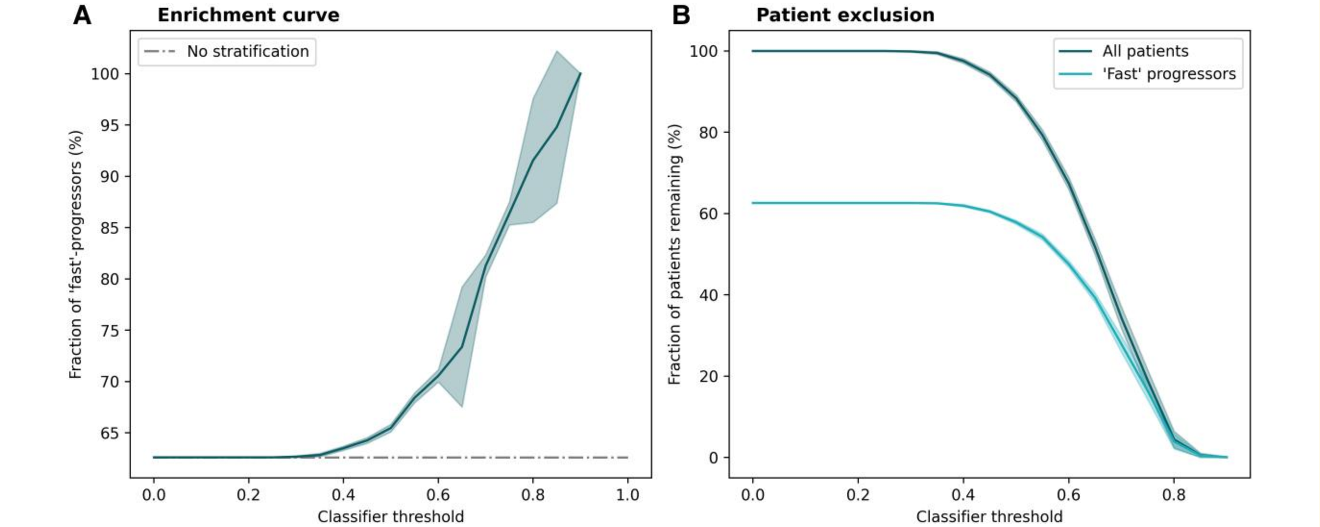


Figure 3 Impact of patient enrichment on a hypothetical trial cohort. ADNI data were used as a hypothetical trial cohort employing our classifier for patient enrichment. Mean trajectories and CIs were calculated across 10 repeats, each with a newly trained model. (**A**) Enrichment of 'fast' progressors with higher classifier thresholds. (**B**) Decrease in sample size with higher classifier thresholds.

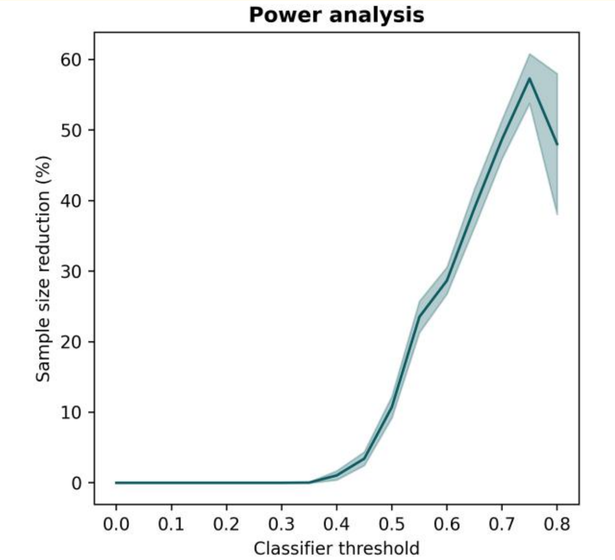


Figure 4 Reducing trial cohort size while maintaining statistical power. Depicted is the possible reduction in trial sample size in relation to the chosen classifier threshold for patient enrichment while maintaining statistical power at 90%. The line displays the mean trajectory calculated across 10 repeats. The shade represents the 95% CI. Larger CIs at higher thresholds are due to lower abundance of individuals with higher scores.

