

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge: Results after 1 Year Follow-up

Razvan V. Marinescu^{1,2}, Neil P. Oxtoby¹, Alexandra L. Young^{1,39}, Esther E. Bron³, Arthur W. Toga⁴, Michael W. Weiner⁵, Frederik Barkhof^{1,6,7}, Nick C. Fox⁷, Arman Eshaghi^{45,1}, Tina Toni[†], Marc Salaterski[†], Veronika Lunina[†], Manon Ansart⁸, Stanley Durrleman⁸, Pascal Lu⁸, Samuel Iddi^{9,10}, Dan Li⁹, Wesley K. Thompson¹¹, Michael C. Donohue⁹, Aviv Nahon¹², Yarden Levy¹², Dan Halbersberg¹², Mariya Cohen¹², Huiling Liao¹³, Tengfei Li¹³, Kaixian Yu¹³, Hongtu Zhu¹³, José G. Tamez-Peña¹⁴, Aya Ismail¹⁵, Timothy Wood¹⁵, Hector Corrada Bravo¹⁵, Minh Nguyen¹⁶, Nanbo Sun¹⁶, Jiashi Feng¹⁶, B.T. Thomas Yeo¹⁶, Gang Chen¹⁷, Ke Qi¹⁸, Shiyang Chen^{18,19}, Deqiang Qiu^{18,19}, Ionut Buciuman²⁰, Alex Kelner²⁰, Raluca Pop²⁰, Denisa Rimoccea²⁰, Mostafa M. Ghazi^{21,22,23,1}, Mads Nielsen^{21,22,23}, Sebastien Ourselin^{24,1}, Lauge Sørensen^{21,22,23}, Vikram Venkatraghavan³, Keli Liu²⁵, Christina Rabe²⁵, Paul Manser²⁵, Steven M. Hill²⁶, James Howlett²⁶, Zhiyue Huang²⁶, Steven Kiddle²⁶, Sach Mukherjee⁴², Anaïs Rouanet²⁶, Bernd Taschler⁴², Brian D. M. Tom²⁶, Simon R. White²⁶, Noel Faux²⁷, Suman Sedai²⁷, Javier de Velasco Oriol¹⁴, Edgar E. V. Clemente¹⁴, Karol Estrada^{28,44}, Leon Aksman¹, Andre Altmann¹, Cynthia M. Stonnington²⁹, Yalin Wang⁴⁰, Jianfeng Wu⁴⁰, Vivek Devadas⁴¹, Clementine Fourier⁸, Lars Lau Raket^{30,31}, Aristeidis Sotiras³², Guray Erus³², Jimit Doshi³², Christos Davatzikos³², Jacob Vogel³³, Andrew Doyle³³, Angela Tam³³, Alex Diaz-Papkovich³³, Emmanuel Jammeh³⁴, Igor Koval⁸, Paul Moore³⁵, Terry J. Lyons³⁵, John Gallacher⁴³, Jussi Tohka³⁶, Robert Cisek³⁶, Bruno Jedynak³⁷, Kruti Pandya³⁷, Murat Bilgel³⁸, William Engels³⁷, Joseph Cole³⁷, Polina Golland², Stefan Klein³, Daniel C. Alexander¹, the EuroPOND Consortium[†], and for the Alzheimer's Disease Neuroimaging Initiative*

¹Centre for Medical Image Computing, University College London, UK, ²Computer Science and Artificial Intelligence Laboratory, Massachusetts Institute of Technology, USA, ³Biomedical Imaging Group Rotterdam, Department of Radiology and Nuclear Medicine, Erasmus MC, Netherlands, ⁴Laboratory of NeuroImaging, University of Southern California, USA, ⁵Center for Imaging of Neurodegenerative Diseases, University of California San Francisco, USA, ⁶Department of Radiology and Nuclear Medicine, VU Medical Centre, Netherlands, ⁷Dementia Research Centre and the UK Dementia Research Institute, UCL Queen Square Institute of Neurology, UK, ⁸Institut du Cerveau et de la Moelle épinière, Paris, France, ⁹Alzheimer's Therapeutic Research Institute, University of Southern California, USA, ¹⁰Department of Statistics and Actuarial Science, University of Ghana, Ghana, ¹¹Department of Family Medicine and Public Health, University of California San Diego, USA, ¹²Ben Gurion University of the Negev, Beersheba, Israel, ¹³The University of Texas Health Science Center at Houston, Houston, USA, ¹⁴Instituto Tecnológico y de Estudios Superiores de Monterrey, Monterrey, Mexico, ¹⁵University of Maryland, College Park, USA, ¹⁶National University of Singapore, Singapore, Singapore, ¹⁷Medical College of Wisconsin, Milwaukee, USA, ¹⁸Emory University, Atlanta, USA, ¹⁹Georgia Institute of Technology, Atlanta, USA, ²⁰Vasile Lucaciu National College, Baia Mare, Romania, ²¹Biomediq A/S, Denmark, ²²Cerebriu A/S, Denmark, ²³University of Copenhagen, Denmark, ²⁴School of Biomedical Engineering and Imaging Sciences, King's College London, UK, ²⁵Genentech, USA, ²⁶MRC Biostatistics Unit, University of Cambridge, UK, ²⁷IBM Research Australia, Melbourne, Australia, ²⁸Brandeis University, Waltham, USA, ²⁹Mayo Clinic, Scottsdale, AZ, USA, ³⁰H. Lundbeck A/S, Denmark, ³¹Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Lund, Sweden, ³²Center for Biomedical Image Computing and Analytics, University of Pennsylvania, ³³McGill University, Montreal, Canada, ³⁴University of Plymouth, UK, ³⁵Mathematical Institute, University of Oxford, UK, ³⁶A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Finland, ³⁷Portland State University, Portland, USA, ³⁸Laboratory of Behavioral Neuroscience, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA, ³⁹Department of Neuroimaging, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK, ⁴⁰School of Computing, Informatics and Decision Systems Engineering, Arizona State University, Tempe, USA, ⁴¹Banner Alzheimer's Institute, Phoenix, USA, ⁴²German Center for Neurodegenerative Diseases, Bonn, Germany, ⁴³Department of Psychiatry, University of Oxford, UK, ⁴⁴Department of Statistical Genetics, Biomarin, San Rafael,

*. 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

USA, ⁴⁵Queen Square Multiple Sclerosis Centre, UCL Queen Square Institute of Neurology, UK, [†]Authors not affiliated with any research institution, Correspondence email: tadpole@cs.ucl.ac.uk, Website: <https://tadpole.grand-challenge.org>

Abstract

Accurate prediction of progression in subjects at risk of Alzheimer’s disease is crucial for enrolling the right subjects in clinical trials. However, a prospective comparison of state-of-the-art algorithms for predicting disease onset and progression is currently lacking. We present the findings of *The Alzheimer’s Disease Prediction Of Longitudinal Evolution* (TADPOLE) Challenge, which compared the performance of 92 algorithms from 33 international teams at predicting the future trajectory of 219 individuals at risk of Alzheimer’s disease. Challenge participants were required to make a prediction, for each month of a 5-year future time period, of three key outcomes: clinical diagnosis, Alzheimer’s Disease Assessment Scale Cognitive Subdomain (ADAS-Cog13), and total volume of the ventricles. The methods used by challenge participants included multivariate linear regression, machine learning methods such as support vector machines and deep neural networks, as well as disease progression models. No single submission was best at predicting all three outcomes. For clinical diagnosis and ventricle volume prediction, the best algorithms strongly outperform simple baselines in predictive ability. However, for ADAS-Cog13 no single submitted prediction method was significantly better than random guesswork. Two ensemble methods based on taking the mean and median over all predictions, obtained top scores on almost all tasks. Better than average performance at diagnosis prediction was generally associated with the additional inclusion of features from cerebrospinal fluid (CSF) samples and diffusion tensor imaging (DTI). On the other hand, better performance at ventricle volume prediction was associated with inclusion of summary statistics, such as the slope or maxima/minima of patient-specific biomarkers. On a limited, cross-sectional subset of the data emulating clinical trials, performance of the best algorithms at predicting clinical diagnosis decreased only slightly (2 percentage points) compared to the full longitudinal dataset. The submission system remains open via the website <https://tadpole.grand-challenge.org>, while TADPOLE SHARE (<https://tadpole-share.github.io/>) collates code for submissions. TADPOLE’s unique results suggest that current prediction algorithms provide sufficient accuracy to exploit biomarkers related to clinical diagnosis and ventricle volume, for cohort refinement in clinical trials for Alzheimer’s disease. However, results call into question the usage of cognitive test scores for patient selection and as a primary endpoint in clinical trials.

Keywords: Alzheimer’s disease prediction, Benchmark, Machine Learning, Statistical Modelling

1. Introduction

Accurate prediction of the onset of Alzheimer’s disease (AD) and its longitudinal progression is important for care planning and for patient selection in clinical trials. Current opinion holds that early detection will be critical for the successful administration of disease modifying treatments during presymptomatic phases of the disease prior to widespread brain damage, e.g. when pathological amyloid and tau start to accumulate (Mehta et al. (2017)). Moreover, accurate prediction of the progression of at-risk subjects will help select homogenous patient groups for clinical trials, thus reducing variability in outcome measures that can obscure positive effects on patients at the right stage to benefit.

A variety of mathematical and computational methods have been developed to predict the onset and progression of AD. Traditional approaches leverage statistical regression to model relationships between target variables (e.g. clinical diagnosis or cognitive/imaging markers)

with other known markers (Scahill et al. (2002); Sabuncu et al. (2011)) or measures derived from these markers such as the rate of cognitive decline (Doody et al. (2010)). More recent approaches involve supervised machine learning techniques such as support vector machines (Klöppel et al. (2008); Salas-Gonzalez et al. (2010); Morra et al. (2009); Alvarez et al. (2009)), random forests (Sarica et al. (2017); Ramírez et al. (2018); Lebedev et al. (2014); Huang et al. (2016); Gray et al. (2013)) and artificial neural networks (Jo et al. (2019); Lee et al. (2019); Lin et al. (2018); Spasov et al. (2019); Li et al. (2019); Duc et al. (2020); Cui et al. (2019)). These approaches have been used to discriminate AD patients from cognitively normal individuals (Klöppel et al. (2008); Zhang et al. (2011)), and for discriminating at-risk individuals who convert to AD in a certain time frame from those who do not (Young et al. (2013); Mattila et al. (2011)). The emerging approach of disease progression modelling aims to reconstruct biomarker trajectories or other disease signatures across the disease progression timeline, without relying on clinical diagnoses or estimates of time to symptom onset. Examples include models built on a set of scalar biomarkers to produce discrete (Fonteijn et al. (2012); Young et al. (2014)) or continuous (Jedynak et al. (2012); Donohue et al. (2014); Lorenzi et al. (2017); Oxtoby et al. (2018); Schiratti et al. (2017); Wang et al. (2014); Villemagne et al. (2013); Lorenzi et al. (2019)) biomarker trajectories; spatio-temporal models that focus on evolving image structure (Bilgel et al. (2016); Marinescu et al. (2019); Abi Nader et al. (2020); Bône et al. (2018)), potentially conditioned by non-imaging variables (Koval et al. (2018)); and models that emulate putative disease mechanisms to estimate trajectories of change (Raj et al. (2012); Iturria-Medina et al. (2016); Zhou et al. (2012); Garbarino et al. (2019)). All these models show promise for predicting AD biomarker progression at group and individual levels. However, previous evaluations within individual publications provide limited information because: (1) they use different data sets or subsets of the same dataset, different processing pipelines, and different evaluation metrics and (2) over-training can occur due to heavy use of popular training datasets. Currently, the field lacks a comprehensive comparison of the capabilities of these methods on standardised tasks relevant to real-world applications.

Community challenges have consistently proved effective in moving forward the state of the art in technology to address specific data-analysis problems by providing platforms for unbiased comparative evaluation and incentives to maximise performance on key tasks (Maier-Hein et al. (2018)). In medical image analysis, for example, such challenges have provided important benchmarks in tasks such as registration (Murphy et al. (2011)) and segmentation (Menze et al. (2014)), and revealed fundamental insights about the problem studied, for example in structural brain-connectivity mapping (Maier-Hein et al. (2017)). Previous challenges in AD include the CADDementia challenge (Bron et al. (2015)), which aimed to identify clinical diagnosis from MRI scans. A similar challenge, the *International challenge for automated prediction of MCI from MRI data* (Castiglioni et al. (2018)), asked participants to predict diagnosis and conversion status from extracted MRI features of subjects from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) study (Weiner et al. (2017)). Yet another challenge, *The Alzheimer’s Disease Big Data DREAM Challenge* (Allen et al. (2016)), asked participants to predict cognitive decline from genetic and MRI data. These challenges have however several limitations: (i) they did not evaluate the ability of algorithms to predict biomarkers at future timepoints (with the exception of one sub-task of DREAM), which is important for patient stratification in clinical trials; (ii) the test data was available to organisers when the competitions were launched, leaving room for potential biases in the design of the challenges; (iii) the training data was drawn from a limited set of modalities.

The Alzheimer’s Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge (<https://tadpole.grand-challenge.org>) aims to identify the data, features and approaches that are the most predictive of future progression of subjects at risk of AD. In contrast to previous challenges, our challenge is designed to inform clinical trials through identification of patients most likely to benefit from an effective treatment, i.e., those at early stages of disease

who are likely to progress over the short-to-medium term (1-5 years). The challenge focuses on forecasting the trajectories of three key features: clinical status, cognitive decline, and neurodegeneration (brain atrophy), over a five-year timescale. It uses “rollover” subjects from the ADNI study (Weiner et al. (2017)) for whom a history of measurements (imaging, psychology, demographics, genetics) is available, and who are expected to continue in the study, providing future measurements for testing. TADPOLE participants were required to predict future measurements from these individuals and submit their predictions before a given submission deadline. Since the test data *did not exist* at the time of forecast submissions, the challenge provides a performance comparison substantially less susceptible to many forms of potential bias than previous studies and challenges. The design choices were published (Marinescu et al. (2018)) before the test set was acquired and analysed. TADPOLE also goes beyond previous challenges by drawing on a vast set of multimodal measurements from ADNI which might support prediction of AD progression.

This article presents the results of the TADPOLE Challenge and documents its key findings. We summarise the challenge design and present the results of the 92 prediction algorithms contributed by 33 participating teams worldwide, evaluated after an 18-month follow-up period. We discuss the results obtained by TADPOLE participants, which represent the current state-of-the-art in Alzheimer’s disease prediction. We also report results on which input data features were most informative, and which feature selection strategies, data imputation methods and classes of algorithms were most effective.

2. Methods

2.1 Predictions

TADPOLE Challenge asked participants to forecast three key biomarkers: (1) clinical diagnosis, which can be either cognitively normal (CN), mild cognitive impairment (MCI), or probable AD; (2) Alzheimer’s Disease Assessment Scale Cognitive Subdomain (ADAS-Cog13) score; and (3) ventricle volume (divided by intra-cranial volume) from MRI. Ventricle volume increase has been shown to be a good predictor of Alzheimer’s disease diagnosis and progression Nestor et al. (2008), notably because portions of the lateral ventricles lie close to the medial temporal lobe, which atrophy during early stages (Ferrarini et al. (2006); Giesel et al. (2006)), and because it is correlated with increases in senile plaques and neurofibrillary tangles (Silbert et al. (2003)).

The exact time of future data acquisitions for any given individual was unknown at forecast time, so participants submitted month-by-month predictions for every individual. Predictions of clinical status comprise relative likelihoods of each option (CN, MCI, and AD) for each individual at each month. Predictions of ADAS-Cog13 and ventricle volume comprise a best-guess estimate as well as a 50% confidence interval for each individual at each month. Full details on challenge design are given in the TADPOLE white paper (Marinescu et al. (2018)).

2.2 Data

The challenge uses data from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (Weiner et al. (2017)). Specifically, the TADPOLE Challenge made four key data sets available to the challenge participants:

- **D1:** The TADPOLE standard training set draws on longitudinal data from the entire ADNI history. The data set contains measurements for every individual that has provided data to ADNI in at least two separate visits (different dates) across three phases of the study: ADNI1, ADNI GO, and ADNI2.
- **D2:** The TADPOLE longitudinal prediction set contains as much available data as we could gather from the ADNI rollover individuals for whom challenge participants are

asked to provide predictions. D2 includes data from all available time-points for these individuals. It defines the set of individuals for which participants are required to provide forecasts.

- **D3:** The TADPOLE cross-sectional prediction set contains a single (most recent) time point and a limited set of variables from each rollover individual in D2. Although we expect worse predictions from this data set than D2, D3 represents the information typically available when selecting a cohort for a clinical trial.
- **D4:** The TADPOLE test set contains visits from ADNI rollover subjects that occurred after 1 Jan 2018 and contain at least one of the three outcome measures: diagnostic status, ADAS-Cog13 score, or ventricle volume.

While participants were free to use any training datasets they wished, we provided the D1-D3 datasets in order to remove the need for participants to pre-process the data themselves, and also to be able to evaluate the performance of different algorithms on the same standardised datasets. Participants that used custom training data sets were asked also to submit results using the standard training data sets to enable direct performance comparison. We also included the D3 cross-sectional prediction set in order to simulate a clinical trial scenario. For information on how we created the D1-D4 datasets, see section A. The software code used to generate the standard datasets is openly available on Github: <https://github.com/noxtoby/TADPOLE>.

Table 1 shows the demographic breakdown of each TADPOLE data set as well as the proportion of biomarker data available in each dataset. Many entries are missing data, especially for certain biomarkers derived from exams performed on only subsets of subjects, such as tau imaging (AV1451). D1 and D2 also included demographic data typically available in ADNI (e.g. education, marital status) as well as standard genetic markers (e.g. Alipoprotein E – APOE epsilon 4 status).

2.3 Forecast Evaluation

For evaluation of clinical status predictions, we used similar metrics to those that proved effective in the CADDementia challenge (Bron et al. (2015)): (i) the multiclass area under the receiver operating characteristic curve (MAUC) and (ii) the overall balanced classification accuracy (BCA). For ADAS-Cog13 and ventricle volume, we used three metrics: (i) mean absolute error (MAE), weighted error score (WES) and coverage probability accuracy (CPA). BCA and MAE focus purely on prediction accuracy ignoring confidence, MAUC and WES account for accuracy and confidence, while CPA assesses the confidence interval only. The formulas for each performance metric are summarised in Table 2. See the TADPOLE white paper (Marinescu et al. (2018)) for further rationale for choosing these performance metrics. In order to characterise the distribution of these metric scores, we compute scores based on 50 bootstraps with replacement on the test dataset.

2.4 Statistical Analysis of Method Attributes with Performance

To identify which features and types of algorithms enable good predictions, we annotated each TADPOLE submission with a set of 21 attributes related to (i) feature selection (manual/automatic and large vs. small number of features), (ii) feature types (e.g. “uses Amyloid PET”), (iii) strategy for data imputation (e.g. “patient-wise forward-fill”) and (iv) prediction method (e.g. “neural network”) for clinical diagnosis and ADAS/Ventricles separately. To understand which of these annotations were associated with increased performance, we applied a general linear model (Kiebel and Holmes (2007)), $Y = X\beta + \epsilon$, where Y is the performance metric (e.g. diagnosis MAUC), X is the nr.submissions x 21 design matrix of binary annotations, and β show the contributions of each of the 21 attributes towards achieving the performance measure Y .

Demographics					
		D1	D2	D3	D4
Overall number of subjects		1667	896	896	219
Controls [†]	Number (% all subjects)	508 (30.5%)	369 (41.2%)	299 (33.4%)	94 (42.9%)
	Visits per subject	8.3 ± 4.5	8.5 ± 4.9	1.0 ± 0.0	1.0 ± 0.2
	Age	74.3 ± 5.8	73.6 ± 5.7	72.3 ± 6.2	78.4 ± 7.0
	Gender (% male)	48.6%	47.2%	43.5%	47.9%
	MMSE	29.1 ± 1.1	29.0 ± 1.2	28.9 ± 1.4	29.1 ± 1.1
	Converters*	18	9	-	-
MCI [†]	Number (% all subjects)	841 (50.4%)	458 (51.1%)	269 (30.0%)	90 (41.1%)
	Visits per subject	8.2 ± 3.7	9.1 ± 3.6	1.0 ± 0.0	1.1 ± 0.3
	Age	73.0 ± 7.5	71.6 ± 7.2	71.9 ± 7.1	79.4 ± 7.0
	Gender (% male)	59.3%	56.3%	58.0%	64.4%
	MMSE	27.6 ± 1.8	28.0 ± 1.7	27.6 ± 2.2	28.1 ± 2.1
	Converters*	117	37	-	9
AD [†]	Number (% all subjects)	318 (19.1%)	69 (7.7%)	136 (15.2%)	29 (13.2%)
	Visits per subject	4.9 ± 1.6	5.2 ± 2.6	1.0 ± 0.0	1.1 ± 0.3
	Age	74.8 ± 7.7	75.1 ± 8.4	72.8 ± 7.1	82.2 ± 7.6
	Gender (% male)	55.3%	68.1%	55.9%	51.7%
	MMSE	23.3 ± 2.0	23.1 ± 2.0	20.5 ± 5.9	19.4 ± 7.2
	Converters*	-	-	-	9
Number of clinical visits for all subjects with data available (% of total visits)					
		D1	D2	D3	D4
Cognitive		8862 (69.9%)	5218 (68.1%)	753 (84.0%)	223 (95.3%)
MRI		7884 (62.2%)	4497 (58.7%)	224 (25.0%)	150 (64.1%)
FDG-PET		2119 (16.7%)	1544 (20.2%)	-	-
AV45		2098 (16.6%)	1758 (23.0%)	-	-
AV1451		89 (0.7%)	89 (1.2%)	-	-
DTI		779 (6.1%)	636 (8.3%)	-	-
CSF		2347 (18.5%)	1458 (19.0%)	-	-

Table 1: Summary of TADPOLE datasets D1-D4. ([†]) Diagnosis at first visit with available data. For D3 and D4, 192 and 6 subjects respectively did not have a diagnosis at any clinical visit, so numbers don't add up to 100%. (*) For D4, converters are ADNI3 subjects who are MCI, but were previously CN, or who are AD, but were previously CN or MCI in their last visit in ADNI2. For D1, D2 and D3, converters are CN or MCI at their earliest available visit, who progress to a later classification of MCI/AD within 1.4 years (same duration as D4)

Formula	Definitions
$MAUC = \frac{1}{L(L-1)} \sum_{i=2}^L \sum_{j=1}^i \hat{A}(c_i c_j) + \hat{A}(c_j c_i)$ where $\hat{A}(c_i c_j) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j}$	n_i, n_j – number of points from class i and j . S_{ij} – the sum of the ranks of the class i test points, after ranking all the class i and j data points in increasing likelihood of belonging to class i . L – number of classes. c_i – class i .
$BCA = \frac{1}{2L} \sum_{i=1}^L \left[\frac{TP_i}{TP_i + FN_i} + \frac{TN_i}{TN_i + FP_i} \right]$	TP_i, FP_i, TN_i, FN_i – the number of true positives, false positives, true negatives and false negatives for class i . L – number of classes
$MAE = \frac{1}{N} \sum_{i=1}^N \left \tilde{M}_i - M_i \right $	M_i is the actual value in individual i in future data. \tilde{M}_i is the participant's best guess at M_i and N is the number of data points
$WES = \frac{\sum_{i=1}^N \tilde{C}_i \tilde{M}_i - M_i }{\sum_{i=1}^N \tilde{C}_i}$	M_i, \tilde{M}_i and N defined as above. $\tilde{C}_i = (C_+ - C_-)^{-1}$, where $[C_-, C_+]$ is the 50% confidence interval
$CPA = ACP - 0.5 $	actual coverage probability (ACP) - the proportion of measurements that fall within the 50% confidence interval.

Table 2: TADPOLE performance metric formulas and definitions for the terms.