In 'Clarity AD', 1795 patients were enrolled from 5967 screened individuals.[4] Applying our classifier during patient recruitment could reduce the sample size by 36.8% (661 patients) while maintaining 90% statistical power (Fig. 4). This would enable an enrichment trial of the same power by enrolling 1135 participants. Recruiting the enriched trial cohort would require screening 867 additional patients, increasing the trial costs by $6 031 719. The 24-month treatment costs for the lecanemab group (50% of the cohort) would amount to ~$47 594 000 for the conventional trial and $30 077 500 for the enrichment trial.

During the original lecanemab trial, 593 participants experienced adverse events[4] while only 375 patients would be affected in an enrichment trial (218 patients reduction). In the enrichment trial, we would further assume 83 fewer serious adverse events than in the conventional trial (144 versus 227). With respect to ARIA, 102 less cases would occur in an enrichment trial, reducing the monitoring costs for ARIA by $157 080 (from $428 120 to $271 040).

The estimated expenses for a conventional lecanemab trial sum up to ~$89 534 539 while the enrichment trial would cost $77,892,678, thus saving 13% ($11 641 861) of the total costs. Notably, this estimate represents a lower bound neglecting the expenses for treating heterogeneous adverse events and longitudinal monitoring procedures, such as regular neuroimaging.

Discussion

In this work, we utilized deep learning to identify two distinct Alzheimer's disease dementia patient subgroups exhibiting 'slow' and 'fast' symptom progression, respectively. The subgroups were robustly discovered in two independent data sets and externally validated. Using a machine learning classifier, we were able to predict the longitudinal progression subtype of an individual patient with good performance relying only on cross-sectional data collected at the time of their dementia diagnosis. By emulating a clinical trial employing prognostic patient enrichment using this classifier,

Table 2 Comparing the estimated monetary expenses of a conventional trial to an enrichment trial, assuming our 'common predictor' classifier was used for patient stratification employing a classifier threshold of 0.65

	Conventional trial	Enrichment trial	Difference
Screened patients	5967	6834	867
Screening costs	$41 512 419	$47 544 138	$6 031 719
Recruited patients	1796	1135	−661
Treatment costs	$47 594 000	$30 077 500	−$17 516 500
Total adverse events	593	375	−218
Serious adverse events	227	144	−83
ARIA cases	278	176	−102
ARIA monitoring costs	$428 120	$271 040	−$157 080
Total costs	$89 534 539	$77 892 678	−$11 641 861

we demonstrated the statistical, economical and patient health-related benefits enrichment trials hold over conventional clinical trials in clinical Alzheimer's disease dementia.

Instead of relying solely on cross-sectional data, as is commonly done for clustering Alzheimer's disease dementia patients,[13-15] we utilized a longitudinal approach that clusters the multivariate progression of patient trajectories.[21] For clustering variables, we deliberately focussed on clinical outcomes of high relevance to clinical trials. Such outcomes could be biased by differences in the pathological disease stage of patients. However, we could not find any significant differences in key biomarkers of Alzheimer's disease pathology between the two subgroups.

Both of our classifiers that predicted the symptom progression subtype of an individual achieved good prediction performance. Using a multimodal classifier over the feature-reduced 'common predictor' classifier yielded no significant benefit in prediction performance. This could be due to the lower number of patients for which the additional biomarker measurements were available and the complexity of this data, which could warrant greater sample sizes to benefit machine learning approaches. In theory, however, additional predictors could improve the sensitivity and specificity of a classifier. Especially for borderline cases, the continuous nature of biomarker values could allow for more nuanced predictions compared with the mainly categorical features used in the 'common predictor' classifier.