2.5 Algorithms

A total of 33 participating teams submitted a total of 58 predictions from the longitudinal prediction set (D2), 34 predictions from the cross-sectional prediction set (D3), and 6 predictions from custom prediction sets (see section 2.2 for description of D2/D3 datasets). A total of 8 D2/D3 submissions from 6 teams did not have predictions for all three target variables, so we computed the performance metrics for only the submitted target variables. Another 3 submissions lacked confidence intervals for either ADAS-Cog13 or ventricle volume, which we imputed using default low-width confidence ranges of 2 for ADAS-Cog13 and 0.002 for Ventricle normalised by intracranial volume (ICV).

Submission	Feature selection	Number of features	Missing data imputation	Diagnosis prediction model	ADAS/Vent. prediction model	Training time	Prediction time (one subject)
AlgosForGood 	manual	16+5*	forward-filling	Aalen model	linear regression	1 min.	1 sec.
Apocalypse 	manual	16	population average	SVM	linear regression	40 min.	3 min.
ARAMIS-Pascal 	manual	20	population average	Aalen model	-	16 sec.	0.02 sec.
ATRI-Biostat-JMM 	automatic	15	random forest	random forest	linear mixed effects model	2 days	1 sec.
ATRI-Biostat-LTJMM 	automatic	15	random forest	random forest	DPM	2 days	1 sec.
ATRI-Biostat-MA 	automatic	15	random forest	random forest	DPM + linear mixed effects model	2 days	1 sec.
BGU-LSTM 	automatic	67	none	feed-forward NN	LSTM	1 day	millisec.
BGU-RF/ BGU-RFFIX 	automatic	\approx 67+1340*	none	semi-temporal RF	semi-temporal RF	a few min.	millisec.
BIGS2 	automatic	all	Iterative Thresholded SVD	RF	linear regression	2.2 sec.	0.001 sec.
Billabong (all) 	manual	15-16	linear regression	linear scale	non-parametric SM	7 hours	0.13 sec.
BORREGOSTECMTY 	automatic	\approx 100 + 400*	nearest-neighbour	regression ensemble	ensemble of regression + hazard models	18 hours	0.001 sec.
BravoLab 	automatic	25	hot deck	LSTM	LSTM	1 hour	a few sec.

CBIL 	manual	21	linear interpolation	LSTM	LSTM	1 hour	one min.
Chen-MCW 	manual	9	none	linear regression	DPM	4 hours	< 1 hour
CN2L-NeuralNetwork 	automatic	all	forward-filling	RNN	RNN	24 hours	a few sec.
CN2L-RandomForest 	manual	>200	forward-filling	RF	RF	15 min.	< 1 min.
CN2L-Average 	automatic	all	forward-filling	RNN/RF	RNN/RF	24 hours	< 1 min.
CyberBrains 	manual	5	population average	linear regression	linear regression	20 sec.	20 sec.
DIKU (all) 	semi-automatic	18	none	Bayesian classifier/LDA + DPM	DPM	290 sec.	0.025 sec.
DIVE 	manual	13	none	KDE+DPM	DPM	20 min.	0.06 sec.
EMC1 	automatic	250	nearest neighbour	DPM + 2D spline + SVM	DPM + 2D spline	80 min.	a few sec.
EMC-EB 	automatic	200-338	nearest-neighbour	SVM classifier	SVM regressor	20 sec.	a few sec.
FortuneTellerFish-Control 	manual	19	nearest neighbour	multiclass ECOC SVM	linear mixed effects model	1 min.	< 1 sec.
FortuneTellerFish-SuStaIn 	manual	19	nearest neighbour	multiclass ECOC SVM + DPM	linear mixed effects model + DPM	5 hours	< 1 sec.
Frog 	automatic	$\approx 70+420^*$	none	gradient boosting	gradient boosting	1 hour	-
GlassFrog-LCMEM-HDR 	semi-automatic	all	forward-fill/nearest-neigh.	multi-state model	DPM + regression	15 min.	2 min.
GlassFrog-SM 	manual	7	linear model	multi-state model	parametric SM	93 sec.	0.1 sec.
GlassFrog-Average 	semi-automatic	all	forward-fill/linear	multi-state model	DPM + SM + regression	15 min.	2 min.
IBM-OZ-Res 	manual	Oct-15	filled with zero	stochastic gradient boosting	stochastic gradient boosting	20 min.	0.1 sec.
ITESMCEM 	manual	48	mean of previous values	RF	LASSO + Bayesian ridge regression	20 min.	0.3 sec.
ImaUCL (all) 	manual	5	regression	multi-task learning	multi-task learning	2 hours	millisec.
Mayo-BAI-ASU 	manual	15	population average	linear mixed effects model	linear mixed effects model	20 min.	1.3 sec.
Orange 	manual	17	none	clinician's decision tree	clinician's decision tree	none	0.2 sec.
Rocket 	manual	6	median of diagnostic group	linear mixed effects model	DPM	5 min.	0.3 sec.
SBIA 	manual	30-70	dropped visits with missing data	SVM + density estimator	linear mixed effects model	1 min.	a few sec.
SPMC-Plymouth (all) 	automatic	20	none	unknown	-	unknown	1 min.
SmallHeads-NeuralNetwork 	automatic	376	nearest neighbour	deep fully-connected NN	deep fully-connected NN	40 min.	0.06 sec.
SmallHeads-LinMixedEffects 	automatic	unknown	nearest neighbour	-	linear mixed effects model	25 min.	0.13 sec.
Sunshine (all) 	semi-automatic	6	population average	SVM	linear model	30 min.	< 1 min.
Threedays 	manual	16	none	RF	-	1 min.	3 sec.
Tohka-Ciszek-SMNSR 	manual	≈ 32	nearest neighbour	-	SMNSR	several hours	a few sec.
Tohka-Ciszek-RandomForestLin 	manual	32	mean patient value	RF	linear model	a few min.	a few sec.
VikingAI (all) 	manual	10	none	DPM + ordered logit model	DPM	10 hours	8 sec.
BenchmarkLastVisit 	None	3	none	constant model	constant model	7 sec.	millisec.
BenchmarkMixedEffects 	None	3	none	Gaussian model	linear mixed effects model	30 sec.	0.003 sec.
BenchmarkMixedEffects-APOE 	None	4	none	Gaussian model	linear mixed effects model	30 sec.	0.003 sec.
BenchmarkSVM 	manual	6	mean of previous values	SVM	support vector regressor (SVR)	20 sec.	0.001 sec.

Table 3: Summary of prediction methods used in the TADPOLE submissions. Keywords: SVM – Support Vector Machine, RF – random forest, LSTM – long short-term memory network, NN – neural network, RNN – recurrent neural network, SMNSR – Sparse Multimodal Neighbourhood Search Regression, DPM – disease progression model, KDE – kernel density estimation, LDA – linear discriminant analysis, SM – slope model, ECOC – error-correcting output codes, SVD – singular value decomposition (*) Augmented features, or summary statistics, such as trends, slope, min/max, moments, generally derived patient-wise using longitudinal data. Color tags denote prediction method category: ■ regression/proportional hazards model, ■ random forest, ■ neural networks, ■ disease progression model, ■ machine learning (other), ■ benchmark, ■ other. The left-side box denotes the category for diagnosis prediction method, while the right-side box denotes the category for ADAS/Ventricle prediction method.

Table 3 summarises the methods used in the submissions in terms of feature selection, handling of missing data, predictive models for clinical diagnosis and ADAS/Ventricles biomarkers, as well as training and prediction times. A detailed description of each method is in section 6. In particular, some entries constructed augmented features (i.e. summary statistics), which are extra features such as slope, min/max or moments that are derived from existing features.

In addition to the forecasts submitted by participants, we also evaluated four benchmark methods, which were made available to participants during the submission phase of the challenge: (i) *BenchmarkLastVisit* uses the measurement of each target from the last available clinical visit as the forecast, (ii) *BenchmarkMixedEffects* uses a mixed effects model with age as predictor variable for ADAS and Ventricle predictions, and Gaussian likelihood model for diagnosis prediction, (iii) *BenchmarkMixedEffectsAPOE* is as (ii) but adds APOE status as a covariate and (iv) *BenchmarkSVM* uses an out-of-the-box support vector machine (SVM) classifier and regressor (SVR) to provide forecasts. More details on these methods can be found in section 6. We also evaluated two ensemble methods based on taking the mean (*ConsensusMean*) and median (*ConsensusMedian*) of the forecasted variables over all submissions. We further evaluated 100 random predictions by adding Gaussian noise to the forecasts of the simplest benchmark model (*BenchmarkLastVisit*), to control for potentially spurious strong performance arising from multiple comparisons. In the subsequent results tables we will show, for each performance metric, only the best score obtained by any of these 100 random predictions (*RandomisedBest*) – See end of section 6 for more information on *RandomisedBest*.

3. Results

3.1 Forecasts from the longitudinal prediction set (D2)

Table 4 compiles all metrics for all TADPOLE submitted forecasts, as well as benchmarks and ensemble forecasts, from the longitudinal D2 prediction set. For details on datasets D2 and D3, see section 2.2, while for details on performance metrics see section 2.3. Box-plots showing the distribution of scores, computed on 50 bootstraps of the test set, are shown in Supplementary Fig. 2, while the distribution of ranks is shown in Supplementary Figs. 9 – 11. Among the benchmark methods, *BenchmarkMixedEffectsAPOE* had the best overall rank of 18, obtaining rank 35 on clinical diagnosis prediction, rank 2 on ADAS-Cog13 and rank 23 on Ventricle volume prediction. Removing the APOE status as covariate proved to significantly increase the predictive performance (*BenchmarkMixedEffects*), although we do not show ranks for this entry as it was found during the evaluation phase. Among participant methods, the submission with the best overall rank was *Frog*, obtaining rank 1 for prediction of clinical diagnosis, rank 4 for ADAS-Cog13 and rank 10 for Ventricle volume prediction.

For clinical diagnosis, the best submitted forecasts (team *Frog*) scored better than all benchmark methods, reducing the error of the best benchmark methods by 0.085 (8.5 percentage points) for the multiclass area under the receiver operating characteristic curve (MAUC) and

Submission	Overall Rank	Diagnosis			ADAS-Cog13				Ventricles (% ICV)			
		Rank	MAUC	BCA	Rank	MAE	WES	CPA	Rank	MAE	WES	CPA
ConsensusMedian	-	-	0.925	0.857	-	5.12	5.01	0.28	-	0.38	0.33	0.09
Frog	1	1	0.931	0.849	4	4.85	4.74	0.44	10	0.45	0.33	0.47
ConsensusMean	-	-	0.920	0.835	-	3.75	3.54	0.00	-	0.48	0.45	0.13
EMC1-Std	2	8	0.898	0.811	23-24	6.05	5.40	0.45	1-2	0.41	0.29	0.43
VikingAI-Sigmoid	3	16	0.875	0.760	7	5.20	5.11	0.02	11-12	0.45	0.35	0.20
EMC1-Custom	4	11	0.892	0.798	23-24	6.05	5.40	0.45	1-2	0.41	0.29	0.43
CBLL	5	9	0.897	0.803	15	5.66	5.65	0.37	13	0.46	0.46	0.09
Apocalypse	6	7	0.902	0.827	14	5.57	5.57	0.50	20	0.52	0.52	0.50
GlassFrog-Average	7	4-6	0.902	0.825	8	5.26	5.27	0.26	29	0.68	0.60	0.33
GlassFrog-SM	8	4-6	0.902	0.825	17	5.77	5.92	0.20	21	0.52	0.33	0.20
BORREGOTECMTY	9	19	0.866	0.808	20	5.90	5.82	0.39	5	0.43	0.37	0.40
BenchmarkMixedEffects	-	-	0.846	0.706	-	4.19	4.19	0.31	-	0.56	0.56	0.50
EMC-EB	10	3	0.907	0.805	39	6.75	6.66	0.50	9	0.45	0.40	0.48
lmaUCL-Covariates	11-12	22	0.852	0.760	27	6.28	6.29	0.28	3	0.42	0.41	0.11
CN2L-Average	11-12	27	0.843	0.792	9	5.31	5.31	0.35	16	0.49	0.49	0.33
VikingAI-Logistic	13	20	0.865	0.754	21	6.02	5.91	0.26	11-12	0.45	0.35	0.20
lmaUCL-Std	14	21	0.859	0.781	28	6.30	6.33	0.26	4	0.42	0.41	0.09
RandomisedBest	-	-	0.800	0.803	-	4.52	4.52	0.27	-	0.46	0.45	0.33
CN2L-RandomForest	15-16	10	0.896	0.792	16	5.73	5.73	0.42	31	0.71	0.71	0.41
FortuneTellerFish-SuStaIn	15-16	40	0.806	0.685	3	4.81	4.81	0.21	14	0.49	0.49	0.18
CN2L-NeuralNetwork	17	41	0.783	0.717	10	5.36	5.36	0.34	7	0.44	0.44	0.27
BenchmarkMixedEffectsAPOE	18	35	0.822	0.749	2	4.75	4.75	0.36	23	0.57	0.57	0.40
Tohka-Ciszek-RandomForestLin	19	17	0.875	0.796	22	6.03	6.03	0.15	22	0.56	0.56	0.37
BGU-LSTM	20	12	0.883	0.779	25	6.09	6.12	0.39	25	0.60	0.60	0.23
DIKU-GeneralisedLog-Custom	21	13	0.878	0.790	11-12	5.40	5.40	0.26	38-39	1.05	1.05	0.05
DIKU-GeneralisedLog-Std	22	14	0.877	0.790	11-12	5.40	5.40	0.26	38-39	1.05	1.05	0.05
CyberBrains	23	34	0.823	0.747	6	5.16	5.16	0.24	26	0.62	0.62	0.12
AlgosForGood	24	24	0.847	0.810	13	5.46	5.11	0.13	30	0.69	3.31	0.19
lmaUCL-halfD1	25	26	0.845	0.753	38	6.53	6.51	0.31	6	0.44	0.42	0.13
BGU-RF	26	28	0.838	0.673	29-30	6.33	6.10	0.35	17-18	0.50	0.38	0.26
Mayo-BAI-ASU	27	52	0.691	0.624	5	4.98	4.98	0.32	19	0.52	0.52	0.40
BGU-RFFIX	28	32	0.831	0.673	29-30	6.33	6.10	0.35	17-18	0.50	0.38	0.26
FortuneTellerFish-Control	29	31	0.834	0.692	1	4.70	4.70	0.22	50	1.38	1.38	0.50
GlassFrog-LCMEM-HDR	30	4-6	0.902	0.825	31	6.34	6.21	0.47	51	1.66	1.59	0.41
SBIA	31	43	0.776	0.721	43	7.10	7.38	0.40	8	0.44	0.31	0.13
Chen-MCW-Stratify	32	23	0.848	0.783	36-37	6.48	6.24	0.23	36-37	1.01	1.00	0.11
Rocket	33	54	0.680	0.519	18	5.81	5.71	0.34	28	0.64	0.64	0.29
BenchmarkSVM	34-35	30	0.836	0.764	40	6.82	6.82	0.42	32	0.86	0.84	0.50
Chen-MCW-Std	34-35	29	0.836	0.778	36-37	6.48	6.24	0.23	36-37	1.01	1.00	0.11
DIKU-ModifiedMri-Custom	36	36-37	0.807	0.670	32-35	6.44	6.44	0.27	34-35	0.92	0.92	0.01
DIKU-ModifiedMri-Std	37	38-39	0.806	0.670	32-35	6.44	6.44	0.27	34-35	0.92	0.92	0.01
DIVE	38	51	0.708	0.568	42	7.10	7.10	0.34	15	0.49	0.49	0.13
ITESMCEM	39	53	0.680	0.657	26	6.26	6.26	0.35	33	0.92	0.92	0.43
BenchmarkLastVisit	40	44-45	0.774	0.792	41	7.05	7.05	0.45	27	0.63	0.61	0.47
Sunshine-Conservative	41	25	0.845	0.816	44-45	7.90	7.90	0.50	43-44	1.12	1.12	0.50
BravoLab	42	46	0.771	0.682	47	8.22	8.22	0.49	24	0.58	0.58	0.41
DIKU-ModifiedLog-Custom	43	36-37	0.807	0.670	32-35	6.44	6.44	0.27	47-48	1.17	1.17	0.06
DIKU-ModifiedLog-Std	44	38-39	0.806	0.670	32-35	6.44	6.44	0.27	47-48	1.17	1.17	0.06
Sunshine-Std	45	33	0.825	0.771	44-45	7.90	7.90	0.50	43-44	1.12	1.12	0.50
Billabong-UniAV45	46	49	0.720	0.616	48-49	9.22	8.82	0.29	41-42	1.09	0.99	0.45
Billabong-Uni	47	50	0.718	0.622	48-49	9.22	8.82	0.29	41-42	1.09	0.99	0.45
ATRI-Biostat-JMM	48	42	0.779	0.710	51	12.88	69.62	0.35	54	1.95	5.12	0.33
Billabong-Multi	49	56	0.541	0.556	55	27.01	19.90	0.46	40	1.07	1.07	0.45
ATRI-Biostat-MA	50	47	0.741	0.671	52	12.88	11.32	0.19	53	1.84	5.27	0.23
BIGS2	51	58	0.455	0.488	50	11.62	14.65	0.50	49	1.20	1.12	0.07
Billabong-MultiAV45	52	57	0.527	0.530	56	28.45	21.22	0.47	45	1.13	1.07	0.47
ATRI-Biostat-LTJMM	53	55	0.636	0.563	54	16.07	74.65	0.33	52	1.80	5.01	0.26
Threedays	-	2	0.921	0.823	-	-	-	-	-	-	-	-
ARAMIS-Pascal	-	15	0.876	0.850	-	-	-	-	-	-	-	-
IBM-OZ-Res	-	18	0.868	0.766	-	-	-	-	46	1.15	1.15	0.50
Orange	-	44-45	0.774	0.792	-	-	-	-	-	-	-	-
SMALLHEADS-NeuralNet	-	48	0.737	0.605	53	13.87	13.87	0.41	-	-	-	-
SMALLHEADS-LinMixedEffects	-	-	-	-	46	8.09	7.94	0.04	-	-	-	-
Tohka-Ciszek-SMSNR	-	-	-	-	19	5.87	5.87	0.14	-	-	-	-

Table 4: Ranked scores for all TADPOLE submissions and benchmarks using the longitudinal prediction data set (D2). Best scores in each category are bolded. Missing numerical entries indicate that submissions did not include forecasts for the corresponding target variable. The “Diagnosis” ranking uses multiclass area under the receiver operating characteristic curve (MAUC), those of ADAS-Cog13 and Ventricles use mean absolute error (MAE). The overall ranking on the left uses the sum of the ranks from the three target variables. The table also lists the secondary metrics: BCA – balanced classification accuracy, WES – weighted error score, CPA – coverage probability accuracy.

by 0.058 (5.8 p.p.) for balanced classification accuracy (BCA). Here, the best benchmarks obtained a MAUC of 0.846 (*BenchmarkMixedEffects*) and a BCA of 0.792 (*BenchmarkLastVisit*). Among participant methods, *Frog* had the best MAUC score of 0.931, significantly better than all entries other than *Threedays* according to the bootstrap test (p-value = 0.24, see section B.1 for details on significance testing). Supplementary Figure 9 further shows the variability in performance ranking over bootstrap samples and highlights that the top two entries consistently remain at the top of the ranking. In terms of BCA, *ARAMIS-Pascal* had the best score of 0.850. Moreover, ensemble methods (*ConsensusMedian*) achieved the second best MAUC score of 0.925 and the best BCA score of 0.857. In contrast, the best randomised prediction (*RandomisedBest*) achieved a much lower MAUC of 0.800 and a BCA of 0.803, suggesting entries below these scores did not perform significantly better than random guessing according to the bootstrap test (p-value = 0.01). MAUC and BCA performance metrics had a relatively high correlation across all submissions ($r = 0.88$, Supplementary Fig. 4).

For Ventricle volume, the best submitted forecasts among participants (team *EMC1*) obtained substantially lower error scores than all benchmark methods, scoring 73% of the lowest benchmark MAE (*BenchmarkMixedEffects* MAE=0.56) and 52% of the lowest benchmark WES (*BenchmarkMixedEffects* WES=0.56). Among participant submissions, *EMC1-Std/-Custom* had the best MAE of 0.41 (% ICV), significantly lower than all entries other than *lmaUCL-Covariates/-Std/-half-D1*, *BORREGOTECMTY* and *SBIA* according to the Wilcoxon signed-rank test (see section B.2) – this is also confirmed in Supplementary Fig. 11 by the variability in performance ranking over bootstrap samples. Team *EMC1* also had the best Ventricle WES of 0.29, while *DIKU-ModifiedMri-Custom/-Std* had the best Ventricle coverage probability accuracy (CPA) of 0.01. Ensemble methods (*ConsensusMean*) achieved the best Ventricle MAE of 0.38. In contrast, the best randomised prediction (*RandomisedBest*) achieved a higher MAE of 0.46, WES of 0.45 and CPA of 0.33. MAE and WES scores showed high correlation ($r = 0.99$, Supplementary Fig. 4) and were often of equal value for many submissions ($n = 24$), as teams set equal weights for all subjects analysed. CPA did not correlate ($r \approx -0.01$, Supplementary Fig. 4) with either MAE or WES.

For ADAS-Cog13, the best submitted forecasts did not score significantly better than the simple benchmarks. Here, the simple *BenchmarkMixedEffects* model obtained the second-best MAE of 4.19, which was significantly lower than all other submitted entries according to the Wilcoxon signed-rank test. *BenchmarkMixedEffects* also had the best ADAS-Cog13 WES of 4.19, while *VikingAI-Sigmoid* had the best ADAS-Cog13 CPA of 0.02. Among participants’ submissions, *FortuneTellerFish-Control* ranked first in ADAS-Cog13 prediction with a MAE of 4.70 (112% of the lowest benchmark score). Moreover, all participants’ forecasts scored worse than the best randomised prediction (*RandomisedBest*), which here achieved a MAE of 4.52 and WES of 4.52. Nevertheless, the ensemble method *ConsensusMean* obtained the best ADAS scores for MAE (3.75), WES (3.54) and CPA (0.0), which along with *BenchmarkMixedEffects* were the only entries that performed significantly better than random guesswork (p-value = 0.01). The MAE and WES scores for ADAS-Cog13 had relatively high correlation ($r = 0.97$, Supplementary Fig. 4) and were often of equal value for many submissions ($n = 25$). CPA had a weak but significant correlation with MAE ($r = 0.37$, p-value < 0.02) and WES ($r = 0.35$, p-value < 0.02).

3.2 Forecasts from the cross-sectional prediction set (D3) and custom prediction sets

Table 5 shows the ranking of the forecasts from the cross-sectional D3 prediction set. Box-plots showing the distribution of scores, computed on 50 bootstraps of the test set, are shown in Supplementary Fig 3, while the distribution of ranks is shown in Supplementary Figs. 12 – 14. Due to the lack of longitudinal data, most submissions had lower performance compared to their equivalents from the D2 longitudinal prediction set. Among submitted forecasts, *GlassFrog-*

Submission	Overall Rank	Diagnosis		ADAS-Cog13			Ventricles (% ICV)					
		Rank	MAUC	BCA	Rank	MAE	WES	CPA	Rank	MAE	WES	CPA
ConsensusMean	-	-	0.917	0.821	-	4.58	4.34	0.12	-	0.73	0.72	0.09
ConsensusMedian	-	-	0.905	0.817	-	5.44	5.37	0.19	-	0.71	0.65	0.10
GlassFrog-Average	1	2-4	0.897	0.826	5	5.86	5.57	0.25	3	0.68	0.55	0.24
GlassFrog-LCMEM-HDR	2	2-4	0.897	0.826	9	6.57	6.56	0.34	1	0.48	0.38	0.24
GlassFrog-SM	3	2-4	0.897	0.826	4	5.77	5.77	0.19	9	0.82	0.55	0.07
Tohka-Ciszek-RandomForestLin	4	11	0.865	0.786	2	4.92	4.92	0.10	10	0.83	0.83	0.35
RandomisedBest	-	-	0.811	0.783	-	4.54	4.50	0.26	-	0.92	0.50	0.00
lmaUCL-Std	5-9	12-14	0.854	0.698	16-18	6.95	6.93	0.05	5-7	0.81	0.81	0.22
lmaUCL-Covariates	5-9	12-14	0.854	0.698	16-18	6.95	6.93	0.05	5-7	0.81	0.81	0.22
lmaUCL-halfD1	5-9	12-14	0.854	0.698	16-18	6.95	6.93	0.05	5-7	0.81	0.81	0.22
Rocket	5-9	10	0.865	0.771	3	5.27	5.14	0.39	23	1.06	1.06	0.27
VikingAI-Logistic	5-9	8	0.876	0.768	6	5.94	5.91	0.22	22	1.04	1.01	0.18
EMC1-Std	10	30	0.705	0.567	7	6.29	6.19	0.47	4	0.80	0.62	0.48
BenchmarkMixedEffects	-	-	0.839	0.728	-	4.23	4.23	0.34	-	1.13	1.13	0.50
SBIA	11	28	0.779	0.782	10	6.63	6.43	0.40	8	0.82	0.75	0.18
BGU-LSTM	12-14	5-7	0.877	0.776	13-15	6.75	6.17	0.39	26-28	1.11	0.79	0.17
BGU-RFFIX	12-14	5-7	0.877	0.776	13-15	6.75	6.17	0.39	26-28	1.11	0.79	0.17
BGU-RF	12-14	5-7	0.877	0.776	13-15	6.75	6.17	0.39	26-28	1.11	0.79	0.17
BravoLab	15	18	0.813	0.730	28	8.02	8.02	0.47	2	0.64	0.64	0.42
BORREGOTECMTY	16-17	15	0.852	0.748	8	6.44	5.86	0.46	30	1.14	1.02	0.49
CyberBrains	16-17	17	0.830	0.755	1	4.72	4.72	0.21	35	1.54	1.54	0.50
ATRI-Biostat-MA	18	19	0.799	0.772	26	7.39	6.63	0.04	11	0.93	0.97	0.10
DIKU-GeneralisedLog-Std	19-20	20	0.798	0.684	20-21	6.99	6.99	0.17	16-17	0.95	0.95	0.05
EMC-EB	19-20	9	0.869	0.765	27	7.71	7.91	0.50	21	1.03	1.07	0.49
DIKU-GeneralisedLog-Custom	21	21	0.798	0.681	20-21	6.99	6.99	0.17	16-17	0.95	0.95	0.05
DIKU-ModifiedLog-Std	22-23	22-23	0.798	0.688	22-25	7.10	7.10	0.17	12-15	0.95	0.95	0.05
DIKU-ModifiedMri-Std	22-23	22-23	0.798	0.688	22-25	7.10	7.10	0.17	12-15	0.95	0.95	0.05
DIKU-ModifiedMri-Custom	24-25	24-25	0.798	0.691	22-25	7.10	7.10	0.17	12-15	0.95	0.95	0.05
DIKU-ModifiedLog-Custom	24-25	24-25	0.798	0.691	22-25	7.10	7.10	0.17	12-15	0.95	0.95	0.05
Billabong-Uni	26	31	0.704	0.626	11-12	6.69	6.69	0.38	19-20	0.98	0.98	0.48
Billabong-UniAV45	27	32	0.703	0.620	11-12	6.69	6.69	0.38	19-20	0.98	0.98	0.48
ATRI-Biostat-JMM	28	26	0.794	0.781	29	8.45	8.12	0.34	18	0.97	1.45	0.37
CBIL	29	16	0.847	0.780	33	10.99	11.65	0.49	29	1.12	1.12	0.39
BenchmarkLastVisit	30	27	0.785	0.771	19	6.97	7.07	0.42	33	1.17	0.64	0.11
Billabong-MultiAV45	31	33	0.682	0.603	30-31	9.30	9.30	0.43	24-25	1.09	1.09	0.49
Billabong-Multi	32	34	0.681	0.605	30-31	9.30	9.30	0.43	24-25	1.09	1.09	0.49
ATRI-Biostat-LTJMM	33	29	0.732	0.675	34	12.74	63.98	0.37	32	1.17	1.07	0.40
BenchmarkSVM	34	36	0.494	0.490	32	10.01	10.01	0.42	31	1.15	1.18	0.50
DIVE	35	35	0.512	0.498	35	16.66	16.74	0.41	34	1.42	1.42	0.34
IBM-OZ-Res	-	1	0.905	0.830	-	-	-	-	36	1.77	1.77	0.50

Table 5: Ranked prediction scores for all TADPOLE submissions that used the cross-sectional prediction data set (D3). Best scores in each category are bolded. Missing numerical entries indicate that submissions did not include predictions for the corresponding target variable. The “Diagnosis” ranking uses multiclass area under the receiver operating characteristic curve (MAUC), those of ADAS-Cog13 and Ventricles use mean absolute error (MAE). The overall ranking on the left uses the sum of the ranks from the three target variables. The table also lists the secondary metrics: BCA – balanced classification accuracy, WES – weighted error score, CPA – coverage probability accuracy. See section 2.3 for details on performance metrics.

Average had the best overall rank, as well as rank 2-4 on diagnosis prediction, rank 5 on ADAS-Cog13 prediction and rank 3 on ventricle prediction.

For clinical diagnosis prediction on D3, the best prediction among TADPOLE participants (team *IBM-OZ-Res*) scored better than all benchmarks, improving over the best benchmark MAUC (BenchmarkMixedEffects, MAUC=0.839) by 6.6 percentage points and the best benchmark BCA (BenchmarkLastVisit, BCA=0.771) by 5.9 percentage points. Among participant methods, *IBM-OZ-Res* had the best MAUC score of 0.905, significantly better than all entries other than *GlassFrog-SM/-Average/-LCMEM-HDR*, *BGU-RF/-RFFIX/-LSTM*, *VikingAI-Logistic*, *EMC-EB*, *Rocket* and *Tohka-Ciszek-RandomForestLin* according to the bootstrap hypothesis test (same methodology as in D2). This is further confirmed in Supplementary Fig. 12 by the variability of ranks under bootstrap samples of the dataset, as these teams often remain at the top of the ranking. *IBM-OZ-Res* also had the best BCA score of 0.830 among participants. Among ensemble methods, *ConsensusMean* obtained the best Diagnosis MAUC of 0.917. In contrast, the best randomised prediction (*RandomisedBest*) obtained an MAUC of 0.811 and a BCA of 0.783. MAUC and BCA performance metrics had a relatively high correlation across all submissions ($r = 0.9$, Supplementary Fig. 5).

For Ventricle volume prediction on D3, the best prediction (*GlassFrog-LCMEM-HDR*) obtained substantially lower error scores than all benchmark methods, scoring 42% of the lowest benchmark MAE (BenchmarkMixedEffects MAE=1.13) and 59% of the lowest benchmark WES (BenchmarkLastVisit WES=0.64), and achieving error rates comparable to the best predictions of D2. Among participant submissions, *GlassFrog-LCMEM-HDR* had the best MAE of 0.48, significantly lower than all other submitted entries according to the Wilcoxon signed-rank test – this is also confirmed in Supplementary Fig. 14 by the rank distribution under dataset bootstraps. *GlassFrog-LCMEM-HDR* also had the best Ventricle WES of 0.38, while submissions by team *DIKU* had the best Ventricle CPA of 0.05. Among ensemble methods, *ConsensusMedian* obtained a Ventricle MAE of 0.71 (4th best) and WES of 0.65 (7th best). In contrast, the best randomised prediction (*RandomisedBest*) obtained a Ventricle MAE of 0.92, WES of 0.50 and CPA of 0. As in D2, MAE and WES scores in D3 for Ventricles had very high correlation ($r = 0.99$, Supplementary Fig. 5), while CPA showed weak correlation with MAE ($r = 0.24$, p-value = 0.17) and WES ($r = 0.37$, p-value < 0.032).

For ADAS-Cog13 on D3, the predictions submitted by participants again did not perform better than the best benchmark methods. *BenchmarkMixedEffects* had the best MAE of 4.23, which was significantly lower than all entries by other challenge participants. Moreover, the MAE of 4.23 was only marginally worse than the equivalent MAE (4.19) by the same model on D2. *BenchmarkMixedEffects* also had the best ADAS-Cog13 WES of 4.23, while *ATRI-Biostat-MA* had the best ADAS-Cog13 CPA of 0.04. Among participants’ submissions, *CyberBrains* ranked first in ADAS-Cog13 prediction with MAE/WES scores of 4.72 (111% of the lowest benchmark score). Among ensemble methods, *ConsensusMean* obtained an ADAS-Cog13 MAE of 4.58, WES of 4.34, better than all participants’ entries. As in D2, the best randomised predictions (*RandomisedBest*) obtained an ADAS-Cog13 MAE of 4.54 (2nd best) and WES of 4.50 (3rd best). As in D2, MAE and WES scores for ADAS-Cog13 had high correlation ($r = 0.97$, Supplementary Fig. 5), while CPA showed weak, non-significant correlation with MAE ($r = 0.34$, p-value ≈ 0.052) or WES ($r = 0.33$, p-value ≈ 0.057).

Results on the custom prediction sets are presented in Supplementary Table 6.

3.3 Algorithm characteristics associated with increased performance

To understand what characteristics of algorithms could have yielded higher performance, we show in Figure 1 associations from a general linear model between predictive performance and feature selection methods, different types of features, methods for data imputation, and methods for forecasting of target variables. For each type of feature/method and each target variable (clinical diagnosis, ADAS-Cog13 and Ventricles), we show the distribution of estimated

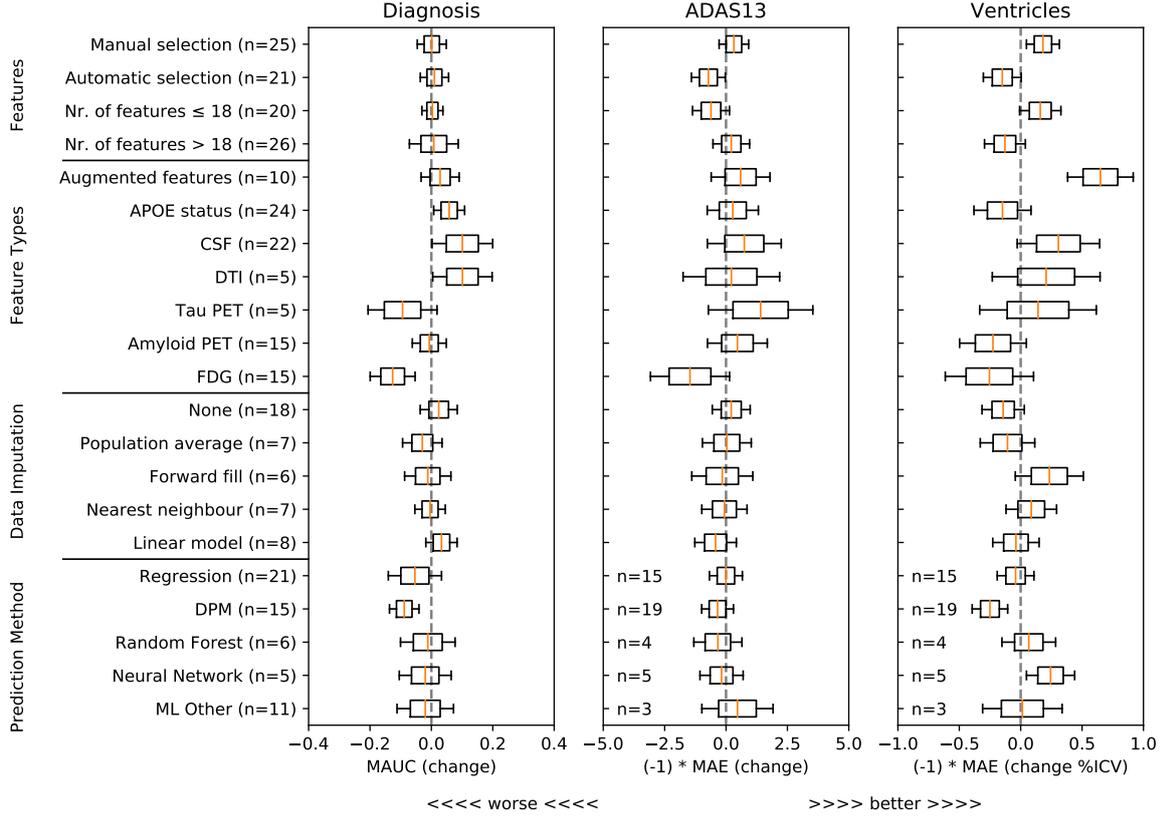


Figure 1: Associations between the prediction of clinical diagnosis, ADAS-Cog13 and Ventricle volume and different strategies of (top) feature selection, (upper-middle) types of features, (lower-middle) data imputation strategies and (bottom) prediction methods for the target variables. For each type of feature/method (rows) and each target variable (columns), we show the distribution of estimated coefficients from a general linear model. Positive coefficients, where distributions lie to the right of the dashed vertical line, indicate better performance than baseline (vertical dashed line). For ADAS-Cog13 and Ventricle prediction, we flipped the sign of the coefficients, to consistently show better performance to the right of the vertical line.

coefficients from a general linear model, derived from the approximated inverse Hessian matrix at the maximum likelihood estimator (see section 2.4). From this analysis we removed outliers, defined as submissions with ADAS MAE higher than 10 and Ventricle MAE higher than 1.15 (%ICV). For all plots, distributions to the right of the gray dashed vertical line denote increased performance compared to baseline (i.e. when those characteristics are not used).

For feature selection, Figure 1 shows that methods with manual selection of features tend to be associated with better predictive performance in ADAS-Cog13 and Ventricles. In terms of feature types, CSF and DTI features were generally associated with an increase in predictive performance for clinical diagnosis, while augmented features were associated with performance improvements for ventricle prediction. In terms of data imputation methods, while some differences can be observed, no clear conclusions can be drawn. In terms of prediction models, the only positive association that indicates increased performance is in the neural networks for ventricle prediction. However, given the small number of methods tested (just under 50) and the large number of degrees of freedom ($= 21$), these results should be interpreted with care.

3.4 External Validation

To verify the performance trends such as ensemble models outperforming individual entries and benchmarks, we performed, in section D, external validation experiments on data from two separate studies: the Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (AIBL) (Ellis et al. (2009)), and the DHA clinical trial (Quinn et al. (2010)).

4. Discussion

The results of the TADPOLE Challenge provide unique and important insights into how well state-of-the-art algorithms can predict progression of AD diagnoses and markers of disease progression both from rich longitudinal data sets and, comparatively, from sparser cross-sectional data sets typical of a clinical trial scenario. The challenge further highlights the algorithms, features and data-handling strategies that tend to lead to improved forecasts. In the following sections we discuss the key conclusions that we draw from our study and highlight important limitations.

4.1 TADPOLE pushed forward the performance on AD clinical diagnosis prediction

In comparison to previous state-of-the-art results in the literature, the best TADPOLE methods show similar or higher performance in AD diagnostic classification while also tackling a harder problem than most previous studies of predicting future, rather than estimating current, classification. A comparison of 15 studies presented by (Moradi et al. (2015)) reported lower performance (maximum AUC of 0.902 vs 0.931 obtained by the best TADPOLE method) for the simpler two-class classification problem of separating MCI-stable from MCI-converters in ADNI. A more recent method by (Long et al. (2017)) reported a maximum AUC of 0.932 and accuracy of 0.88 at the same MCI-stable vs -converter classification task. However, a) TADPOLE’s discrimination of CN-converters from CN-stable subjects is harder as disease signal is weaker at such early stages, and b) the predictive performance drops in three-class problems like TADPOLE compared to two-class. Furthermore, the best out of 19 algorithms in the CADDementia Challenge (Bron et al. (2015)) obtained an MAUC of 0.78.

We are unaware of previous studies forecasting future ventricle volume or ADAS-Cog13, so TADPOLE sets a new benchmark state-of-the-art performance on these important prediction tasks.

4.2 No one-size-fits-all prediction algorithm

The results on the longitudinal D2 prediction set suggest no clear winner on predicting all target variables – no single method performed best on all tasks. While *Frog* had the best overall submission with the lowest sum of ranks, for each performance metric individually different winners emerge: *Frog* (clinical diagnosis MAUC of 0.931), *ARAMIS-Pascal* (clinical diagnosis BCA of 0.850), *BenchmarkMixedEffects* (ADAS-Cog13 MAE and WES of 4.19), *VikingAI-Sigmoid* (ADAS-Cog13 CPA of 0.02), *EMC1-Std/EMC1-Custom* (ventricle MAE of 0.41 and WES of 0.29), and *DIKU-ModifiedMri-Std/-Custom* (ventricle CPA of 0.01). Moreover, on the cross-sectional D3 prediction set, the methods by *Glass-Frog* had the best performance. Associations of method-type with increased performance in Fig. 1 confirm no clear increase in performance for any types of prediction methods (with the exception of neural networks for ventricle volume prediction). This suggests performance depends more on data quality and feature choice. Substantially larger data sets may reveal differences arising from algorithmic choices, but the results we present here are representative of realistic clinical-trial scenarios.

4.3 Characteristics of top-5 algorithms

The top-5 algorithms had several common characteristics. In the clinical diagnosis prediction category (top 5: *Frog*, *Threedays*, *EMC-EB*, *GlassFrog* and *Apocalypse*), 4/5 used Machine Learning methods, 4/5 used APOE status as an input feature and 4/5 predicted the clinical diagnosis directly from the input at each future timestep, whereas 1/5 first predicted the ADAS-Cog/Ventricle measures at each future timestep, then predicted the clinical diagnosis from the ADAS-Cog/Ventricle measures. In the ADAS-Cog13 prediction category (top 5: *BenchmarkMixedEffects*, *FortuneTellerFish*, *Frog*, *Mayo-BAI-ASU* and *CyberBrains*), we find that 4/5 used manual feature selection and 4/5 used linear regression methods. One potential reason why linear regression models performed the best out of all models here is due to their implicit regularization, although we note that none of them performed significantly better than random guessing (with the exception of *BenchmarkMixedEffects* – see section 4.5). In the Ventricle prediction category (top 5: *EMC1*, *ImaUCL*, *BORREGOTECMTY*, *CN2L*, *SBIA*), 4/5 used APOE status as an input feature, 3/5 used an automatic feature selection mechanism (which selected >250 features), 3/5 generated augmented features, and 3/5 used (parametric) regression (while 1/5 used a neural network and 1/5 a disease progression model).

4.4 Ensemble methods perform strongly

Consistently strong results from ensemble methods (*ConsensusMean/ConsensusMedian* outperformed all others on most tasks) might suggest that the varying assumptions in different methods cause different biases: some consistently over-estimate, some consistently under-estimate, and thus averaging aligns more closely with the truth. This is confirmed by plots of the difference between true and estimated measures (Supplementary Figures 6–8), where most methods systematically under- or over-estimate in *all subjects*. The consistent over-estimation and under-estimation of individual methods is likely due to the effect of their inductive biases on the predictions, in the presence of domain shifts from the D1-D2 training sets to the D4 test set (e.g. older individuals in D4 compared to D1-D3). However, even if methods were completely unbiased, averaging over all methods could also help predictions by reducing the variance in the estimated target variables.

4.5 Predictability of ADAS-Cog13 scores

ADAS-Cog13 scores were more difficult to forecast than clinical diagnosis or ventricle volume. The only single method able to forecast ADAS-Cog13 better than informed random guessing (*RandomisedBest*) was the *BenchmarkMixedEffects*, a simple mixed effects model with no covariates and age as a regressor. One possible explanation is the complex multi-effect relationship between the acquired data (imaging, protein markers, etc.) and the composite cognitive test score. However, that relationship is no less complex for clinical diagnosis where prediction appears much more feasible. Alternatively, the difficulty may arise from variability in administering the cognitive tests, or practice effects. Treatment trials often use a change of 4 or more in ADAS-Cog13 as a threshold to identify responders/non-responders ([Grochowalski et al. \(2016\)](#)), so error scores of 4 or less provide a sensible target performance level. With the exception of the ensemble method, all submitted forecasts failed to produce mean error below 4, highlighting the substantial challenge of estimating change in ADAS-Cog13 over the 1.4 year interval. The difficulty in forecasting ADAS-Cog13 calls into question the usage of cognitive test scores in patient selection and as primary endpoint.

4.6 Prediction errors from limited cross-sectional dataset mimicking clinical trials are similar to those from longitudinal dataset

For clinical diagnosis, the best performance on the limited, cross-sectional D3 prediction set was similar to the best performance on the D2 longitudinal prediction set: 0.917 vs 0.931 for MAUC (p-value = 0.14), representing 2 p.p. decrease for D3 compared to D2. Slightly larger and significant differences were observed for ADAS MAE (3.75 vs 4.23, p-value < 0.01) and Ventricle MAE (0.38 vs 0.48, p-value < 0.01). It should be noted that Ventricle predictions for D3 were extremely difficult, given that only 25% of subjects to be forecasted had MRI data in D3. This suggests that, for clinical diagnosis, current forecast algorithms are reasonably robust to lack of longitudinal data and missing inputs, while for ADAS and Ventricle volume prediction, some degree of performance is lost. Future work is also required to determine the optimal balance of input data quality and quantity versus cost of acquisition.

4.7 DTI and CSF features appear informative for clinical diagnosis prediction, augmented features appear informative for ventricle prediction

DTI and CSF features are most associated with increases in clinical diagnosis forecast performance. CSF, in particular, is well established as an early marker of AD ([Jack Jr et al. \(2010\)](#)) and likely to help predictions for early-stage subjects, while DTI, measuring microstructure damage, may be informative for middle-stage subjects. On the other hand, for prediction of ventricle volume, augmented features had the highest association with increases in prediction performance. Future work is required to confirm the added value of these features and others in a more systematic way.

4.8 Challenge design and limitations

TADPOLE Challenge has several limitations that future editions of the challenge may consider addressing. One limitation is the reliability of the three target variables: clinical diagnosis, ADAS-Cog13 and Ventricle volume. First of all, clinical diagnosis has only moderate agreement with gold-standard neuropathological post-mortem diagnosis. In particular, one study ([Beach et al. \(2012\)](#)) has shown that a clinical diagnosis of probable AD has sensitivity between 70.9% and 87.3% and specificity between 44.3% and 70.8%. With the advent of post-mortem confirmation in ADNI, future challenges might address this by evaluating the algorithms on subjects with pathological confirmation. Similarly, ADAS-Cog13 is known to suffer from low reliability across consecutive visits ([Grochowalski et al. \(2016\)](#)), and TADPOLE algorithms fail to forecast it reliably. However, this might be related to the short time-window (1.4 years), and more accurate predictions might be possible over longer time-windows, when there is more significant cognitive decline. Ventricle volume measurements depend on MRI scanner factors such as field strength, manufacturer and pulse sequences ([Han et al. \(2006\)](#)), although these effects have been removed to some extent by ADNI through data preprocessing and protocol harmonization. TADPOLE Challenge also assumes all subjects either remain stable or convert to Alzheimer’s disease, whereas in practice some of them might develop other types of neurodegenerative diseases.

For performance evaluation, we elected to use very simple yet reliable metrics as the primary performance scores: the multiclass area under the curve (mAUC) for the clinical categorical variable and the mean absolute error (MAE) for the two numerical variables. While the mAUC accounts for decision confidence, the MAE does not, which means that the confidence intervals submitted by participants do not contribute to the rankings computed in [Tables 4 and 5](#). While the weighted error score (WES) takes confidence intervals into account, we consider it susceptible to “hacking”, e.g. participants might assign high confidence to only one or two data points and thereby skew the score to ignore most of the predictions – in practice, we did not observe this behaviour in any submission. For clinical relevance, we believe that confidence intervals are an

extremely important part of such predictions and urge future studies to consider performance metrics that require and take account of participant-calculated confidence measures.