Our developed machine learning classifiers were not designed for decision support in a clinical care setting, and we do not believe that their performance is sufficient for this task. Instead, we deliberately aimed at building classifiers for patient screening in clinical enrichment trials. In such a setting, the patient enrichment using a machine learning classifier gains its power through an application across a substantial number of potential trial participants. While any classifier that performs above chance level will lead to an enrichment of a sought-after patient subgroup given that enough patients are screened, better-performing classifiers are more cost-effective as fewer patients need to be screened to achieve the required cohort size and homogeneity. Similar conditions apply to the threshold placed on classifier predictions, which represents an arbitrary decision that weights the expanse of patient screening against the achieved homogeneity of the resulting trial cohort.

Here, we focussed on prognostic enrichment; however, another promising route would be predictive enrichment based on disease subtypes.[28] The discrimination of patients on a mechanistic level could enable novel clinical trial designs such as umbrella trials[29] for Alzheimer's disease, but would not guarantee an increased homogeneity in the outcome of interest.

Our results indicate that the inter-patient heterogeneity in cognitive symptom progression could hamper clinical trials especially when their duration is shorter than 2 years. During this period, the identified 'slow' progression subgroup experienced only minute cognitive decline and even highly efficacious treatments would show small effect sizes. Previously, the solution to this statistical challenge was often considered to involve increasing sample sizes[30,31]; however, we argue that enrichment trials present a promising alternative with additional benefits.

Economically, enrichment trials increase the screening demands to recruit a cohort of sufficient size but decrease the costs during trial runtime. Accordingly, if treatment becomes considerably cheaper, or screening more expensive, the economic benefit of patient enrichment could vanish. In our simulated hypothetical enrichment trial, however, we found that an enrichment trial aiming at a cognitive outcome could be conducted with significantly smaller cohort sizes as compared with currently ongoing and previously successful trials. In the dementia context, this will likely lead to cheaper clinical trials and less participant harm caused.

Our presented cost analysis comparing conventional clinical trials against enrichment trials is limited in many ways and the presented amounts are only approximate. Additional costs, such as follow-up care and treatments for non-ARIA side-effects, have been neglected, and no additional expenses for applying the classifier and eventually prolonged screening phases were considered. Furthermore, the marketed costs of drugs differ from the factual costs covered by trial sponsors and employed statistical analyses could differ, depending on the trial analysis plans.

Supplementary material

Supplementary material is available at *Brain Communications* online.

Acknowledgements

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) and P50 AG047270 (PI Stephen Strittmatter, MD, PhD).

Funding

This project was partially funded by the European Union's Horizon 2020 research and innovation programme under grant agreement no. 826421, 'TheVirtualBrain-Cloud'.

Competing interests

J.D.J. is an employee of Boehringer Ingelheim Pharma GmbH & Co. KG. The company had no influence on the scientific results of this study.

Data availability

The used data sets are publicly available at https://adni.loni.usc.edu and https://naccdata.org. The code used for the analysis was uploaded to https://github.com/Cojabi/Progression_subtypes.

References

1. Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimers Dement*. 2019;15:321-387.
2. Budd Haeberlein S, Aisen PS, Barkhof F, *et al.* Two randomized Phase 3 studies of aducanumab in early Alzheimer's disease. *J Prev Alzheimers Dis*. 2022;9:197-210.
3. Sims JR, Zimmer JA, Evans CD, *et al.* Donanemab in early symptomatic Alzheimer disease: The TRAILBLAZER-ALZ 2 randomized clinical trial. *JAMA*. 2023;330:512-527.
4. van Dyck CH, Swanson CJ, Aisen P, *et al.* Lecanemab in early Alzheimer's disease. *N Engl J Med*. 2023;388:9-21.
5. Kim CK, Lee YR, Ong L, *et al.* Alzheimer's disease: Key insights from two decades of clinical trial failures. *J Alzheimers Dis*. 2022; 87:83-100.
6. Anderson RM, Hadjichrysanthou C, Evans S, Wong MM. Why do so many clinical trials of therapies for Alzheimer's disease fail? *Lancet*. 2017;390:2327-2329.
7. Birkenbihl C, Salimi Y, Fröhlich H, Initiative, for the J. A. D. N. & Initiative, the A. D. N. Unraveling the heterogeneity in Alzheimer's disease progression across multiple cohorts and the implications for data-driven disease modeling. *Alzheimers Dement*. 2022;18: 251-261.
8. Lonergan M, Senn SJ, McNamee C, *et al.* Defining drug response for stratified medicine. *Drug Discov Today*. 2017;22: 173-179.
9. FDA. *Enrichment strategies for clinical trials to support approval of human drugs and biological products*. U.S. Food and Drug Administration; 2019. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/enrichment-strategies-clinical-trials-support-approval-human-drugs-and-biological-products.
10. Seyhan AA, Carini C. Are innovation and new technologies in precision medicine paving a new era in patients centric care? *J Transl Med*. 2019;17:114.