TADPOLE has some limitations related to the algorithms’ comparability and generalisability. We could only compare full methods submissions and not different types of features, and strategies for data imputation and prediction used within the full method. While we tried to evaluate the effect of these characteristics in Figure 1, in practice the numbers were small and hence most effects did not reach statistical significance. The analysis in Figure 1 also assumes a linear correspondence between method characteristics and performance, which was necessitated due to the small number of methods tested (just under 50) and the large number of degrees of freedom (= 21). Moreover, the challenge format does not provide an exhaustive comparison of all combinations of data processing, predictive model, features, etc., so does not lead to firm conclusions on the best combinations but rather provides hypotheses for future testing. In future work, we plan to test inclusion of features and strategies for data imputation and prediction independently, by changing one such characteristic at a time. Future challenges might also consider how to provide stronger external validation of findings, e.g. by evaluating all submissions directly on prescribed independent data sets. However, this presents substantial difficulties, as comprehensive external data consistent with the internal, especially for which follow-up occurs on the same timescale, is difficult to find, and extra demands on participants present barriers to entry so the trade-off with engagement must be considered carefully.

Another limitation is that the number of controls and MCI converters in the D4 test set is low (9 MCI converters and 9 control converters). However, these numbers will increase over time as ADNI acquires more data, and we plan to re-run the evaluation at a later stage with the additional data acquired after April 2019. A subsequent evaluation will also enable us to evaluate the TADPOLE methods on longer time-horizons, over which the effects of putative drugs would be higher.

5. Conclusion

In this work we presented the results of the TADPOLE Challenge. The results of the challenge provide important insights into the current state of the art in AD forecasting, such as performance levels achievable with current data and technology as well as specific algorithms, features and data-handling strategies that support the best forecasts. The developments and outcomes of TADPOLE Challenge can aid refinement of cohorts and endpoint assessment for clinical trials, and can support accurate prognostic information in clinical settings. The challenge website (<https://tadpole.grand-challenge.org>) will stay open for submissions, which can be added to our current ranking. The open test set remains available on the ADNI LONI website and also allows individual participants to evaluate future submissions. Through TADPOLE-SHARE <https://tadpole-share.github.io/>, we further plan to implement many TADPOLE methods in a common framework, to be made publicly available. TADPOLE provides a standard benchmark for evaluation of future AD prediction algorithms.

6. Prediction Algorithms

Team: AlgosForGood  (Members: Tina Toni, Marcin Salaterski, Veronika Lunina, Institution: N/A)

Overall Ranking: 24

Feature selection: Manual + Automatic: Manual selection of uncorrelated variables from correlation matrix and automatic selection of variables that have highest cumulative hazard rates in survival regression from MCI to AD.

Selected features: Demographics (age, education, gender, race, marital status), cognitive tests (ADAS-Cog13, RAVLT immediate, RAVLT forgetting, CDRSOB, ADAS11, FDG), Ventricles, AV45, ICV, APOE4.

Missing data: Fill-in using last available value from corresponding patient

Confounder correction: none

Method category: Statistical Regression / Proportional hazards model

Prediction method:

- **Diagnosis:** Aalen additive regression constructing a cumulative hazard rate for progressing to AD.
- **ADAS-Cog13:** regression using change in ventricles/ICV as predictive variable, stratified by last known diagnosis.
- **Ventricles:** regression over month, with several pre-processing steps: 1. Enforced monotonicity by accumulating maximum value, 2. For APOE positive patients used only last three visits due to non-linearity 3. Stratified by diagnosis

Team: Apocalypse  (Members: Manon Ansart, Stanley Durreleman, Institution: Institut du Cerveau et de la Moelle épinière, ICM, Paris, France)

Overall Ranking: 6

Feature selection: Manual – important features were identified by looking at the correlations with the diagnosis. Personal knowledge of the disease was also used to complement those results and select relevant features. Different feature sets were compared using cross-validation.

Selected features: Cognitive features (ADAS-Cog13, MMSE, RAVLT immediate, FAQ, CDRSOB), MRI features (WholeBrain, Entorhinal, Fusiform, MidTemp, Ventricles, Entorhinal, Hippocampus), APOE4, education, age, clinical diagnosis

Missing data: Filled in using the mean feature value

Confounder correction: none

Method category: Machine learning / Regression

Prediction method: Linear regression is used to first predict the future of a set of features (MMSE, ADAS-Cog13, CDRSOB, RAVLT, Ventricles) at the prediction dates. Afterwards, an SVM is used to predict the current diagnosis for each prediction date, based on the forecasted features as well as other features which are constant for the subject (APOE4, education, age at last known visit).

Team: ARAMIS-Pascal  (Members: Pascal Lu, Institution: Institut du Cerveau et de la Moelle épinière, ICM, Paris, France)

Clinical Diagnosis Ranking: 15

Feature selection: Manual, based on known biomarkers from the literature.

Selected features: APOE4, cognitive tests (CDRSB, ADAS11, ADAS-Cog13, MMSE, RAVLT immediate, FAQ), volumetric MRI (hippocampus, ventricles, whole brain, entorhinal, fusiform, middle temporal, ICV), whole brain FDG, CSF biomarkers (amyloid-beta, tau, phosphorylated tau), education and age.

Missing data: Imputed using the average biomarker values across the population.

Confounder correction: none

Method category: Statistical Regression / Proportional hazards model

Prediction method: For diagnosis prediction, the Aalen model for survival analysis was used to predict the conversion from MCI to AD, which returns the probability of a subject remaining MCI as a function of time. The method assumes cognitively normal and dementia subjects will not convert and thus will remain constant. The method did not predict ADAS-Cog13 or Ventricles.

Publication link: <https://hal.inria.fr/tel-02433613/document>

Team: ATRI_Biostat (Members: Samuel Iddi^{1,2}, Dan Li¹, Wesley K. Thompson³ and Michael C. Donohue¹. Institutions: ¹Alzheimer’s Therapeutic Research Institute, USC, USA; ²Department of Statistics and Actuarial Science, University of Ghana, Ghana; ³Department of Family Medicine and Public Health, University of California, USA)

Overall Ranking: 48-53

Feature selection: Automatic - features were ranked by their importance in classifying diagnostic status using a random forest algorithm. All cognitive tests, imaging biomarkers, demographic information and APOE status were considered as potential features

Selected features: ADAS-Cog13, EcogTotal, CDRSOB, FAQ, MOCA, MMSE, RAVLT immediate, Ventricles/ICV, Entorhinal, Hippocampus/ICV and FDG Pet. Age, gender and APOE status were included as covariates. The interaction between diagnosis at first available visit and years since first visit was also considered.

Missing data: Imputed using the MissForest Algorithm, based on a non-parametric random forest methodology (Stekhoven and Bühlmann (2011)). The algorithm was chosen based on its ability to handle mixed-type outcomes, complex interactions and non-linear relationships between variables.

Confounder correction: APOE status, last known clinical status, age and gender.

Method category: Machine learning and data-driven disease progression models

Prediction method: The method applied different types of mixed-effects models to forecast ADAS-Cog13 and Ventricles, and then used a Random Forest classifier to predict the clinical diagnosis from the forecasted continuous scores.

- JMM  – Joint Mixed-effect Modelling with subject-specific intercept and slope.
- LTJMM  – Latent Time Joint Mixed-effect Modelling with subject-specific intercept, slope and time-shift
- MA  – Model average of the two models above, as well as a third model where random intercepts are shared across outcomes.

Confidence Intervals: The 50% prediction intervals for ADAS-Cog13 and Ventricles were obtained by taking the 25th and 75th percentile of the posterior predicted samples.

Publication link: <https://braininformatics.springeropen.com/articles/10.1186/s40708-019-0099-0>

Team: BGU (Members: Aviv Nahon, Yarden Levy, Dan Halbersberg, Mariya Cohen, Institution: Ben Gurion University of the Negev, Beersheba, Israel)

Overall Ranking: 20-28

Feature selection: Automatic – used the following algorithm: 1. Find the two variables with highest correlation (Spearman for continuous variables and Mutual information for discrete variables). 2. Compute the correlation of each variable with the target variables separately and remove the variable with the lower correlation. 3. If there are still pairs of variables with a correlation of more than 80%, repeat from step 1.

Selected features: Cognitive tests (CDRSOB, MMSE, RAVLT, MOCA, all Ecog), MRI biomarkers (Freesurfer cross-sectional and longitudinal), FDG- PET (hypometabolic convergence index), AV45 PET (Ventricles, Corpus Callosum, Hippocampus), White-matter hypointensities’ volume, CSF biomarkers (amyloid-beta, tau, phosphorylated tau). For each continuous variable, an additional set of 20 augmented features was used, representing changes and trends in variables (e.g. mean, standard deviation, trend mean, trend standard deviation, minimum, mean minus global mean, baseline value, last observed value). This resulted in 233 features, which were used for prediction.

Missing data: Random forest can deal automatically with missing data. LSTM network used indicator that was set to zero for missing data.

Confounder correction: None

Method category: Machine learning

Prediction method:

- BGU-LSTM █: This model consisted of two integrated neural networks: an LSTM network for modelling continuous variables and a feed-forward neural network for the static variables.
- BGU-RF █: A semi-temporal Random Forest was used which contained the augmented features.
- BGU-RFFIX █: Same as BGU-RF, but with small correction for the prediction of diagnosis: whenever the model predicted AD with probability higher than 80%, the probability of CN was changed to zero and vice versa.

Team: BIGS2 █ (Members: Huling Liao, Tengfei Li, Kaixian Yu, Hongtu Zhu, Yue Wang, Binxin Zhao, Institution: University of Texas, Houston, USA)

Overall Ranking: 51

Feature selection: Automatic – used auto-encoder to extract aggregated features.

Selected features: All continuous features in D1/D2, which represented the input for the autoencoder. Apart from the autoencoder-extracted features, other features used for the classifier were demographic information, APOE status, whole brain biomarkers from MRI (volume) and PET (FDG, PIB and AV45), and MMSE.

Missing data: SoftImpute method ([Mazumder et al. \(2010\)](#)) was used for imputing missing data. The complete dataset was then used as input to the autoencoder.

Confounder correction: none

Method category: Regression and Machine Learning

Prediction method: Linear models were used to predict ADAS-Cog13 and Ventricle scores independently. For prediction of clinical diagnosis, a random forest was used based on the autoencoder-extracted features and the other selected features.

Team: Billabong █ (Members: Neil Oxtoby, Institution: University College London, UK)

Overall Ranking: 46-52

Feature selection: Manual, using knowledge from literature

Selected features: MRI biomarkers normalised by ICV (ventricles, hippocampus, whole brain, entorhinal, fusiform, middle temporal), FDG, AV45, CSF biomarkers (amyloid beta, tau, phosphorylated tau) and cognitive tests (ADAS-Cog13, MMSE, MOCA, RAVLT immediate). Separate submissions (Billabong-UniAV45, Billabong-MultiAV45) were made which also included AV45, that was initially excluded due to noise.

Missing data: Only imputed during staging via linear regression against age. The method can deal with missing data during training.

Confounder correction: None

Method category: Data-driven disease progression model

Prediction method: For each selected feature independently, a data-driven longitudinal trajectory was estimated using a differential equation model based on Gaussian Process Regression ([Oxtoby et al. \(2018\)](#)). Subjects were staged using either a multivariate or univariate approach:

- Billabong-Uni: Univariate staging which estimates disease stage for each target variable independently.
- Billabong-Multi: Multivariate staging that combines all selected features, producing an average disease stage.

For the prediction of clinical diagnosis, the historical ADNI diagnoses were mapped to a linear scale using partially-overlapping squared-exponential distribution functions. The linear scale and the three distributions were used to forecast the future diagnoses.

Custom prediction set: Predictions were made also for a custom dataset, which was similar to D3 but missing data was filled in using the last available biomarker data.

Confidence Intervals: The 25th and 75th percentiles of the GPR posterior were each integrated into a trajectory to obtain 50% confidence (credible) intervals for the forecasts of ADAS-Cog13 and Ventricles/ICV.

Publication link: <https://doi.org/10.1093/brain/awy050>

Team: BORREGOSTECMTY  (Members: José Gerardo Tamez-Peña, Institution: Tecnológico de Monterrey, Monterrey, Mexico)

Overall Ranking: 9

Feature selection: Automatic, using bootstrapped stage-wise selection.

Selected features: Main cognitive tests (excluding subtypes), MRI biomarkers, APOE status, demographic information (age, gender, education) and diagnosis status. Augmented features were further constructed from the MRI set: the cubic root of all volumes, the square root of all surface areas, the compactness, the coefficient of variation, as well as the mean value and absolute difference between the left and right measurements.

Missing data: Imputed using nearest-neighbourhood strategy based on L1 norm.

Confounder correction: Gender and intracranial volume (ICV) adjustments relative to controls.

Method category: Regression (ensemble of statistical models)

Prediction method: ADAS-Cog13 and Ventricles were predicted using an ensemble of 50 linear regression models, one set for each diagnostic category. The best models were selected using Bootstrap Stage-Wise Model Selection, using statistical fitness (Pencina et al. (2008)) tests to evaluate models and features to use within the models. All selected models were then averaged in a final prediction using bagging. For prediction, the last known diagnosis of the subject was used to select the category of models for forecasting.

For the prediction of clinical diagnosis, a two-stage approach was used based on prognosis and time-to-event estimation. The prognosis approach used an ensemble of 50 regression models to estimate the future diagnosis, while the time-to-event method used an ensemble of 25 models to estimate the square root of the time it took for a subject to convert to MCI or AD. These approaches were performed independently for CN-to-MCI, MCI-to-AD and CN-to-AD conversion.

Confidence Intervals: The 50% confidence intervals for ADAS-Cog13 and Ventricle volume were estimated by extracting the interquartile range of the 50 regression estimates.

Repository link: <https://github.com/joseTamezPena/TADPOLE>

Team: BravoLab  (Members: Aya Ismail, Timothy Wood, Hector Corrada Bravo, Institution: University of Maryland, USA)

Overall Ranking: 42

Feature selection: Automatic, using a random forest to select features with highest cross-entropy or GINI impurity reduction.

Selected features:

- Ventricle prediction: MRI volumes of ventricular sub-regions (Freesurfer cross-sectional and longitudinal)
- ADAS-Cog13 prediction: RAVLT, Diagnosis, MMSE, CDRSOB
- Diagnosis prediction: ADAS-Cog13, ADAS11, MMSE, CSRSOB

Missing data: Imputation using Hot Deck (Andridge and Little (2010)) was done only for data missing at random.

Confounder correction: None

Method category: Machine learning

Prediction method: A long-short term memory network (LSTM) with target replication was trained independently for each category: Diagnosis, ADA13 and Ventricles. All existing data was used for the first forecast, after which the output of the last prediction was used as input for the next prediction, along with other features that remain constant over time. Since subjects had a different number of visits and available biomarker data, the network was adapted to accept inputs of variable length. For predictions, the network used a weighted mean absolute error as a loss function. In addition, for the prediction of the clinical diagnosis, a soft-max function was used to get the final prediction.

Team: CBIL  (Members: Minh Nguyen, Nanbo Sun, Jiashi Feng, Thomas Yeo, Institution: National University of Singapore, Singapore)

Overall Ranking: 5

Feature selection: Manual, based on model performance on D1 subset.

Selected features: Cognitive tests (CDRSOB, ADAS11, ADAS-Cog13, MMSE, RAVLT immediate, learning, forgetting and percent forgetting, MOCA, FAQ), MRI biomarkers (entorhinal, fusiform, hippocampus, ICV, middle temporal, ventricles, whole brain), whole brain AV45 and FDG, CSF biomarkers (amyloid-beta, tau, phosphorylated tau).

Missing data: Imputation using interpolation.

Confounder correction: None

Method category: Machine learning and data-driven disease progression model

Prediction method: Recurrent neural network adapted for variable duration between time-points. A special loss function was designed, which ensured forecasts at timepoints close together are more correlated than those at timepoints further apart.

Confidence Intervals: hardcoded values

Publication link: <https://www.biorxiv.org/content/10.1101/755058v1>

Repository link: https://github.com/ThomasYeoLab/CBIG/tree/master/stable_projects/predict_phenotypes/Nguyen2020_RNNAD

Team: Chen-MCW  (Members: Gang Chen, Institution: Medical College of Wisconsin, Milwaukee, USA)

Overall Ranking: 32-35

Feature selection: Manual

Selected features: ADAS-Cog13, MMSE, MRI volumes (hippocampus, whole brain, entorhinal, fusiform and middle temporal), APOE status, gender and education.

Missing data: No imputation performed.

Confounder correction: None

Method category: Regression and data-driven disease progression model

Prediction method: Prediction of ADAS-Cog13 and Ventricles was made using linear regression using age, APOE status, gender and education as covariates. Different models were estimated for CN, MCI and AD subjects. For diagnosis prediction, an AD risk stage was calculated based on the Event-based probability (EBP) model (Chen et al. (2016)). Prediction of clinical diagnosis was then made based on two approaches:

- Chen-MCW-Std: Predict diagnosis based on AD stage as well as APOE4, gender and education using a Cox proportional hazards model.

- Chen-MCW-Stratify: As above, but the model was further stratified based on AD risk stages, into low risk and high-risk.

Team: CN2L (Members: Ke Qi¹, Shiyang Chen^{1,2}, Deqiang Qiu^{1,2}, Institutions: ¹Emory University, ²Georgia Institute of Technology)

Overall Ranking: 11-17

Feature selection: Automatic

Selected features: For the neural network, all features in D1/D2 are used; For the random forest, the main cognitive tests, MRI biomarkers (cross-sectional only), FDG, AV45, AV1451, DTI, CSF, APOE, demographics and clinical diagnosis are used.

Missing data: Imputation in a forward filled manner (i.e. using last available value).

Confounder correction: None

Method category: Machine learning

Prediction method:

- CN2L-NeuralNetwork ■: 3-layer recurrent neural network, 1024 units/layer. Dropout layers (dropout rate:0.1) were added to output and state connections to prevent overfitting. Adam method was used for training. Validation was performed using a leave-last-time-point-out approach.
- CN2L-RandomForest ■: Random forest method was used, where features of small importance for diagnosis prediction were filtered out. For the prediction of clinical diagnosis, an ensemble of 200 trees was trained on a class-balanced bootstrap sample of the training set. For the prediction of ADAS-Cog13 and Ventricles, an ensemble of 100 trees was used. Different predictions are made for different previous visits of a patient, and the final prediction is taken as the average of all predictions.
- CN2L-Average ■: The average of the above two methods.

Confidence Intervals: Confidence intervals are estimated based on probabilities output of the model.

Publication link: <https://cds.ismrm.org/protected/18MPresentations/abstracts/3668.html> Chen et al., ISMRM, 2018 (S et al. (2018))

Team: CyberBrains ■ (Members: Ionut Buciuman, Alex Kelner, Raluca Pop, Denisa Rimocea, Kruk Zsolt, Institution: Vasile Lucaciu College, Baia Mare, Romania)

Overall Ranking: 23

Feature selection: Manual

Selected features: MRI volumes (Ventricles, middle temporal), ADAS-Cog13, APOE status

Missing data: For subjects with no ventricle measurements, authors computed an average value based on ADAS-Cog13 tests. This was used especially for D3 predictions.

Confounder correction: None

Method category: Regression

Prediction method: Fit a linear model of monthly difference in ventricle volume, as a function of ventricle volume, stratified by clinical diagnosis and ventricle volumes smaller and larger than 140,000 mm³. A similar model is applied for ADAS-Cog13 prediction, but stratified by APOE status and middle temporal volume smaller or greater than 16,000 mm³. Prediction of clinical diagnosis also used a linear model that was stratified based on the ADAS-Cog13, for ADAS-Cog13 ranges between 10 and 45. For ADAS-Cog13 greater than 45 and smaller than 10, pre-defined values were used for the probabilities of each diagnosis.

Team: DIKU ■ (Members: Mostafa Mehdipour Ghazi^{1,2,3,5}, Mads Nielsen^{1,2,3}, Ak-

shay Pai^{1,2,3}, Marc Modat^{4,5}, M. Jorge Cardoso^{4,5}, Sebastien Ourselin^{4,5}, Lauge Sørensen^{1,2,3}; Institutions: ¹Biomediq A/S, ²Cerebriu A/S, ³University of Copenhagen, Denmark, ⁴King's College London, UK, ⁵University College London, UK)

Overall Ranking: 21-44

Feature selection: Semi-automatic; linear discriminant analysis (LDA) was applied to select the top most-informative biomarkers, and ventricular volume and a few other MRI measures were subsequently manually added.

Selected features: Cognitive tests (CDR-SB, ADAS-11, ADAS-13, MMSE, FAQ, MOCA, RAVLT-Immediate, RAVLT-Learning, RAVLT-Percent-Forgetting), CSF measures (amyloid-beta, phosphorylated tau), MRI volumetric measures divided by ICV (ventricles, hippocampus, whole brain, entorhinal, fusiform, middle temporal).

Missing data: Method automatically deals with missing data.

Confounder correction: Linear transformation of age as part of the algorithm.

Method category: Data-driven disease progression model.

Prediction method: For predicting ADAS-13 and Ventricles, a data-driven disease progression model was used, which estimated a parametric trajectory for each selected feature over a common disease progression axis reflecting an estimated latent disease progression score (DPS). The chosen parametric function was generalised logistic function (Richard's curve), and the DPS was a linear transformation of the age of subjects representing the subject-specific time shift and progression speed. The constrained fitting was performed alternating between estimation of subject-specific DPS transformations and global biomarker trajectories using L2-norm loss functions. The authors made three submissions:

- DIKU-GeneralisedLog-Std: constrained, generalised logistic function for the trajectory model;
- DIKU-ModifiedLog-Std: constrained sigmoid function for the trajectory model;
- DIKU-ModifiedMri-Std: as above, but separately fitting MRI biomarkers for Ventricles prediction.

The above models were trained on D1 data only. Authors also made predictions from a custom training set (D1+D2 together), named DIKU-***-Custom. Clinical diagnosis was predicted based on the DPS scores using both a Bayesian classifier with likelihoods modeling using Gaussian mixture models, as well as an ensemble of LDAs. The final prediction was obtained through bagging of the two classifiers' predictions. The whole method and a robust extension developed post-TADPOLE is described in (Ghazi et al. (2019)).

Confidence Intervals: They were obtained by using bootstrapping via Monte Carlo resampling and evaluating the model performance assuming a Gaussian distribution.

Publication link: <https://arxiv.org/abs/1908.05338>

Team: DIVE ████ (Members: Razvan Marinescu, Institution: University College London, UK, Massachusetts Institute of Technology, USA)

Overall Ranking: 38

Feature selection: Manual

Selected features: FDG, AV45, CDRSOB, ADAS-Cog13, MRI volumes (ventricles, hippocampus, whole brain, entorhinal, middle temporal), CSF (amyloid-beta, tau, phosphorylated tau)

Missing data: Method automatically deals with missing data

Confounder correction: None

Method category: Data-driven disease progression model

Prediction method:

For predicting the ADAS-Cog13 and Ventricle volume, the "Data-Driven Inference of Ver-

texwise Evolution” (DIVE) algorithm was used (Marinescu et al. (2019)), which clusters the input biomarkers based on how similar their progression is over the disease time-course. While the original DIVE method was a spatio-temporal model, for TADPOLE it was applied on extracted features directly. The model estimates a parametric, sigmoidal trajectory of the biomarkers, which are a function of subjects’ disease progression scores (DPS), representing a linear transformation of their age. Subject-specific parameters included the latent time-shift and progression speed, as well as an intercept. For the prediction of clinical diagnosis, the posterior probability of each class was computed given the future DPS scores using non-parametric Kernel Density Estimators (KDE), fitted on the DPS scores for each diagnostic class independently. Code for the model is available online:

Confidence Intervals: The model estimates a variance parameter under a gaussian noise model, which was scaled accordingly to obtain the 50% confidence intervals for ADAS-Cog13 and Ventricle volume.

Publication link: <https://www.sciencedirect.com/science/article/pii/S1053811919301491>

Repository link: <https://github.com/mrazvan22/dive>

Team: EMC1  (Members: Vikram Venkatraghavan, Esther Bron, Stefan Klein, Institution: Erasmus MC, The Netherlands)

Overall Ranking: 2-4

Feature selection: Automatic – Only the subjects who had converted to AD were used for feature selection. Features with the largest changes over time after correcting for age, gender, education and ICV were selected

Selected features: 250 features from the set of FDG, AV45, DTI, MRI (cross-sectional Freesurfer volumes), Arterial Spin Labelling (ASL) MRI, CSF and cognitive tests.

Missing data: Imputed using nearest-neighbour interpolation. For D2, visits with missing diagnosis were excluded. For the D3 subjects with no known diagnosis, this was estimated using a nearest-neighbour search based on disease severity

Confounder correction: Corrected for age, gender, education and ICV using linear regression based on data from controls.

Method category: Data-driven disease progression model and machine learning

Prediction method: Authors hypothesize that aging and progression of AD are the primary causes for the change in biomarker values with time and that these changes eventually lead to a change in clinical status. To predict biomarker values at future timepoints, the rate of AD progression is estimated in each subject. This is followed by estimating the interactions of aging and AD progression in the progression of different biomarkers. Lastly, authors use the biomarkers estimated at the future timepoint to predict the change in clinical status. These steps are elaborated below:

Rate of Progression of AD: To assess the severity of AD, we estimated the sequence in which the selected features became abnormal in AD using a Discriminative Event-Based Model (Venkatraghavan et al. (2019)) and used it to estimate the disease severity at all the timepoints for each subject. A linear mixed effect model was fit to estimate the rate of change of disease severity for different subjects. This model was used for predicting the disease severity at all the future timepoints.

Interactions of aging and AD progression: For predicting the biomarker values at the future timepoint, we fit linear mixed effect models for each biomarker considering interactions between the estimated disease severity and age, with gender and ICV as additional covariates. This model was used to forecast the future values of all 250 selected features, including ADAS-Cog13 scores and Ventricle volumes.

Predicting the change in clinical status: For the diagnosis prediction, the forecasted values of

the biomarkers and the last known clinical diagnosis of the subject were used as inputs for a soft-margin SVM classifier with a radial basis function kernel. Two separate submissions were made:

- EMC1-Std (ID 1): ASL based features were excluded in this model
- EMC1-Custom (ID 2): ASL based features were included in this model

Confidence Intervals: Standard errors of the predicted values of Ventricles and ADAS-Cog-13 were estimated by repeating the prediction procedure, including the estimation of disease severity, for 10 repetitions of bootstrap sampling. These standard errors were used to define the confidence intervals.

Publication link: <https://doi.org/10.1016/j.neuroimage.2018.11.024>

Repository link: https://github.com/88vikram/TADPOLE_submission_with_debm

Team: EMC-EB ■ (Members: Esther E. Bron, Vikram Venkatraghavan, Stefan Klein, Institution: Erasmus MC, The Netherlands)

Overall Ranking: 10

Feature selection: Automatic – For the D2 prediction, features were selected that had the largest change over time in subjects who converted to AD using corrections for age, gender, education and ICV, i.e. the same approach as EMC1. For the D3 prediction, features with less than 50% missing data were selected.

Selected features: 200 (D2 prediction) and 338 (D3 prediction) from the set of clinical diagnosis, cognitive tests, MRI volumes (Freesurfer cross-sectional), FDG PET, DTI measures (FA, MD, RD, AD) and CSF features.

Missing data: Imputation using nearest-neighbour interpolation based on the subject's earlier timepoints. If not possible, imputation by the mean of training set was used. Visits with no clinical diagnosis were excluded for classifier training.

Confounder correction: None

Method category: Machine learning

Prediction method: All predictions were based on SVMs. For diagnosis and ADAS-Cog13 prediction, respectively a classifier with balanced class weights and a regressor were trained to predict the target measures at the next visit. These predictions do not explicitly take account of time, but assume that the times between two visits are roughly equal. For Ventricle prediction, a regressor was trained to predict the change of ventricle volume per year. Ventricle volumes were normalized using ICV at baseline (ICV at current time point for D3). Using the predicted change, the normalized ventricle volume at each future visit was computed. For all predictions, authors used a radial basis function (RBF) kernel SVM, of which the C-parameter was set to $C = 0.5$ and gamma to the reciprocal of the number of features. All features were normalized to zero mean and unit standard deviation.

Confidence Intervals: Bootstrap resampling ($n = 100$) of the training set.

Repository link: https://github.com/tadpole-share/tadpole-algorithms/tree/master/tadpole_algorithms/models/ecmeb

Team: FortuneTellerFish (Members: Alexandra Young, Institutions: University College London, UK, King's College London, UK)

Overall Ranking: 15-29

Feature selection: Manual

Selected features: Age at assessment, age at scan, APOE4 status, education, gender, MRI volumes (ventricles, hippocampus, whole brain, entorhinal, fusiform, middle temporal, all major lobes, insula, and basal ganglia). The probability of being amyloid positive, obtained from joint mixture modelling of CSF amyloid-beta and global AV45, was also included as a

feature. Two key features, disease subtype and stage, were derived from the Subtype and Stage Inference (SuStaIn) model based on the MRI features (Young et al. (2018)).

Missing data: Imputed by averaging over the k -nearest neighbours with $k = 5$

Confounder correction: Brain volumes were corrected for age, intracranial volume and field strength using linear regression. Parameters for the linear regression were estimated based on amyloid-negative controls.

Method category: Data-driven disease progression models + statistical regression

Prediction method: For ADAS-Cog13 and Ventricle prediction, a linear mixed effects model was used which used a different set of fixed/random effects, depending on the submission:

- FortuneTellerFish-Control ■■: For Ventricles, fixed effects were age at scan and gender. For ADAS-Cog13 and MMSE the fixed effects were age at scan, education, APOE status and amyloid positivity. For all target measures, there was one random effect per subject.
- FortuneTellerFish-SuStaIn ■■: two additional fixed effects from the SuStaIn model: subtype and stage.

For the prediction of clinical diagnosis, a multiclass error-correcting output codes (ECOC) classifier based on SVMs was trained with the following inputs: age at assessment, age at scan, APOE status, amyloid positivity, gender, education, SuStaIn subtype and stage, ADAS-Cog13, MMSE and ventricle volume. For diagnosis prediction at future timepoints, the forecasted values for ADAS-Cog13, MMSE and Ventricle volume were used as input to the classifier.

Team: Frog ■ (Members: Keli Liu, Christina Rabe, Paul Manser Institution: Genentech, USA)

Overall Ranking: 1

Feature selection: Automatic using the Xgboost package (Chen and Guestrin (2016))

Selected features: Cognitive tests (ADAS-Cog13, CDRSB, MMSE, RAVLT), clinical diagnosis, MRI measurements, FDG PET measurements, APOE status and CSF measurements. For each longitudinal measurement (e.g. test scores and MRI), the following transformations were computed and used to augment the original feature set: most recent measurement, time since most recent measurement, the historical highest (lowest) measurement, time since the historical highest and lowest measurement, and the most recent change in measurement.

Missing data: Xgboost package automatically deals with missing data through inference based on reduction of training loss.

Confounder correction: None

Method category: Statistical prediction using regression

Prediction method: Flexible models and features were chosen automatically using gradient boosting (Xgboost package). Different models were trained for the following forecast windows: 0-8 months, 9-15, 16-27, 28-39, 40-60, >60 (given windows are for clinical status prediction, slightly different windows used for ADAS-Cog13 and ventricular volume prediction). Variable importance scores from Xgboost suggest that MRI features play a bigger role in models trained for longer forecast windows.

Confidence Intervals: Standard deviation for prediction error was estimated based on cross validation (on training set). Normality of prediction errors was then assumed to construct prediction intervals based on estimated standard deviation.

Team: GlassFrog-LCMEM-HDR ■■ (Members: Steven Hill¹, James Howlett¹, Robin Huang¹, Steven Kiddle¹, Sach Mukherjee², Anaïs Rouanet¹, Bernd Taschler², Brian Tom¹, Simon White¹, Institutions: ¹MRC Biostatistics Unit, University of Cambridge, UK; ²German Center for Neurodegenerative Diseases (DZNE), Bonn, Germany)

Overall Ranking: 30

Feature selection:

- MSM: Automatic, by selecting features which passed a likelihood ratio test when compared against a model with no covariates.
- LCMEM: Manual
- HDR: Automatic, selected via sparse Lasso regression.

Selected features:

- MSM: D2 - gender, age, education, ADAS-Cog13, diagnosis, MMSE, CDRSOB, APOE status, first 5 principal components from imaging, amyloid positivity, tau level. D3 - gender, age, education, ADAS-Cog13, diagnosis, MMSE, Ventricles/ICV
- LCMEM: ADAS-Cog13, gender, education, age at baseline
- HDR: all features were provided to the method, excluding some features with many missing values

Missing data:

- MSM: Filling with last known value, or nearest neighbour if feature was never observed.
- LCMEM: Complete case analysis (assumption that missing data are missing at random)
- HDR: Imputation using within-subject interpolation and nearest neighbour matching

Confounder correction: None

Method category: Combination of statistical regression and data-driven disease progression models

Prediction method: The prediction of clinical diagnosis was done using a Multi-State Model (MSM). Multi-state models (MSMs) ([Kalbfleisch and Lawless \(1985\)](#)) are continuous-time Markov chain models, here with states corresponding to CN, MCI, AD, and transition rates estimated from the data using covariates selected as described above. The model accounts for noise in the historical diagnostic labels. Predictions for a given forecast month were made using the last observed disease state and associated covariates.

Prediction of ADAS-Cog13 was done using a Latent class mixed effects model (LCMEM). The model used four latent classes, where class membership probability was modelled via a multinomial logistic. For each latent class a specific linear mixed effects model defined a Gaussian latent process, with class-specific fixed effects, random effects for intercept, slope and square slope, and Gaussian noise. Age at baseline, gender and education were also included as covariates. Finally, a Beta cumulative distribution link function was used (and estimated simultaneously) between ADAS-Cog 13 and the latent process, to account for the departure from the Gaussian assumption on the outcome. The number of latent classes was optimised with the Bayesian Information Criterion (BIC).

Prediction of Ventricle volume was done using high-dimensional regression (HDR) and disease state-specific slope models: Subject-specific slopes were obtained by a combination of Lasso regression and shrinkage towards disease-state-specific shrinkage targets. Conversion times, from one disease state to another, were forecasted using the MSM model.

Confidence Intervals: LCMEM: the 50% confidence intervals for ADAS-Cog13 were obtained using a bootstrap approach. HDR: Confidence intervals were set as percentages of the predicted values.

Team: GlassFrog-SM  (members and affiliations as above)

Overall Ranking: 8

Feature selection: Manual

Selected features: ADAS-Cog13, Ventricles/ICV, age at visit, APOE status, education, diagnosis

Missing data: For training, complete case analysis was performed. For prediction, imputation

was performed for missing outcomes using a linear model with age, education, diagnosis and APOE status as covariates.

Confounder correction: None

Method category: Combination of statistical regression and data-driven disease progression models

Prediction method: The prediction of clinical diagnosis was using MSM models as in GlassFrog-LCMEM-HDR. The prediction for ADAS-Cog13 and Ventricles used a Slope Model (SM), which used a quadratic function to model the slope of the outcome variable as a function of the current outcome value and covariates. Covariates used were age at visit, education and APOE status.

Confidence Intervals: SM: Confidence intervals were set as percentages of the predicted values. The percentages used were manually selected and depended on the missingness of covariates for each individual.

Team: GlassFrog-Average  (members and affiliations as above)

Overall Ranking: 7

Prediction method: The prediction of clinical diagnosis was using MSM models as in GlassFrog-LCMEM-HDR. For ADAS-Cog13 and Ventricles an ensemble approach was used that averaged the predictions from three methods: LCMEM, HDR and SM, as described above. Confidence interval bounds were also averaged.

Team: IBM-OZ-Res (Members: Noel Faux, Suman Sedai, Institution: IBM Research Australia, Melbourne, Australia)

Clinical Diagnosis Ranking: 18

Feature selection: using boosting regression

Selected features: Ventricle volume, AV45, FDG PET, cognitive tests, clinical diagnosis, age

Missing data: Imputed with zero; observations with missing ventricle volume are dropped.

Confounder correction: None

Method category: Machine Learning

Prediction method: A stochastic gradient boosting regression machine (GBM) was used to predict Ventricle Volume, with a Huber loss function. To reduce overfitting, a shrinkage mechanism was adopted, where the response of each tree is reduced by a factor of 0.01. Independent predictions were made for each individual visit, and averaged when a subject had more than one visit. For the prediction of clinical status, a similar GBM model was adopted, but with a multinomial deviance loss function.

Team: ITESMCEM  (Members: Javier de Velasco Oriol¹, Edgar Emmanuel Vallejo Clemente¹, Karol Estrada². Institution: ¹Instituto Tecnológico y de Estudios Superiores de Monterrey, ²Brandeis University)

Overall Ranking: 39

Feature selection: Manual

Selected features: Demographics, MRI volumes, FDG PET and all cognitive tests

Missing data: Imputation using the mean of previous values of that patient, otherwise mean across all patients.

Confounder correction: None

Method category: Machine learning

Prediction method: ADAS-Cog13 is predicted with a Lasso model with $\alpha = 0.1$. Ventricles are predicted using a Bayesian ridge regression. Clinical diagnosis is predicted using two different Random Forest models, one which selected between CN and either MCI or AD and

the second which in turn predicted between MCI and AD for those selected as non-CN. The predictions for all target variables made use of a “transition model”, which predicted the next timepoint given the current one, until all 60 monthly predictions were made. The transition model was implemented using a total of 29 Lasso models.

Confidence Intervals: They were calculated by sampling the test samples multiple times, evaluating the performance of the model with those samples and analyzing the CIs supposing a Gaussian distribution and the corresponding t-distribution.

Team: lmaUCL  (Members: Leon Aksman, Institution: University College London, UK)

Overall Ranking: 11-25

Feature selection: Manual

Selected features: Diagnosis, gender, education, APOE4 status and MMSE.

Missing data: Imputation using regression over ventricles and demographics (age, gender, education)

Confounder correction: None

Method category: Statistical regression and machine learning

Prediction method: For ADAS-Cog13 and Ventricles, a multi-task learning model was used with similar trajectories across subjects. The regression model was a linear model over age, but dependencies between different subjects were modelled through a special prior structure over the coefficients of the linear model. The prior structure has hyperparameters that control for the amount of coupling across subjects, and are optimised through empirical Bayes. Clinical diagnosis was predicted using the ADAS-Cog13 trajectory estimates plus a simple estimate of the mean and standard deviation of ADAS-Cog13 in each diagnostic group (AD/MCI/CN). Each ADAS-Cog13 prediction was then assigned a probability of belonging to each group. Three different submissions were made:

- lmaUCL-Std: used only last available diagnosis as covariate
- lmaUCL-Covariates: as above, but also used gender, education, APOE status and MMSE as covariates
- lmaUCL-halfD1: trained only on half of the D1 dataset, but allowed for longer training time of 10 hours.

Confidence Intervals: Both the multi-task learning and clinical diagnosis models are probabilistic, providing estimates of predictive mean and standard deviation assuming a normal distribution. These were converted to confidence intervals using the inverse normal CDF.

Publication link: <https://doi.org/10.1002/hbm.24682>

Repository link: <https://github.com/LeonAksman/bayes-mtl-traj>

Team: Mayo-BAI-ASU  (Members: Cynthia M. Stonnington¹, Yalin Wang², Jianfeng Wu², Vivek Devadas³, Institution: ¹Mayo Clinic, Scottsdale, AZ, USA, ²School of Computing, Informatics and Decision Systems Engineering, Arizona State University, USA, ³Banner Alzheimer's Institute, Phoenix, AZ, USA)

Overall Ranking: 27

Feature selection: Manual, from clinical experience

Selected features: Age, PET(AV45, AV1451, FDG), hippocampal volume/ICV, ventricle volume/ICV, diagnosis, cognitive tests (ADAS11, ADAS-Cog13, MMSE, RAVLT, MOCA, Ecog), amyloid-beta, tau, phosphorylated tau, APOE status. All features except age were z-score normalised.

Missing data: Imputation with zero.

Confounder correction: None

Method category: Statistical regression

Prediction method: ADAS-Cog13 and Ventricles were forecasted using a linear mixed effects model, using all features as fixed effects and one random effect per subject (intercept). Training used all visits, but the forecasts only used the last visit. Clinical diagnosis was predicted with a similar model, by converting to CN/MCI/AD to a categorical variable (1/2/3).

Team: Orange  (Members: Clementine Fourrier, Institution: Institut du Cerveau et de la Moelle épinière, ICM, Paris, France)

Clinical Diagnosis Ranking: 44-45

Feature selection: Manual, through knowledge from literature

Selected features: demographics (age, education, gender), cognitive tests (ADAS11, CDSRB, MMSE), imaging (AV45 PET, FDG PET, hippocampus size, cortical thickness) and molecular markers (phosphorylated tau to amyloid-beta ratio, total tau, CMRgl, HCI)

Method category: Decision tree of a clinician

Prediction method: This method is based on the decision tree of a clinician. It looks at the latest available visit for a patient, and based on the value of the selected features, it predicts a duration to conversion. The duration to conversion depends on the initial clinical diagnosis and the other available data. Depending on the initial diagnosis, the algorithm assumes the patient will convert within a certain time period, and this time period is modulated by the available data about the patient. In that regard, the algorithm does not need to account for missing values. The ADAS-Cog13 and Ventricle measures at each month are computed assuming a linear evolution between the current time point and the conversion date.

Team: Rocket  (Members: Lars Lau Raket, Institutions: H. Lundbeck A/S, Denmark; Clinical Memory Research Unit, Department of Clinical Sciences Malmö, Lund University, Lund, Sweden)

Overall Ranking: 33

Feature selection: Manual

Selected features: ADAS-Cog13, baseline diagnosis, and APOE4 carrier status.

Missing data: APOE4: Imputed using median number of alleles per baseline diagnostic group. Missing ADAS-Cog13 scores were imputed using multivariate imputation by chained equations based on age, sex, diagnosis and cognitive tests.

Confounder correction: None

Method category: Statistical regression and data-driven disease progression modelling

Prediction method: Prediction of ADAS-Cog13 is done through a latent-time non-linear mixed-effects model, where the trajectory is parameterised using an exponential function. The time shift is built from a fixed effect shift relative to time since baseline for different diagnostic groups (e.g. AD patients will be shifted to be later in the course of cognitive decline) and a random effect shift for each subject. The model also includes APOE status as a fixed effect that modifies rate of decline. This disease progression modeling methodology along with several extensions is presented in (Raket (2019)). Prediction of Ventricles/ICV uses a linear mixed-effects model using an integrated B-spline basis (5 knots + intercept) in predicted ADAS-Cog13 disease time. A random intercept is included per subject. Prediction of clinical diagnosis is based on kernel density estimation of the states (CN/MCI/AD) across the disease time from the ADAS-Cog13 model.

Confidence Intervals: Prediction intervals conditioned on the predicted disease time of the subject were derived based on the estimated variance-covariance matrix of the model. Because of the monotone nature of cognitive decline, and to heuristically compensate for the conditioning on disease time, the upper limit of the prediction interval was multiplied by 1.5.

Publication link: <https://doi.org/10.1101/2019.12.13.19014860>

Team: SBIA  (Members: Aristeidis Sotiras, Guray Erus, Jimit Doshi, Christos Davatzikos, Institution: Center for Biomedical Image Computing and Analytics, University of Pennsylvania)

Overall Ranking: 31

Feature selection: Manual

Selected features: demographics, cognitive tests, diagnosis, MRI features (Freesurfer cross-sectional). Imaging indices (SPARE-AD and SPARE-MCI) that summarise brain atrophy patterns were estimated through support vector machines with linear kernels. Another index representing brain age (SPARE-BA) was estimated using a regressor model applied to imaging features.

Missing data: Features and time-points with missing data were not included

Confounder correction: Age, gender, APOE4, education, and SPARE scores were used as covariates in the linear mixed effects models. Also, the regression model was applied separately on different diagnosis groups.

Method category: Statistical regression and machine learning

Prediction method: A linear mixed effects model was used to forecast the SPARE indices for future timepoints. For diagnosis predictions, authors used class probability distribution estimations based on the forecasted SPARE-AD score. For the prediction of ADAS-Cog13 and Ventricles, linear mixed effects models were used, with age, gender and SPARE scores as covariates.

Team: SmallHeads – BigBrains (Members: Jacob Vogel, Andrew Doyle, Angela Tam, Alex Diaz-Papkovich, Institution: McGill University, Montreal, Canada)

ADAS-Cog13 Ranking: 46-53

Feature selection: Automatic

Selected features:

- SmallHeads-NeuralNetwork : All features in the TADPOLE spreadsheet were considered, as long as they had less than 50% missing data. This resulted in a final set of 376 features. Features were normalised to zero mean and unit variance.
- SmallHeads-LinMixedEffects : All features were normalised to zero mean and unit variance. A LASSO feature selection algorithm with ADAS-Cog13 and “Y” variable was used to select the best features, which were required to have a weight greater than 0.001. 10-fold cross-validation was used to estimate the best LASSO parameters.

Missing data: Imputed using 5-nearest neighbour method, using Euclidean distance (Fancy-Impute 0.0.4)

Confounder correction: None

Method category: Machine learning

Prediction method:

- SmallHeads-NeuralNetwork: A Deep fully connected neural network was trained to predict the future diagnosis using the selected features and time until future timepoint as input. Network has 5 fully-connected layers with Leaky ReLU activations. Each layer has 512, 512, 1024, 1024 and 256 neurons, with softmax layer at output. Training used P(0.5) dropout using the Adam optimiser, based on a class-unweighted categorical cross-entropy loss function.
- SmallHeads-LinMixedEffects: Only ADAS-Cog13 was predicted with a linear mixed effects model, using months since baseline as an interaction term. A random (subject — time) effect was also added, allowing variable subject-specific slopes over time.

Repository link: <https://github.com/SmallHeads/tadpole>

Team: SPMC-Plymouth  (Members: Emmanuel Jammeh, Institution: University of Plymouth, UK)

Overall Ranking: N/A

Feature selection: Automatic – Authors used the WEKA (<https://www.cs.waikato.ac.nz/ml/weka/>) machine learning tool.

Selected features: Age, gender, education, ApoE4, CDRSB, ADAS11, MMSE, RAVLT, Moca, Ecog, Hippocampus, WholeBrain, Entorhinal, MidTemp, FDG, AV45, PIB, ABETA, TAU, PTAU

Missing data:

Confounder correction: None

Method category: Machine learning

Prediction method: A machine learning classifier based on k-nearest neighbours, Naive Bayes, Random Forest and SVM was used to predict the clinical diagnosis. ADAS-Cog13 and Ventricle volume were not predicted. Two predictions were made, SPMC-Plymouth1 and SPMC-Plymouth2, but the authors could not be contacted to provide details on the differences between the two.

Team: Sunshine  (Members: Igor Koval, Stanley Durrleman, Institution: Institut du Cerveau et de la Moelle épinière, ICM, Paris, France)

Overall Ranking: 41-45

Feature selection: Semi-automatic – An initial set of 60 features was selected by a clinical expert. Out of this set, the features that had more than 30% missing data were removed. A final subset of features was chosen based on cross-validation results using trial and error.