11. Maheux E, Koval I, Ortholand J, *et al.* Forecasting individual progression trajectories in Alzheimer's disease. *Nat Commun.* 2023; 14:761.
12. Fröhlich H, Balling R, Beerenwinkel N, *et al.* From hype to reality: Data science enabling personalized medicine. *BMC Med.* 2018;16: 150.
13. Scheltens NME, Tijms BM, Koene T, *et al.* Cognitive subtypes of probable Alzheimer's disease robustly identified in four cohorts. *Alzheimers Dement.* 2017;13:1226-1236.
14. Whitwell JL, Graff-Radford J, Tosakulwong N, *et al.* [18f]AV-1451 clustering of entorhinal and cortical uptake in Alzheimer's disease. *Ann Neurol.* 2018;83:248-257.
15. Levin F, Ferreira D, Lange C, *et al.* Data-driven FDG-PET subtypes of Alzheimer's disease-related neurodegeneration. *Alzheimers Res Ther.* 2021;13:49.
16. Vogel JW, Young AL, Oxtoby NP, *et al.* Four distinct trajectories of tau deposition identified in Alzheimer's disease. *Nat Med.* 2021;27: 871-881.
17. Mueller SG, Weiner MW, Thal LJ, *et al.* The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clin.* 2005;15: 869-877.
18. 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;32:351.
19. 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;11:367-376.
20. Birkenbihl C, Salimi Y, Domingo-Fernándéz D, Lovestone S, Fröhlich H, Hofmann-Apitius M. Evaluating the Alzheimer's disease data landscape. *Alzheimer's Dementia.* 2020;6(1):e12102.
21. de Jong J, Emon MA, Wu P, *et al.* Deep learning for clustering of multivariate clinical patient trajectories with missing values. *GigaScience.* 2019;8:giz134.
22. Tibshirani R, Walther G. Cluster validation by prediction strength. *J Comput Graph Stat.* 2005;14:511-528.
23. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.* 2016:785-794. doi:10.1145/2939672.2939785.
24. Salimi Y, Domingo-Fernández D, Bobis-Álvarez C, *et al.* ADataViewer: Exploring semantically harmonized Alzheimer's disease cohort datasets. *Alzheimers Res Ther.* 2022;14:69.
25. Lancet T. Lecanemab for Alzheimer's disease: Tempering hype and hope. *The Lancet.* 2022;400:1899.
26. Jönsson L, Wimo A, Handels R, *et al.* The affordability of lecanemab, an amyloid-targeting therapy for Alzheimer's disease: An EADC-EC viewpoint. *Lancet Reg. Health—Eur.* 2023;29:100657.
27. Ross EL, Weinberg MS, Arnold SE. Cost-effectiveness of aducanumab and donanemab for early Alzheimer disease in the US. *JAMA Neurol.* 2022;79:478-487.
28. de Jong J, Cutcutache I, Page M, *et al.* Towards realizing the vision of precision medicine: AI based prediction of clinical drug response. *Brain.* 2021;144:1738-1750.
29. Fountzilas E, Tsimberidou AM, Vo HH, Kurzrock R. Clinical trial design in the era of precision medicine. *Genome Med.* 2022;14:101.
30. Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K. Alzheimer's disease drug development pipeline: 2020. *Alzheimers Dement. Transl. Res. Clin. Interv.* 2020;6:e12050.
31. Brookmeyer R, Abdalla N. Design and sample size considerations for Alzheimer's disease prevention trials using multistate models. *Clin. Trials.* 2019;16:111-119.