Selected features: Age, APOE status, MMSE, ADAS-Cog13, RAVLT immediate and CDRSOB

Missing data: Imputed using mean value

Confounder correction: None

Method category: Statistical regression and machine learning

Prediction method: A linear model was used to predict future values of ADAS-Cog13 and Ventricles, as well as other cognitive tests: MMSE, RAVLT and CDRSOB. For the prediction of clinical diagnosis, forecasted values of the previous five measures were used as input to an SVM. APOE4 status and education were also used as inputs to the SVM. Adding extra features did not seem to increase prediction scores based on cross-validation. Two submissions were made:

- Sunshine-Conservative: CN and AD subjects were forecasted to have the same diagnosis (i.e. no conversion) for all future timepoints, after observing that a small proportion of them convert after 1 year.
- Sunshine-Std: Without the above modification.

Team: Threedays  (Members: Paul Moore¹, Terry J. Lyons¹, John Gallacher², Institution: ¹Mathematical Institute, University of Oxford, ²Department of Psychiatry, University of Oxford, UK)

Clinical Diagnosis Ranking: 2

Feature selection: Manual

Selected features: Age, months since baseline, gender, race, marital status, diagnosis, cognitive tests (MMSE, CDRSB, ADAS11, ADAS-Cog13, RAVLT immediate, learning, forgetting and percent forgetting, FAQ) and APOE status.

Missing data: Random forest method deals with missing data automatically, by finding optimal splits with existing data only.

Confounder correction: None

Method category: Machine learning

Prediction method: For the prediction of clinical diagnosis, two random forest models are trained, the first for transitions from a healthy diagnosis, and the second for transitions from an MCI diagnosis. For AD individuals, authors assume that the diagnosis will not change. The training data was generated by ordering each participant's data by time, then associating the feature vector x with diagnosis y for each time horizon available for the participant. ADAS-Cog13 and Ventricles were not predicted. The PLOS paper describes a method similar to the original, but using different predictors and a single random forest.

Publication link: <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0211558>

Team: Tohka-Ciszek (Members: Jussi Tohka, Robert Ciszek Institution: A.I. Virtanen Institute for Molecular Sciences, University of Eastern Finland, Finland)

Overall Ranking: 19

Feature selection: Manual

Selected features:

- D2: diagnosis, gender, education, race, marital and APOE status, age, cognitive tests (CDRSB, ADAS11, ADAS-Cog13, MMSE, RAVLT learning, immediate and perc. forgetting, FAQ, MOCA, all Ecog), MRI volumes (ventricles, hippocampus, whole brain, entorhinal, fusiform, middle temporal, ICV)
- D3: diagnosis, age, gender, education, ethnicity, race, marital status, ADAS-Cog13, MMSE, MRI volumes as above

Missing data:

- SMNSR : A sub-model is trained for each subset of features which occurs without missing values. A specific catch-all subset is used for patients for which well performing whole measurement set cannot be found. For this data set, values are imputed using the median of k-nearest neighbours or replaced with -1.
- RandomForestLin : Imputation using mean values from the timepoints of the same subject (D2) or diagnostic category (for D3).

Confounder correction: None

Method category: Machine learning

Prediction method:

- Tohka-Ciszek-SMNSR : For ADAS-Cog13 prediction, a Sparse Multimodal Neighborhood Search Regression was used. This method first uses a linear regression model to estimate ADAS-Cog13 from the selected features belonging to the current subject and neighbour subjects, estimated based on a K-nearest neighbour algorithm. The forecasts from this model were passed to a gradient-boosted tree model, providing the final prediction. Ventricles and clinical diagnosis were not predicted.
- Tohka-Ciszek-RandomForestLin : To predict ADAS-Cog13 and Ventricles, a weighted average of two models was used: 1) a unimodal linear model 2) a linear model taking the response variable from the final time point and predictor variables from a time point before that. For diagnosis prediction, a random forest was trained using ADAS-Cog13, Ventricle/ICV, age and APOE status.

Confidence Intervals: Predicted score +/- cross-validation MAE.

Repository link: [https://github.com/jussitohka/tadpole\(RandomForestLin\)](https://github.com/jussitohka/tadpole(RandomForestLin)), <https://github.com/rciszek/SMNSR> (SMNSR)

Team: VikingAI ■ (Members: Bruno Jedynak¹, Kruti Pandya¹, Murat Bilgel², William Engels¹, Joseph Cole¹, Institutions: ¹Portland State University, USA, ²Laboratory of Behavioral Neuroscience, National Institute on Aging, National Institutes of Health, Baltimore, MD, USA)

Overall Ranking: 3-13

Feature selection: Manual

Selected features: Diagnosis, age, ADAS-Cog13, CDRSOB, MMSE, RAVLT immediate, tau, ventricles/ICV, hippocampus/ICV. Features were normalized prior to model fitting.

Missing data: Method automatically deals with missing data through Bayesian inference

Confounder correction: None

Method category: Data-driven disease progression model

Prediction method: For the prediction of ADAS-Cog13 and Ventricles, a latent-time parametric model was used, estimating a linear subject-specific model over the age of subjects, resulting in a disease progression score (DPS). The trajectories of biomarkers were assumed to be either sigmoidal functions or a sum of logistic basis functions over the DPS space. Some feature-specific parameters are estimated from the features' histograms, while the rest are optimized. Priors over the parameters to be optimized are set a-priori. Two submissions were made:

- VikingAI-Sigmoid: sigmoidal function as biomarker trajectory
- VikingAI-Logistic: the sum of 15 logistic basis functions as the biomarker trajectory

Confidence Intervals: Bayesian predictive intervals

In addition to the above entries, the organisers also included several benchmark algorithms as well as some extra predictions: (1) two ensemble predictions averaging all the predictions submitted by the participants, and (2) 62 randomised predictions which can tell how likely it is that top submissions obtained their scores due to chance, by building a null distribution of the performance metrics. The

Source code of some benchmarks (*BenchmarkLastVisit*, *BenchmarkMixedEffectsAPOE*, *BenchmarkSVM*) was offered to participants before the conference deadline, as a starting point for making predictions.

Benchmark: BenchmarkLastVisit ■ (Authors: Daniel Alexander, Razvan Marinescu, Institutions: University College London, UK, Massachusetts Institute of Technology, USA)

Overall Ranking: 40

Feature selection: None

Selected features: ADAS-Cog13, ventricle volume, diagnosis

Missing data: Not required

Confounder correction: None

Method category: Regression

Prediction method: For ADAS-Cog13 and Ventricles, the last available measure is used, otherwise the average for the current diagnostic group is used. Confidence intervals are set to default widths of 2 for ADAD13 and 0.001 for Ventricles/ICV. For prediction of clinical diagnosis, the last available diagnosis is used with probability 100%, and 0% probability for the other diagnoses.

Confidence Intervals: hard-coded

Repository link: <https://github.com/noxtoby/TADPOLE/blob/master/evaluation>

Benchmark: BenchmarkMixedEffects ■ (Author: Razvan Marinescu, Daniel Alexander, Institution: University College London, UK, Massachusetts Institute of Technology, USA)

Overall Ranking: 10-18

Feature selection: None

Selected features: ADAS-Cog13, ventricle volume, diagnosis (, APOE status)

Missing data: Automatic, since model is univariate.

Confounder correction: APOE status was used as covariate in the linear mixed effects model

Method category: Regression

Prediction method: Linear Mixed Effects Model with age at visit as the predictor variable. Model was fitted independently for ADAS-Cog13 and Ventricles. Predictions for clinical diagnosis were derived from the corresponding ADAS-Cog13 forecasts, using three Gaussian likelihood models for CN, MCI and AD. The likelihoods for diagnostic classes were finally converted to probabilities by normalisation. Default values were used for confidence intervals. Two predictions were made:

- BenchmarkMixedEffectsAPOE: the slope of the population trajectory was stratified by APOE status
- BenchmarkMixedEffects: as above but without APOE

Confidence Intervals: hard-coded

Repository link: <https://github.com/noxtoby/TADPOLE/blob/master/evaluation>

Benchmark: BenchmarkSVM ■ (Author: Esther Bron, Institution: Erasmus MC)

Overall Ranking: 34-35

Feature selection: Manual

Selected features: Diagnosis, age, ADAS-Cog13, Ventricles, ICV, APOE

Missing data: Fill-in using average value of biomarker from past visits of the same subject, otherwise population average.

Confounder correction: None

Method category: Machine Learning

Prediction method: For the prediction of clinical diagnosis, a probabilistic SVM was used based on the selected features, while for the prediction of ADAS-Cog13 and Ventricles, a Support Vector Regressor (SVR) was used. All SVM/SVRs used linear kernels. Default values were used for confidence intervals.

Confidence Intervals: hard coded

Repository link: <https://github.com/noxtoby/TADPOLE/blob/master/evaluation>

Benchmark: RandomisedBest ■ (Author: Razvan Marinescu, Institutions: University College London, UK, Massachusetts Institute of Technology, USA)

Overall Ranking: 15

Feature selection: Manual

Selected features: Diagnosis, age, ADAS-Cog13, Ventricles, ICV

Missing data: Fill-in using last available measurement

Confounder correction: None

Method category: Regression

Prediction method: The method aims to construct a null distribution of values, to check how likely a high score could be obtained by chance alone. Starting from the simplest prediction method, i.e. BenchmarkLastVisit which simply takes the last available measure, 62 randomised predictions were created (as many as the total number of predictions in D2) by adding random perturbations to the predictions. For Diagnosis prediction, the probability of controls and MCI subjects to convert within 1 year was computed from ADNI historical data, and then each control and MCI was randomly chosen to convert with those probabilities. For ADAS-Cog13 and Ventricles, random uniform noise was added to the predictions as follows:

- ADAS: $\text{new_adas} \sim \text{last_available_adas} + U(0, 7)$
- Ventricles: $\text{new_ventricles} \sim \text{last_available_ventricles} + U(0, 0.01)$

The new values `new_adas` and `new_ventricles`, as well as the new diagnosis, were assigned to all 60 months, thus assuming no change across the 60 month predictions. All 62 different predictions were evaluated, and the RandomisedBest entry shows the best scores obtained by all 62 submissions in each category separately.

Confidence Intervals: hard-coded

Repository link: <https://github.com/noxtoby/TADPOLE/blob/master/evaluation>

7. Acknowledgements

TADPOLE Challenge has been organised by the European Progression Of Neurological Disease (EuroPOND) consortium, in collaboration with the ADNI. We thank all the participants and advisors, in particular Clifford R. Jack Jr. from Mayo Clinic, Rochester, United States and Bruno M. Jedynak from Portland State University, Portland, United States for useful input and feedback.

The organisers are extremely grateful to The Alzheimer’s Association, The Alzheimer’s Society and Alzheimer’s Research UK for sponsoring the challenge by providing the £30,000 prize fund and providing invaluable advice into its construction and organisation. Similarly, we thank the ADNI leadership and members of our advisory board and other members of the EuroPOND consortium for their valuable advice and support.

RVM was supported by the EPSRC Centre For Doctoral Training in Medical Imaging with grant EP/L016478/1 and by the Neuroimaging Analysis Center with grant NIH NIBIB NAC P41EB015902. NPO, FB, SK, and DCA are supported by EuroPOND, which is an EU Horizon 2020 project. ALY was supported by an EPSRC Doctoral Prize fellowship and by EPSRC grant EP/J020990/01. PG was supported by NIH grant NIBIB NAC P41EB015902 and by grant NINDS R01NS086905. DCA was supported by EPSRC grants J020990, M006093 and M020533. FB was supported by the NIHR UCLH Biomedical Research Centre and the AMY-PAD project, which has received support from the EU-EFPIA Innovative Medicines Initiatives 2 Joint Undertaking (AMYPAD project, grant 115952). EEB was supported by the Dutch Heart Foundation (PPP Allowance, 2018B011) and Medical Delta Diagnostics 3.0: Dementia and Stroke. The UCL-affiliated researchers received support from the NIHR UCLH Biomedical Research Centre. This project has received funding from the European Union Horizon 2020 research and innovation programme under grant agreement No 666992. Data collection and sharing for this project was funded by the Alzheimer’s Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012), and the Alzheimer’s Disease Cooperative Study (ADCS), funded by the National Institutes of Health Grant U19 AG010483. NPO is a UKRI Future Leaders Fellow (MRC MR/S03546X/1).

The work of VikingAI was facilitated in part by the Portland Institute for Computational Science and its resources acquired using NSF Grant DMS 1624776. MB was supported by the Intramural Research Program of the National Institute on Aging, NIH. BTTY, NS and MN were supported by the Singapore National Research Foundation (NRF) Fellowship (Class of 2017). Team SBIA was supported by grant R01 AG054409. Team DIKU has received funding from the European Union’s Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 721820.

8. Author Contributions

DCA, RVM, NPO, ALY, EEB and SK designed the challenge. RVM performed the evaluation of the algorithms and the analysis of the results, and drafted the manuscript. NPO performed the external validation analysis and drafted the corresponding text. All authors provided feedback on the manuscript. RVM, NPO, ALY, EB and DCA created the TADPOLE website. RVM

and NPO constructed the TADPOLE D1-D4 datasets. RVM, DCA and EB wrote benchmark scripts which were offered to participants before the deadline. DCA and NPO organised and ran the TADPOLE webinars. NF and FB provided valuable suggestions on the challenge design. AT and MW offered access to the ADNI database. All other co-authors participated in the challenge. EB lead the TADPOLE-SHARE effort to make the algorithms openly available for further reuse, in a standardised format.

9. Conflicts of Interest

SJK received fees for participation in a Roche Diagnostics advisory board and does paid consulting for DIADEM outside of the submitted work.

10. Ethical Standards

ADNI obtained all IRB approvals and met all ethical standards in the collection of data. The ADNI protocol was approved by the Institutional Review Boards of all of the participating institutions.

The fifty-one centers in the DHA clinical trial obtained approval from their local institutional review boards. Written informed consent was obtained from study participants, legally authorized representatives, or both, according to local guidelines.

The AIBL study, including the follow-up protocol and subsequent amendments and revisions to the protocol, was approved by the institutional human research ethics committees of Austin Health, St Vincent's Health, Hollywood Private Hospital and Edith Cowan University. All volunteers gave written informed consent before participating in study assessments, and the study was conducted in accordance with the Helsinki Declaration of 1975.

References

- Clement Abi Nader, Nicholas Ayache, Philippe Robert, Marco Lorenzi, Alzheimer's Disease Neuroimaging Initiative, et al. Monotonic gaussian process for spatio-temporal disease progression modeling in brain imaging data. *Neuroimage*, 205:116266, 2020.
- Genevera I Allen, Nicola Amoroso, Catalina Anghel, Venkat Balagurusamy, Christopher J Bare, Derek Beaton, Roberto Bellotti, David A Bennett, Kevin L Boehme, Paul C Boutros, et al. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. *Alzheimer's & Dementia*, 12(6):645–653, 2016.
- Ignacio Alvarez, Juan Manuel Górriz, Javier Ramírez, Diego Salas-Gonzalez, Míriam López, Fermín Segovia, Carlos García Puntónet, and Beatriz Prieto. Alzheimer's diagnosis using eigenbrains and support vector machines. In *International Work-Conference on Artificial Neural Networks*, pages 973–980. Springer, 2009.
- Rebecca R Andridge and Roderick JA Little. A review of hot deck imputation for survey non-response. *International statistical review*, 78(1):40–64, 2010.
- Thomas G Beach, Sarah E Monsell, Leslie E Phillips, and Walter Kukull. Accuracy of the clinical diagnosis of Alzheimer disease at National Institute on Aging Alzheimer Disease Centers, 2005–2010. *Journal of neuropathology and experimental neurology*, 71(4):266–273, 2012.
- Murat Bilgel, Jerry L Prince, Dean F Wong, Susan M Resnick, and Bruno M Jernigan. A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. *Neuroimage*, 134:658–670, 2016.

- Alexandre Bône, Olivier Colliot, and Stanley Durrleman. Learning distributions of shape trajectories from longitudinal datasets: a hierarchical model on a manifold of diffeomorphisms. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 9271–9280, 2018.
- Esther E Bron, Marion Smits, Wiesje M Van Der Flier, Hugo Vrenken, Frederik Barkhof, Philip Scheltens, Janne M Papma, Rebecca ME Steketee, Carolina Méndez Orellana, Rozanna Meijboom, et al. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. *NeuroImage*, 111:562–579, 2015.
- Isabella Castiglioni, Christian Salvatore, Javier Ramírez, and Juan Manuel Górriz. Machine-learning neuroimaging challenge for automated diagnosis of mild cognitive impairment: Lessons learnt. *Journal of neuroscience methods*, 302:10, 2018.
- Guangyu Chen, Hao Shu, Gang Chen, B Douglas Ward, Piero G Antuono, Zhijun Zhang, and Shi-Jiang Li. Staging Alzheimer’s disease risk by sequencing brain function and structure, cerebrospinal fluid, and cognition biomarkers. *Journal of Alzheimer’s Disease*, 54(3):983–993, 2016.
- Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, pages 785–794. ACM, 2016.
- Ruoxuan Cui, Manhua Liu, Alzheimer’s Disease Neuroimaging Initiative, et al. Rnn-based longitudinal analysis for diagnosis of alzheimer’s disease. *Computerized Medical Imaging and Graphics*, 73:1–10, 2019.
- Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845, 1988.
- Michael C Donohue, Hélène Jacqmin-Gadda, Mélanie Le Goff, Ronald G Thomas, Rema Ramman, Anthony C Gamst, Laurel A Beckett, Clifford R Jack Jr, Michael W Weiner, Jean-François Dartigues, et al. Estimating long-term multivariate progression from short-term data. *Alzheimer’s & Dementia*, 10(5):S400–S410, 2014.
- Rachelle S Doody, Valory Pavlik, Paul Massman, Susan Rountree, Eveleen Darby, and Wenyaw Chan. Erratum to: Predicting progression of Alzheimer’s disease. *Alzheimer’s research & therapy*, 2(3):14, 2010.
- Nguyen Thanh Duc, Seungjun Ryu, Muhammad Naveed Iqbal Qureshi, Min Choi, Kun Ho Lee, and Boreom Lee. 3d-deep learning based automatic diagnosis of alzheimer’s disease with joint mmse prediction using resting-state fmri. *Neuroinformatics*, 18(1):71–86, 2020.
- Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.
- Kathryn A Ellis, Ashley I Bush, David Darby, Daniela De Fazio, Jonathan Foster, Peter Hudson, Nicola T Lautenschlager, Nat Lenzo, Ralph N Martins, Paul Maruff, 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. *International psychogeriatrics*, 21(4):672–687, 2009.
- Luca Ferrarini, Walter M Palm, Hans Olofsen, Mark A van Buchem, Johan HC Reiber, and Faiza Admiraal-Behloul. Shape differences of the brain ventricles in alzheimer’s disease. *Neuroimage*, 32(3):1060–1069, 2006.

- Hubert M Fonteijn, Marc Modat, Matthew J Clarkson, Josephine Barnes, Manja Lehmann, Nicola Z Hobbs, Rachael I Scahill, Sarah J Tabrizi, Sebastien Ourselin, Nick C Fox, et al. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. *NeuroImage*, 60(3):1880–1889, 2012.
- Karl J Friston, Andrew P Holmes, Keith J Worsley, J-P Poline, Chris D Frith, and Richard SJ Frackowiak. Statistical parametric maps in functional imaging: a general linear approach. *Human brain mapping*, 2(4):189–210, 1994.
- Sara Garbarino, Marco Lorenzi, Neil P Oxtoby, Elisabeth J Vinke, Razvan V Marinescu, Arman Eshaghi, M Arfan Ikram, Wiro J Niessen, Olga Ciccarelli, Frederik Barkhof, et al. Differences in topological progression profile among neurodegenerative diseases from imaging data. *Elife*, 8:e49298, 2019.
- Mostafa Mehdipour Ghazi, Mads Nielsen, Akshay Pai, Marc Modat, M Jorge Cardoso, Sébastien Ourselin, and Lauge Sørensen. Robust parametric modeling of Alzheimer's disease progression. *arXiv preprint arXiv:1908.05338*, 2019.
- FL Giesel, HK Hahn, PA Thomann, E Widjaja, E Wignall, H von Tengg-Kobligk, J Pantel, PD Griffiths, HO Peitgen, J Schroder, et al. Temporal horn index and volume of medial temporal lobe atrophy using a new semiautomated method for rapid and precise assessment. *American journal of neuroradiology*, 27(7):1454–1458, 2006.
- Katherine R Gray, Paul Aljabar, Rolf A Heckemann, Alexander Hammers, Daniel Rueckert, Alzheimer's Disease Neuroimaging Initiative, et al. Random forest-based similarity measures for multi-modal classification of alzheimer's disease. *NeuroImage*, 65:167–175, 2013.
- Joseph H Grochowalski, Ying Liu, and Karen L Siedlecki. Examining the reliability of ADAS-Cog change scores. *Aging, Neuropsychology, and Cognition*, 23(5):513–529, 2016.
- Xiao Han, Jorge Jovicich, David Salat, Andre van der Kouwe, Brian Quinn, Silvester Czanner, Evelina Busa, Jenni Pacheco, Marilyn Albert, Ronald Killiany, et al. Reliability of MRI-derived measurements of human cerebral cortical thickness: the effects of field strength, scanner upgrade and manufacturer. *Neuroimage*, 32(1):180–194, 2006.
- Lei Huang, Yan Jin, Yaozong Gao, Kim-Han Thung, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Longitudinal clinical score prediction in alzheimer's disease with soft-split sparse regression based random forest. *Neurobiology of aging*, 46:180–191, 2016.
- Yasser Iturria-Medina, Roberto C Sotero, Paule J Toussaint, José María Mateos-Pérez, Alan C Evans, Michael W Weiner, Paul Aisen, Ronald Petersen, Clifford R Jack, William Jagust, et al. Early role of vascular dysregulation on late-onset Alzheimer's disease based on multi-factorial data-driven analysis. *Nature communications*, 7:11934, 2016.
- Clifford R Jack Jr, David S Knopman, William J Jagust, Leslie M Shaw, Paul S Aisen, Michael W Weiner, Ronald C Petersen, and John Q Trojanowski. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. *The Lancet Neurology*, 9(1):119–128, 2010.
- Bruno M Jernak, Andrew Lang, Bo Liu, Elyse Katz, Yanwei Zhang, Bradley T Wyman, David Raunig, C Pierre Jernak, Brian Caffo, Jerry L Prince, et al. A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease Neuroimaging Initiative cohort. *Neuroimage*, 63(3):1478–1486, 2012.
- Taeho Jo, Kwangsik Nho, and Andrew J Saykin. Deep learning in alzheimer's disease: diagnostic classification and prognostic prediction using neuroimaging data. *Frontiers in aging neuroscience*, 11:220, 2019.

- JD Kalbfleisch and Jerald Franklin Lawless. The analysis of panel data under a Markov assumption. *Journal of the American Statistical Association*, 80(392):863–871, 1985.
- S Kiebel and AP Holmes. *The general linear model*, volume 8. chapter, 2007.
- Stefan Klöppel, Cynthia M Stonnington, Carlton Chu, Bogdan Draganski, Rachael I Scahill, Jonathan D Rohrer, Nick C Fox, Clifford R Jack Jr, John Ashburner, and Richard SJ Frackowiak. Automatic classification of MR scans in Alzheimer’s disease. *Brain*, 131(3):681–689, 2008.
- Igor Koval, Alexandre Bône, Maxime Louis, Simona Bottani, Arnaud Marcoux, Jorge Samper-Gonzalez, Ninon Burgos, Benjamin Charlier, Anne Bertrand, Stéphane Epelbaum, et al. Simulating Alzheimer’s disease progression with person-alised digital brain models. *Inria preprint*, 2018.
- AV Lebedev, Eric Westman, GJP Van Westen, MG Kramberger, Arvid Lundervold, Dag Aarsland, H Soininen, I Kłoszewska, P Mecocci, M Tsolaki, et al. Random forest ensembles for detection and prediction of alzheimer’s disease with a good between-cohort robustness. *NeuroImage: Clinical*, 6:115–125, 2014.
- Garam Lee, Kwangsik Nho, Byungkon Kang, Kyung-Ah Sohn, and Dokyoon Kim. Predicting alzheimer’s disease progression using multi-modal deep learning approach. *Scientific reports*, 9(1):1–12, 2019.
- Hongming Li, Mohamad Habes, David A Wolk, Yong Fan, Alzheimer’s Disease Neuroimaging Initiative, et al. A deep learning model for early prediction of alzheimer’s disease dementia based on hippocampal magnetic resonance imaging data. *Alzheimer’s & Dementia*, 15(8):1059–1070, 2019.
- Weiming Lin, Tong Tong, Qinquan Gao, Di Guo, Xiaofeng Du, Yonggui Yang, Gang Guo, Min Xiao, Min Du, Xiaobo Qu, et al. Convolutional neural networks-based mri image analysis for the alzheimer’s disease prediction from mild cognitive impairment. *Frontiers in neuroscience*, 12:777, 2018.
- Xiaojing Long, Lifang Chen, Chunxiang Jiang, Lijuan Zhang, Alzheimer’s Disease Neuroimaging Initiative, et al. Prediction and classification of Alzheimer disease based on quantification of MRI deformation. *PloS one*, 12(3):e0173372, 2017.
- Marco Lorenzi, Maurizio Filippone, Giovanni B Frisoni, Daniel C Alexander, Sébastien Ourselin, Alzheimer’s Disease Neuroimaging Initiative, et al. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer’s disease. *NeuroImage*, 2017.
- Marco Lorenzi, Maurizio Filippone, Giovanni B Frisoni, Daniel C Alexander, Sébastien Ourselin, Alzheimer’s Disease Neuroimaging Initiative, et al. Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer’s disease. *NeuroImage*, 190:56–68, 2019.
- Klaus H Maier-Hein, Peter F Neher, Jean-Christophe Houde, Marc-Alexandre Côté, Eleftherios Garyfallidis, Jidan Zhong, Maxime Chamberland, Fang-Cheng Yeh, Ying-Chia Lin, Qing Ji, et al. The challenge of mapping the human connectome based on diffusion tractography. *Nature communications*, 8(1):1349, 2017.
- Lena Maier-Hein, Matthias Eisenmann, Annika Reinke, Sinan Onogur, Marko Stankovic, Patrick Scholz, Tal Arbel, Hrvoje Bogunovic, Andrew P Bradley, Aaron Carass, et al. Why rankings of biomedical image analysis competitions should be interpreted with care. *Nature communications*, 9(1):5217, 2018.

- Razvan V Marinescu, Neil P Oxtoby, Alexandra L Young, Esther E Bron, Arthur W Toga, Michael W Weiner, Frederik Barkhof, Nick C Fox, Stefan Klein, Daniel C Alexander, et al. Tadpole Challenge: prediction of longitudinal evolution in Alzheimer's disease. *arXiv preprint arXiv:1805.03909*, 2018.
- Răzvan V Marinescu, Arman Eshaghi, Marco Lorenzi, Alexandra L Young, Neil P Oxtoby, Sara Garbarino, Sebastian J Crutch, Daniel C Alexander, Alzheimer's Disease Neuroimaging Initiative, et al. Dive: A spatiotemporal progression model of brain pathology in neurodegenerative disorders. *NeuroImage*, 192:166–177, 2019.
- Jussi Mattila, Juha Koikkalainen, Arho Virkki, Anja Simonsen, Mark van Gils, Gunhild Walde- mar, Hilka Soininen, Jyrki Lötjönen, Alzheimer's Disease Neuroimaging Initiative, et al. A disease state fingerprint for evaluation of Alzheimer's disease. *Journal of Alzheimer's Disease*, 27(1):163–176, 2011.
- Rahul Mazumder, Trevor Hastie, and Robert Tibshirani. Spectral regularization algorithms for learning large incomplete matrices. *Journal of machine learning research*, 11(Aug):2287–2322, 2010.
- Dev Mehta, Robert Jackson, Gaurav Paul, Jiong Shi, and Marwan Sabbagh. Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective eps for 2010-2015. *Expert opinion on investigational drugs*, 26(6):735–739, 2017.
- Bjoern H Menze, Andras Jakab, Stefan Bauer, Jayashree Kalpathy-Cramer, Keyvan Farahani, Justin Kirby, Yuliya Burren, Nicole Porz, Johannes Slotboom, Roland Wiest, et al. The multimodal brain tumor image segmentation benchmark (BRATS). *IEEE Transactions on Medical Imaging*, 34(10):1993–2024, 2014.
- Elaheh Moradi, Antonietta Pepe, Christian Gaser, Heikki Huttunen, Jussi Tohka, Alzheimer's Disease Neuroimaging Initiative, et al. Machine learning framework for early MRI-based Alzheimer's conversion prediction in MCI subjects. *Neuroimage*, 104:398–412, 2015.
- Jonathan H Morra, Zhuowen Tu, Liana G Apostolova, Amity E Green, Arthur W Toga, and Paul M Thompson. Comparison of adaboost and support vector machines for detecting alzheimer's disease through automated hippocampal segmentation. *IEEE transactions on medical imaging*, 29(1):30–43, 2009.
- Keelin Murphy, Bram Van Ginneken, Joseph M Reinhardt, Sven Kabus, Kai Ding, Xiang Deng, Kunlin Cao, Kaifang Du, Gary E Christensen, Vincent Garcia, et al. Evaluation of registration methods on thoracic CT: the EMPIRE10 challenge. *IEEE Transactions on Medical Imaging*, 30(11):1901–1920, 2011.
- Sean M Nestor, Raul Rupsingh, Michael Borrie, Matthew Smith, Vittorio Accomazzi, Jennie L Wells, Jennifer Fogarty, Robert Bartha, and Alzheimer's Disease Neuroimaging Initiative. Ventricular enlargement as a possible measure of alzheimer's disease progression validated using the alzheimer's disease neuroimaging initiative database. *Brain*, 131(9):2443–2454, 2008.
- Kenichi Oishi, Andreia Faria, Hangyi Jiang, Xin Li, Kazi Akhter, Jiangyang Zhang, John T Hsu, Michael I Miller, Peter CM van Zijl, Marilyn Albert, et al. Atlas-based whole brain white matter analysis using large deformation diffeomorphic metric mapping: application to normal elderly and Alzheimer's disease participants. *Neuroimage*, 46(2):486–499, 2009.
- Neil P Oxtoby, Alexandra L Young, David M Cash, Tammie LS Benzinger, Anne M Fagan, John C Morris, Randall J Bateman, Nick C Fox, Jonathan M Schott, and Daniel C Alexander.

- Data-driven models of dominantly-inherited Alzheimer’s disease progression. *Brain*, 141(5): 1529–1544, 2018.
- Michael J Pencina, Ralph B D’Agostino Sr, Ralph B D’Agostino Jr, and Ramachandran S Vasan. Evaluating the added predictive ability of a new marker: from area under the ROC curve to reclassification and beyond. *Statistics in medicine*, 27(2):157–172, 2008.
- Gautam Prasad, Talia M Nir, Arthur W Toga, and Paul M Thompson. Tractography density and network measures in Alzheimer’s disease. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 692–695. IEEE, 2013.
- Joseph F Quinn, Rema Raman, Ronald G Thomas, Karin Yurko-Mauro, Edward B Nelson, Christopher Van Dyck, James E Galvin, Jennifer Emond, Clifford R Jack, Michael Weiner, et al. Docosahexaenoic acid supplementation and cognitive decline in alzheimer disease: a randomized trial. *Jama*, 304(17):1903–1911, 2010.
- Ashish Raj, Amy Kuceyeski, and Michael Weiner. A network diffusion model of disease progression in dementia. *Neuron*, 73(6):1204–1215, 2012.
- Lars Lau Raket. Disease progression modeling in alzheimer’s disease: insights from the shape of cognitive decline. *medRxiv*, 2019.
- J Ramírez, JM Górriz, A Ortiz, FJ Martínez-Murcia, F Segovia, D Salas-Gonzalez, D Castillo-Barnes, IA Illán, CG Puntonet, Alzheimer’s Disease Neuroimaging Initiative, et al. Ensemble of random forests one vs. rest classifiers for mci and ad prediction using anova cortical and subcortical feature selection and partial least squares. *Journal of neuroscience methods*, 302: 47–57, 2018.
- Martin Reuter, Nicholas J Schmansky, H Diana Rosas, and Bruce Fischl. Within-subject template estimation for unbiased longitudinal image analysis. *Neuroimage*, 61(4):1402–1418, 2012.
- Chen S, Qi K, and Qiu D. Is MRI all we need? Prediction of conversion between normal cognitive function, mild cognitive impairment and Alzheimer’s disease. *ISMRM*, 2018.
- Mert R Sabuncu, Rahul S Desikan, Jorge Sepulcre, Boon Thye T Yeo, Hesheng Liu, Nicholas J Schmansky, Martin Reuter, Michael W Weiner, Randy L Buckner, Reisa A Sperling, et al. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. *Archives of neurology*, 68(8):1040–1048, 2011.
- Diego Salas-Gonzalez, Juan Manuel Górriz, Javier Ramírez, M López, I Alvarez, F Segovia, R Chaves, and CG Puntonet. Computer-aided diagnosis of alzheimer’s disease using support vector machines and classification trees. *Physics in Medicine & Biology*, 55(10):2807, 2010.
- Alessia Sarica, Antonio Cerasa, and Aldo Quattrone. Random forest algorithm for the classification of neuroimaging data in alzheimer’s disease: a systematic review. *Frontiers in aging neuroscience*, 9:329, 2017.
- Rachael I Scahill, Jonathan M Schott, John M Stevens, Martin N Rossor, and Nick C Fox. Mapping the evolution of regional atrophy in Alzheimer’s disease: unbiased analysis of fluid-registered serial MRI. *Proceedings of the National Academy of Sciences*, 99(7):4703–4707, 2002.
- Jean-Baptiste Schiratti, Stéphanie Allasonnière, Olivier Colliot, and Stanley Durrleman. A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. *Journal of Machine Learning Research*, 18(133):1–33, 2017.

- Lisa C Silbert, JF Quinn, MM Moore, E Corbridge, MJ Ball, G Murdoch, G Sexton, and JA Kaye. Changes in premorbid brain volume predict alzheimer's disease pathology. *Neurology*, 61(4):487–492, 2003.
- Simeon Spasov, Luca Passamonti, Andrea Duggento, Pietro Lio, Nicola Toschi, Alzheimer's Disease Neuroimaging Initiative, et al. A parameter-efficient deep learning approach to predict conversion from mild cognitive impairment to alzheimer's disease. *Neuroimage*, 189:276–287, 2019.
- Daniel J Stekhoven and Peter Bühlmann. Missforest—non-parametric missing value imputation for mixed-type data. *Bioinformatics*, 28(1):112–118, 2011.
- Vikram Venkatraghavan, Esther E Bron, Wiro J Niessen, Stefan Klein, and Alzheimer's Disease Neuroimaging Initiative. Disease progression timeline estimation for Alzheimer's disease using discriminative event based modeling. *NeuroImage*, 186:518–532, 2019.
- Victor L Villemagne, Samantha Burnham, Pierrick Bourgeat, Belinda Brown, Kathryn A Ellis, Olivier Salvado, Cassandra Szoek, S Lance Macaulay, Ralph Martins, Paul Maruff, et al. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic alzheimer's disease: a prospective cohort study. *The Lancet Neurology*, 12(4):357–367, 2013.
- Xiang Wang, David Sontag, and Fei Wang. Unsupervised learning of disease progression models. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 85–94, 2014.
- Michael W Weiner, Dallas P Veitch, Paul S Aisen, Laurel A Beckett, Nigel J Cairns, Robert C Green, Danielle Harvey, Clifford R Jack Jr, William Jagust, John C Morris, et al. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. *Alzheimer's & Dementia*, 13(4):e1–e85, 2017.
- Alexandra L Young, Neil P Oxtoby, Pankaj Daga, David M Cash, Nick C Fox, Sebastien Ourselin, Jonathan M Schott, and Daniel C Alexander. A data-driven model of biomarker changes in sporadic Alzheimer's disease. *Brain*, 137(9):2564–2577, 2014.
- Alexandra L Young, Razvan V Marinescu, Neil P Oxtoby, Martina Bocchetta, Keir Yong, Nicholas C Firth, David M Cash, David L Thomas, Katrina M Dick, Jorge Cardoso, et al. Uncovering the heterogeneity and temporal complexity of neurodegenerative diseases with Subtype and Stage Inference. *Nature communications*, 9(1):4273, 2018.
- Jonathan Young, Marc Modat, Manuel J Cardoso, Alex Mendelson, Dave Cash, Sebastien Ourselin, Alzheimer's Disease Neuroimaging Initiative, et al. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical*, 2:735–745, 2013.
- Daoqiang Zhang, Yaping Wang, Luping Zhou, Hong Yuan, Dinggang Shen, Alzheimer's Disease Neuroimaging Initiative, et al. Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage*, 55(3):856–867, 2011.
- Juan Zhou, Efsthathios D Gennatas, Joel H Kramer, Bruce L Miller, and William W Seeley. Predicting regional neurodegeneration from the healthy brain functional connectome. *Neuron*, 73(6):1216–1227, 2012.

Appendix A. Creating the D1-D4 datasets

The data used from ADNI consists of: (1) CSF markers of amyloid-beta and tau deposition; (2) various imaging modalities such as magnetic resonance imaging (MRI), positron emission tomography (PET) using several tracers: Fluorodeoxyglucose (FDG, hypometabolism), AV45 (amyloid), AV1451 (tau) as well as diffusion tensor imaging (DTI); (3) cognitive assessments acquired in the presence of a clinical expert; (4) genetic information such as apolipoprotein E4 (APOE4) status extracted from DNA samples; and (5) general demographic information. Extracted features from this data were merged together into a final spreadsheet and made available on the LONI ADNI website.

The imaging data has been pre-processed with standard ADNI pipelines. For MRI scans, this included correction for gradient non-linearity, B1 non-uniformity correction and peak sharpening. [ADNI MRI pre-processing]. Meaningful regional features such as volume and cortical thickness were extracted using the Freesurfer cross-sectional and longitudinal pipelines (Reuter et al. (2012)). Each PET image (FDG, AV45, AV1451) had their frames co-registered, averaged across the six five-minute frames, standardised with respect to the orientation and voxel size, and smoothed to produce a uniform resolution of 8mm full-width/half-max (FWHM) (see <http://adni.loni.usc.edu/methods/pet-analysis/pre-processing/>). Standardised uptake value ratio (SUVR) measures for relevant regions-of-interest were extracted after registering the PET images to corresponding MR images using the SPM5 software (Friston et al. (1994)). Further details have been provided in the ADNI procedures manual. DTI scans were corrected for head motion and eddy-current distortion, skull-stripped, EPI-corrected, and finally aligned to the T1 scans using the pipeline from (Prasad et al. (2013)). Diffusion tensor summary measures were estimated based on the Eve white-matter atlas (Oishi et al. (2009)).

In addition to the standard datasets, we also created three leaderboard datasets LB1, LB2 and LB3 which mimic the D1, D2 and D4 datasets. These datasets were used by participants to preliminarily evaluate their algorithms before the competition deadline, and to compare their results on the leaderboard system (<https://tadpole.grand-challenge.org/Leaderboard/>).

Appendix B. Statistical testing

B.1 Differences in MAUC scores

For analysing whether the MAUC scores obtained by top algorithms are significantly different, we performed a bootstrapped hypothesis test (Efron and Tibshirani (1994)), since the significance test for comparing two AUC scores (DeLong et al. (1988)) does not extend to multiple classes. Let A and B be two TADPOLE algorithms and M_A and M_B be their associated MAUC scores. If $M_A > M_B$ on the full D4 test set, we want to confirm if algorithm A was significantly better than B , or if this was likely due to chance. We define the null hypothesis $H_0 : M_A = M_B$ and the alternative hypothesis $H_1 : M_A > M_B$. We then proceed as follows:

- Sample $N = 100$ random bootstraps D_i of the D4 test set with replacement.
- Compute the $M_A^{D_i}$ and $M_B^{D_i}$ based on the bootstrapped dataset. Repeat for all N bootstraps.
- Compute the p-value as $\sum_i I[M_A^{D_i} < M_B^{D_i}]/N$, which is the proportion of bootstrapped datasets where A performed worse than B .
- Accept/reject H_0 based on a 5% significance level.

B.2 Differences in MAE scores

For comparing differences in MAE scores, we applied the non-parametric Wilcoxon signed-rank test on paired samples of absolute errors across all visits of the D4 subjects. We chose the non-parametric Wilcoxon test because the input samples are not normally distributed, as they

represent absolute errors and are always positive. We report results based on a 5% significance level.

B.3 Differences between D2 and D3 forecasts

For comparing differences between the scores obtained by two algorithms on D2 vs D3 forecasts, we use an approach similar to comparing MAUC scores (section 8.4.2).

B.4 Comparisons with random guessing model

We used predictions from the *RandomisedBest* model to test whether a TADPOLE algorithm was significantly better performance than random guessing. If the MAE error of a TADPOLE algorithm was X and the performance of a random guess model was R , we wanted to test whether $H_0 : X = R$ (no difference in performance) or $H_1 : X < R$ (there is a difference in performance). For this, we performed the following:

- Generate 100 random predictions $R_i, i = [1, \dots, 100]$.
- Compute the p-value as $\sum_i I[X < R_i]$

Appendix C. Supplementary Results

Submission	Overall Rank	Diagnosis		ADAS-Cog13			Ventricles (% ICV)					
		Rank	MAUC	BCA	Rank	MAE	WES	CPA	Rank	MAE	WES	CPA
Billabong-UniAV45 	1	1	0.719	0.624	1-2	8.71	8.55	0.33	3-4	3.49	3.40	0.50
Billabong-Uni 	2	2	0.717	0.621	1-2	8.71	8.55	0.33	3-4	3.49	3.40	0.50
Billabong-MultiAV45 	3	3	0.661	0.562	3-4	12.95	12.71	0.42	1-2	3.16	3.08	0.47
Billabong-Multi 	4	4	0.658	0.552	3-4	12.95	12.71	0.42	1-2	3.16	3.08	0.47
Simple-SPMC-Plymouth2 	-	5	0.500	0.504	-	-	-	-	-	-	-	-
Simple-SPMC-Plymouth1 	-	6	0.500	0.499	-	-	-	-	-	-	-	-

Table 6: Results on custom prediction sets from two teams: Billabong and SPMC-Plymouth. SPMC-Plymouth predicted fewer subjects due to an incomplete submission, while Billabong used a prediction set similar to D3, but filled in missing data for cognitive tests and MRI with the last available measurement. SPMC-Plymouth only submitted predictions for clinical diagnosis, and obtained an MAUC score of 0.5. Results from Billabong show higher MAUC and BCA in diagnosis prediction compared to D3, but lower performance for ADAS-Cog13 and Ventricle volume prediction. Bold entries show best scores in this category.

Appendix D. External validation

This section reports an external validation experiment to verify, on an independent data set, the general finding that TADPOLE participants’ algorithms perform better than simple baselines and that consensus approaches perform as well or better than individual entries. To this end, we evaluate our primary performance metrics for each prediction task on external data for the *ConsensusMean* and *ConsensusMedian* algorithms, all benchmark algorithms, and the challenge-winning submission (Team Frog).

D.1 Data

External data comes from a clinical trial and an observational study. Data from the completed DHA clinical trial (Quinn et al. (2010)); <https://www.adcs.org/dha/>) was obtained from the Alzheimer’s Disease Cooperative Study (<https://www.adcs.org/>), including MMSE and ADAS-Cog scores from 402 AD participants over 18 months. A subset of approximately 90 also had T1-weighted MRI from which we calculated Ventricle volume (normalised by ICV)

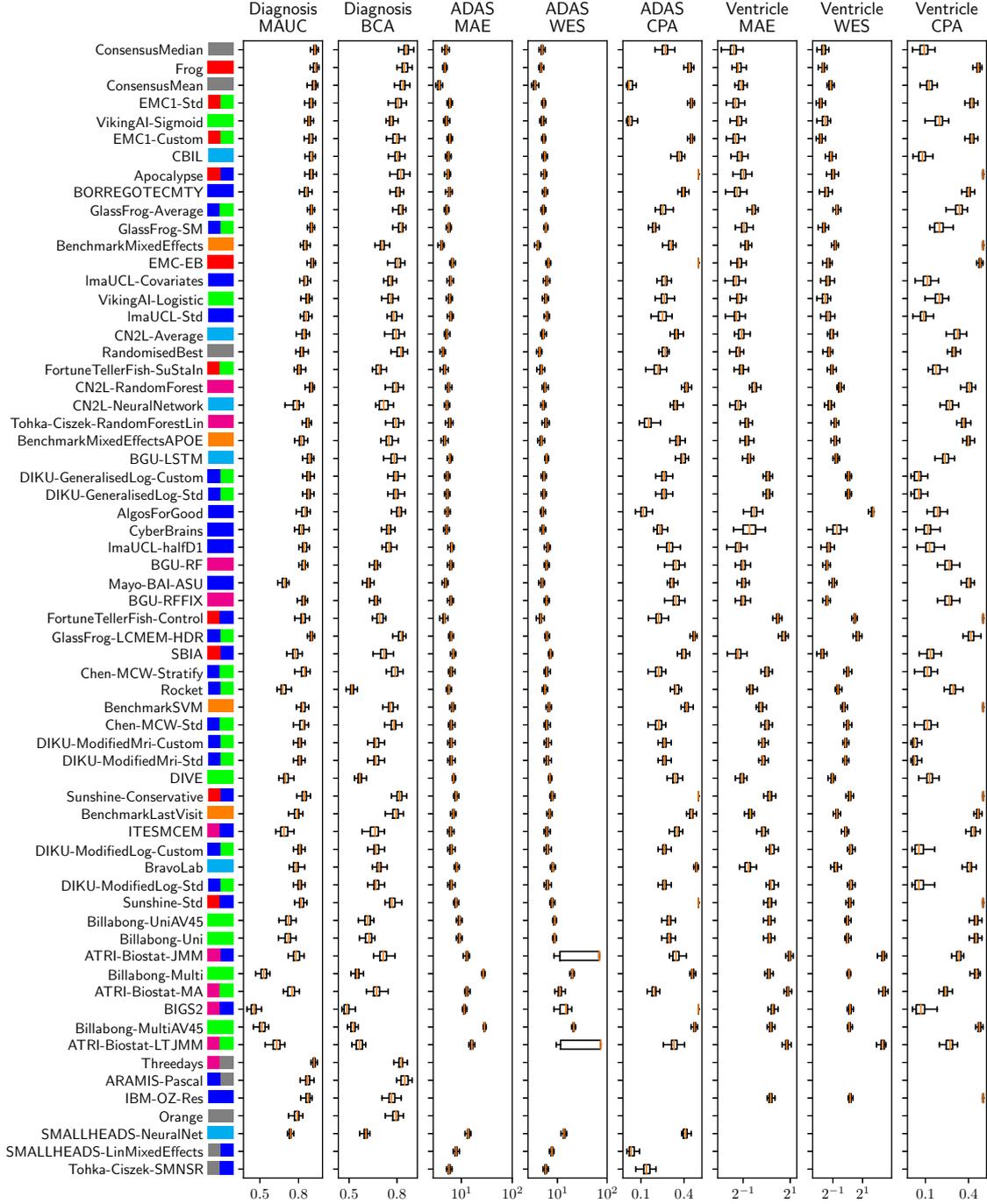


Figure 2: Distribution of performance metrics for clinical diagnosis (MAUC and BCA), ADAS-Cog13 (MAE, WES and CPA) and ventricle volume (MAE, WES and CPA) on the longitudinal D2 prediction set. For each entry, we plot the distribution of performance metrics derived using 50 bootstrap data sets drawn from the D4 test set. The submissions (rows) are in the same order as in Table 4. Entries are missing where teams did not make predictions for a particular target variable.

using FreeSurfer version 6.0.0. The ADAS-Cog11 scores in DHA were mapped to ADAS-Cog13 using a simple regression model trained on ADNI data. Clinical data was obtained from the Australian Imaging, Biomarkers, and Lifestyle Flagship Study of Ageing (AIBL) (Ellis et al.

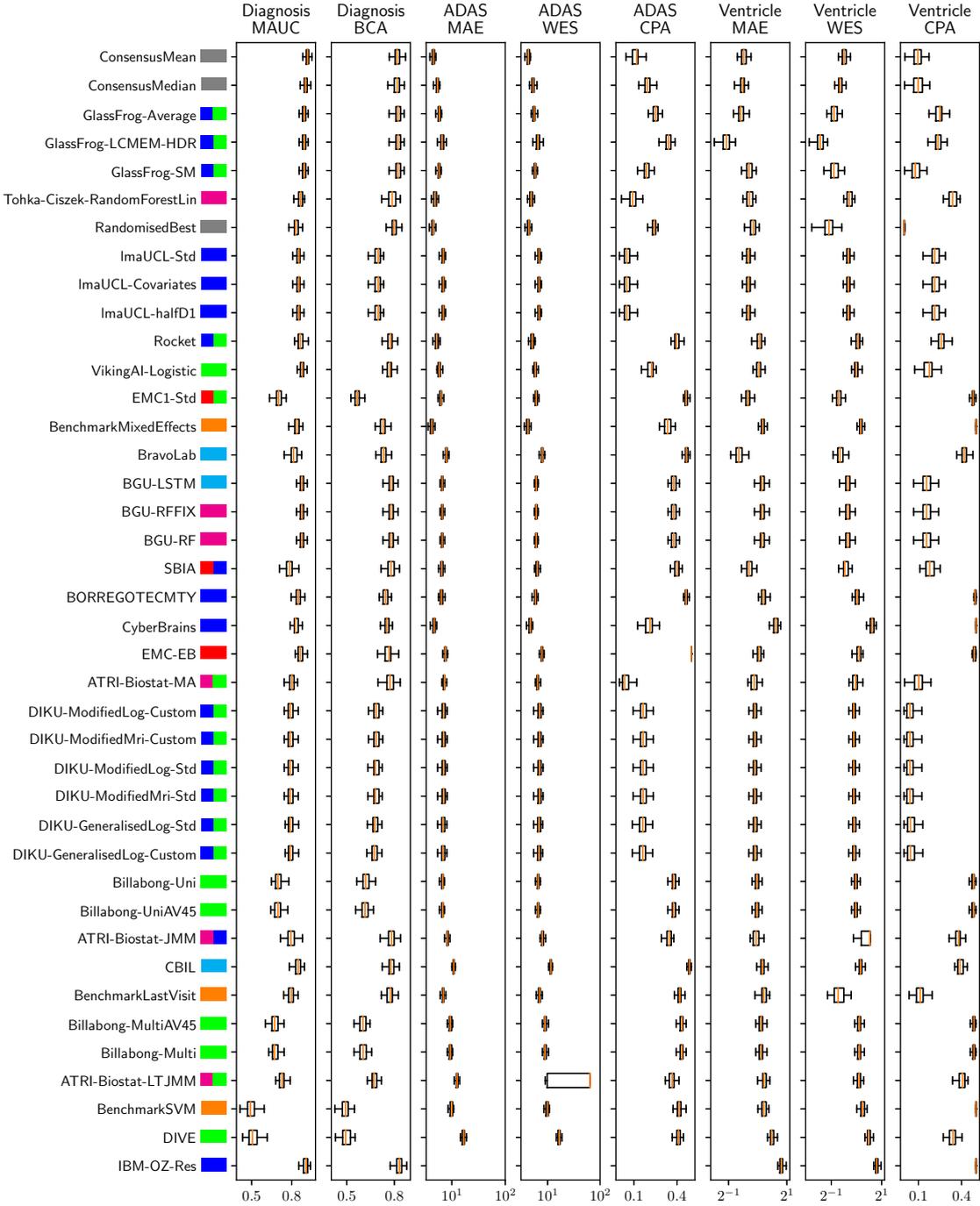


Figure 3: Box plots of performance metrics for clinical diagnosis (MAUC and BCA), ADAS-Cog13 (MAE, WES and CPA) and ventricle volume (MAE, WES and CPA) on the cross-sectional D3 prediction set. The submissions (rows) are in the same order as in Table 5. Some entries are missing because teams did not make predictions for those target variables.

(2009)), including MMSE and diagnosis from 857 participants (608 CN, 144 MCI, 105 AD) over a 5-year period (average followup interval 1.5 ± 1.9 years). The data from both studies was assembled into three external datasets as follows:

- **D5** (analogous to D3): The external cross-sectional prediction set containing *baseline data* from the DHA clinical trial and AIBL observational study.

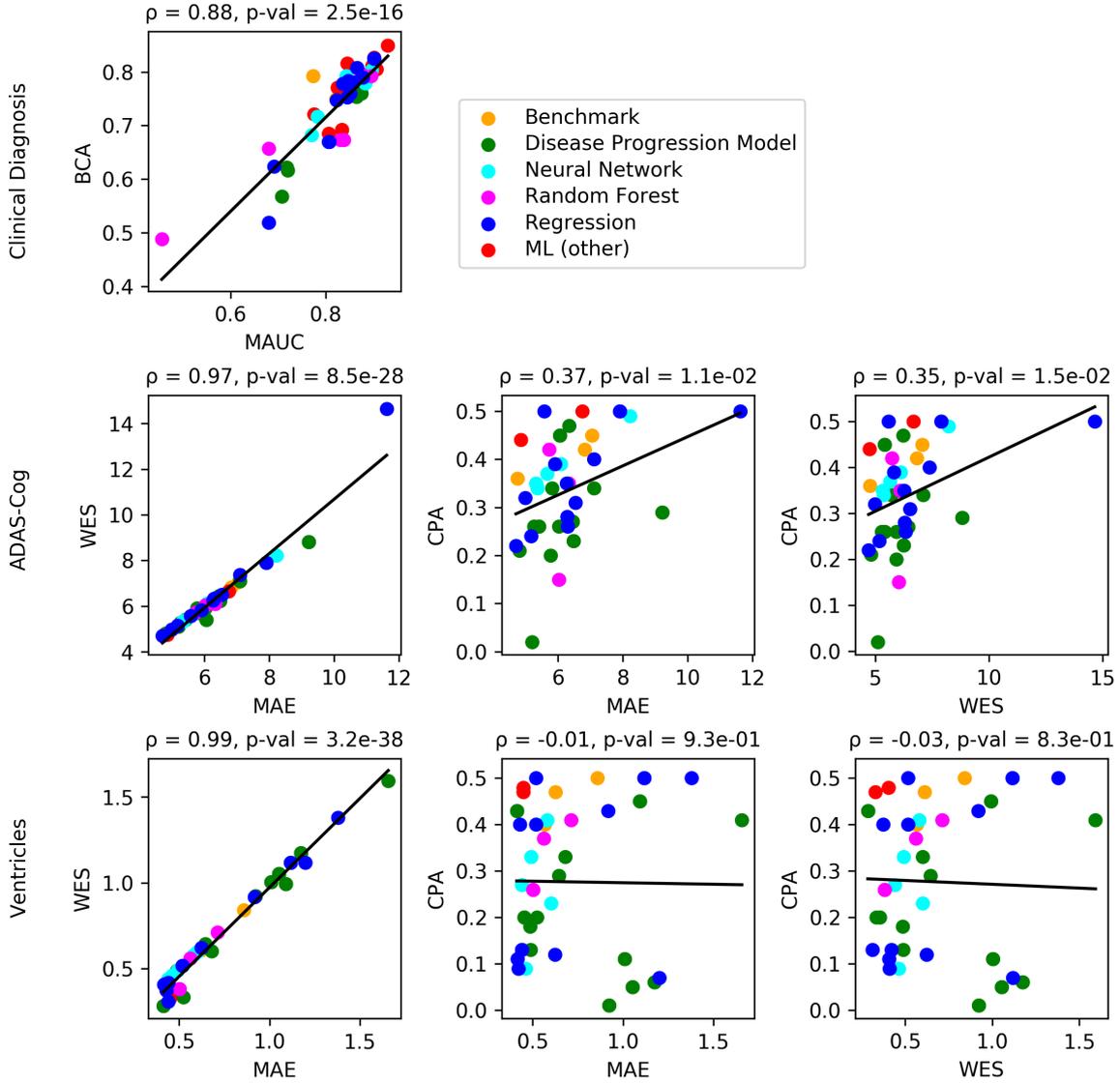


Figure 4: For D2 submissions, we show scatter plots of pairs of performance metrics for (top row) clinical diagnosis, (middle row) ADAS-Cog13 and (bottom row) Ventricles. Each dot is a participant submission, coloured according to the type of prediction algorithm used. Correlation coefficients and p-values are given above each subplot. A few outlier submissions with ADAS MAE > 20, ADAS WES > 40 or Ventricle WES > 3 were excluded from the analysis.

- **D6** (analogous to D4): The external test set containing data from the n=361 DHA participants with follow-up visits providing at least one of the following variables: (i) ADAS-Cog13 score, (ii) Ventricle Volume.
- **D7** (analogous to D4): The external test set containing data from the n=399 AIBL participants (318 CN, 47 MCI, 34 AD) with follow-up visits, which include only clinical diagnosis.

We could not run all participants' methods directly on the external data sets, since the external evaluation necessarily occurred after the original challenge was complete. Instead, we obtain predictions for each subject in the external data sets by copying predictions from the closest matching TADPOLE data point. We used a weighted L2-norm to find the closest match for each D5 case among TADPOLE subjects with the same sex and diagnosis. We eliminated 93

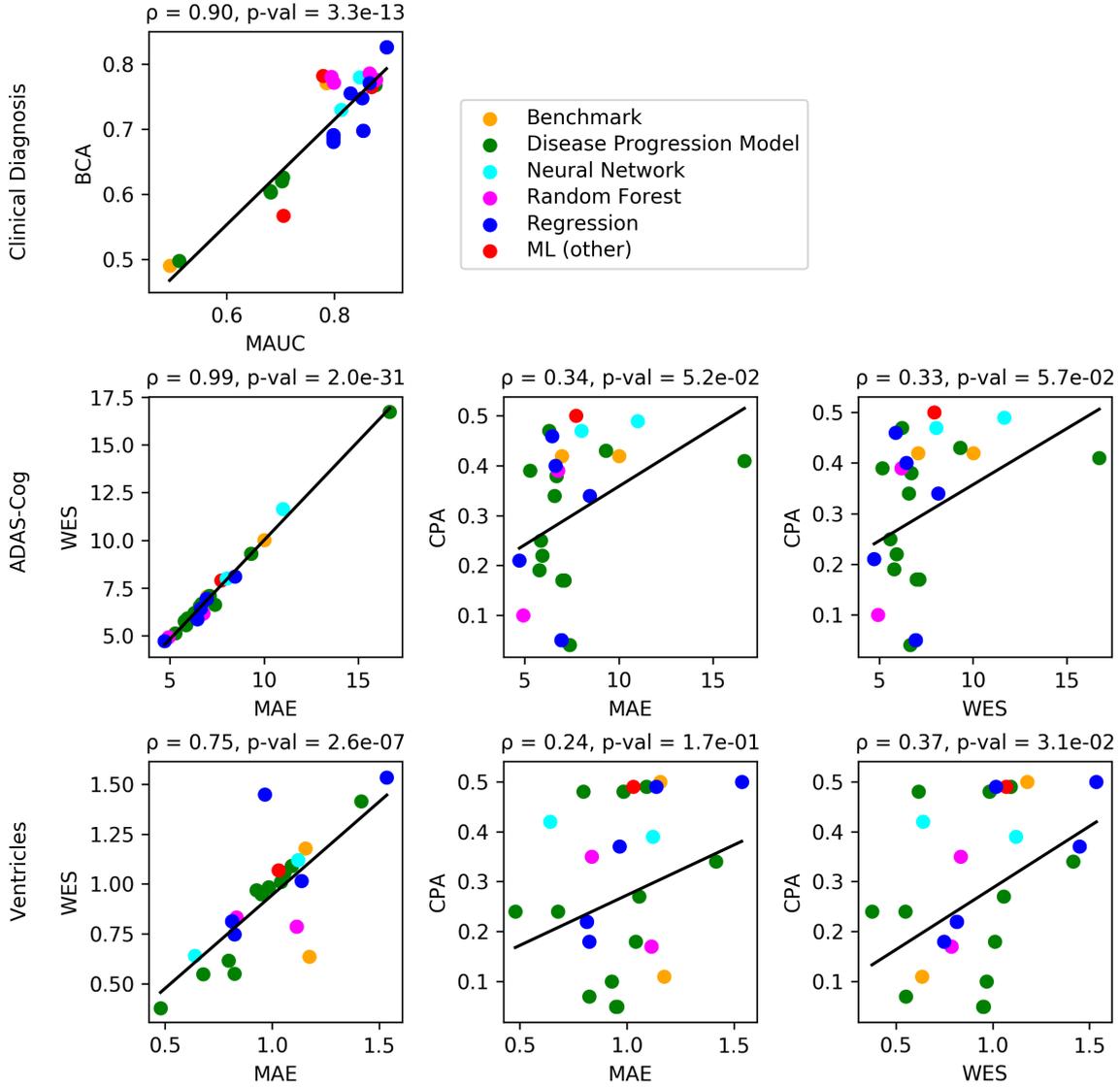


Figure 5: For D3 submissions, we show scatter plots of pairs of performance metrics for (top row) clinical diagnosis, (middle row) ADAS-Cog13 and (bottom row) Ventricles. Each dot is a participant submission, coloured according to the type of prediction algorithm used. Correlation coefficients and p-values are given above each subplot. A few outlier submissions with ADAS MAE > 20 , ADAS WES > 40 or Ventricle WES > 3 were excluded from the analysis.

subjects from D5 for whom we could not find a sufficiently close match in D3, i.e. within 7 years in age, 3 points on the MMSE test, 7 points on ADAS-Cog13, and Ventricles volume within $\pm 29\%$. These cutoff values are $\sigma/\sqrt{2}$ for each variable, where σ is the standard deviation of that variable in D5 (DHA and AIBL participants at baseline). The prediction of each quantity (diagnosis, ADAS-Cog13, ventricle volume) for a matched D5 case using a particular algorithm is then the prediction of the same quantity using the same algorithm on the matched D3 case; the matches do not vary among algorithms.

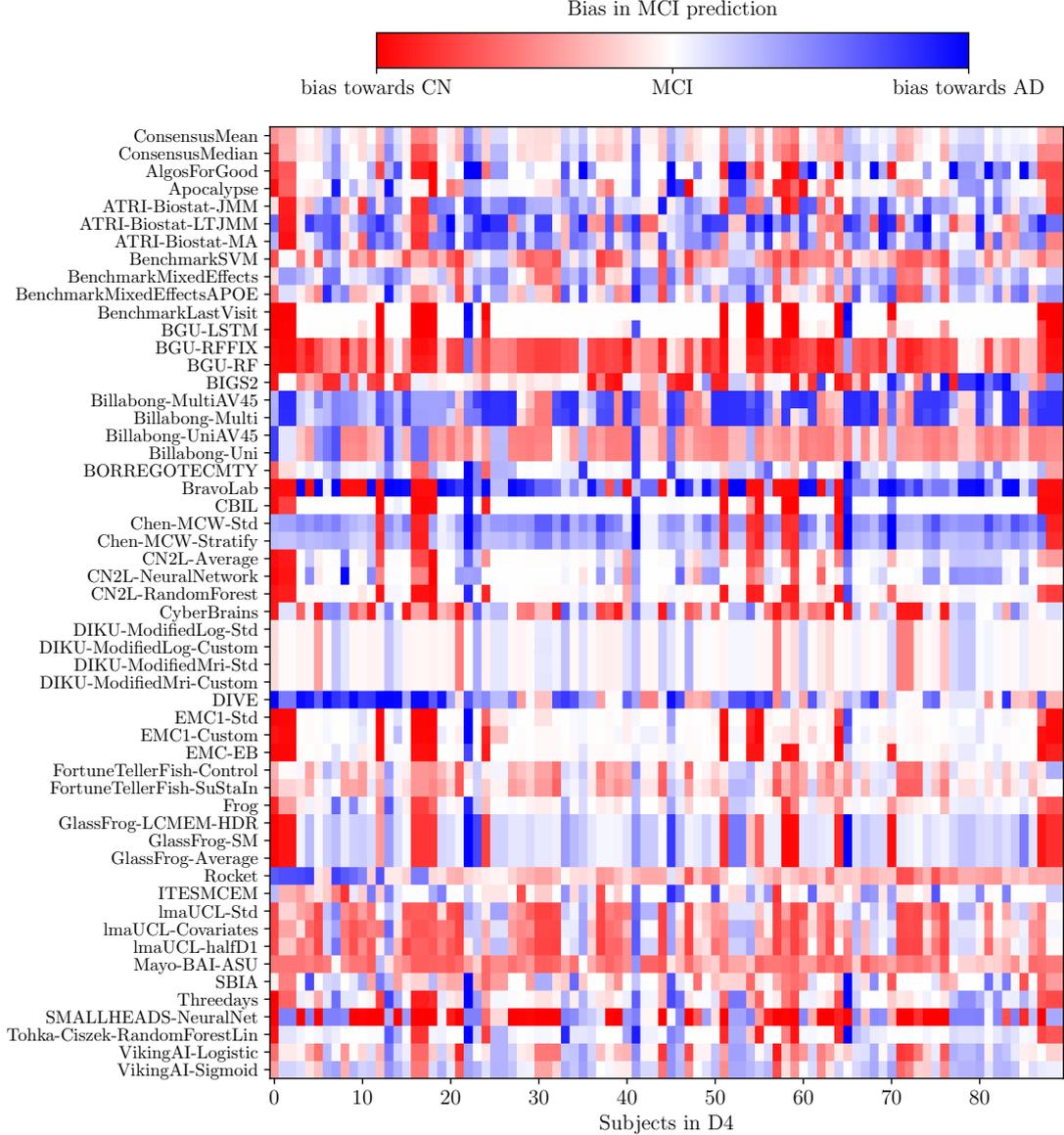


Figure 6: Bias in prediction of clinical diagnosis for MCI subjects only. X-axis shows individual subjects with designated MCI status at the clinical visit in D4, while the Y-axis shows TADPOLE algorithms. Red represents subjects which were predicted as CN with true diagnosis of MCI, while blue represents subjects predicted as AD with true diagnosis of MCI. Some algorithms show systematic biases either towards CN or AD.

D.2 External test results

In D3 we found acceptable matches for 268 of 361 DHA participants and all 399 AIBL participants in D5. Supplementary Table 7 shows prediction performance metrics for the test sets (D6 and D7), together with corresponding metrics from the internal test set D4 (reproduced from Figure 3.1).

Similar trends in predictive performance among methods arise on the external test sets as on the internal: consensus methods perform better overall than the best individual method, and substantially outperform all benchmarks; the representative strongly performing submission overall outperforms benchmarks. Results for diagnosis directly reflect that trend; consensus methods now outperform even Frog and Frog remains substantially ahead of all benchmarks.

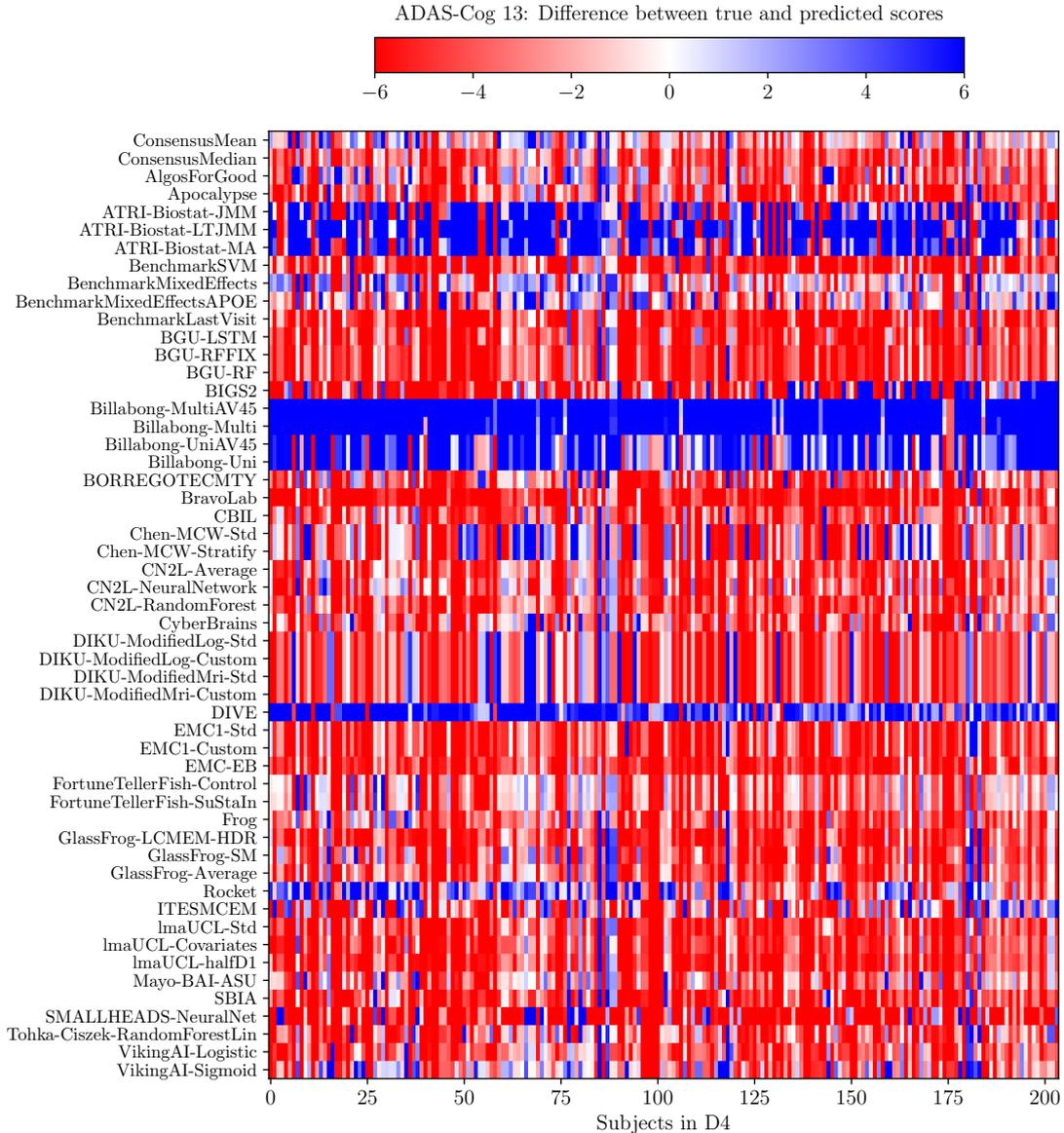


Figure 7: Bias in prediction of ADAS-Cog13. X-axis shows individual subjects with ADAS-Cog measurements in D4, while Y-axis shows TADPOLE algorithms. Red represents under-estimates while blue represents over-estimates. Most algorithms under-estimate ADAS-Cog measurements.

For ADAS-Cog13, as with the internal test set, Frog fails to outperform the simple benchmarks and only consensus methods approach viable (better than simple default) performance. Slightly anomalous results arise for ventricle volume prediction (Frog attains the worst MAE), which likely arises from the small sample; consensus methods still perform best. In comparison with internal performance metrics, clinical diagnosis MAUC is lower (indicating lower predictive accuracy) for D7 than D4; ADAS-Cog13 MAE is higher (lower predictive accuracy) on D6 than D4; ventricle volume MAE is slightly lower (higher predictive accuracy) on D6 than D4. In general, we expect the imperfect matching process to reduce performance on the unseen external data set compared to the internal data set, as we observe in diagnosis and ADAS-Cog13 prediction. However, the elimination of D5 subjects for which no good match is found may push average performance up by avoiding difficult/unusual subjects, as we observe in ventricle-volume prediction.

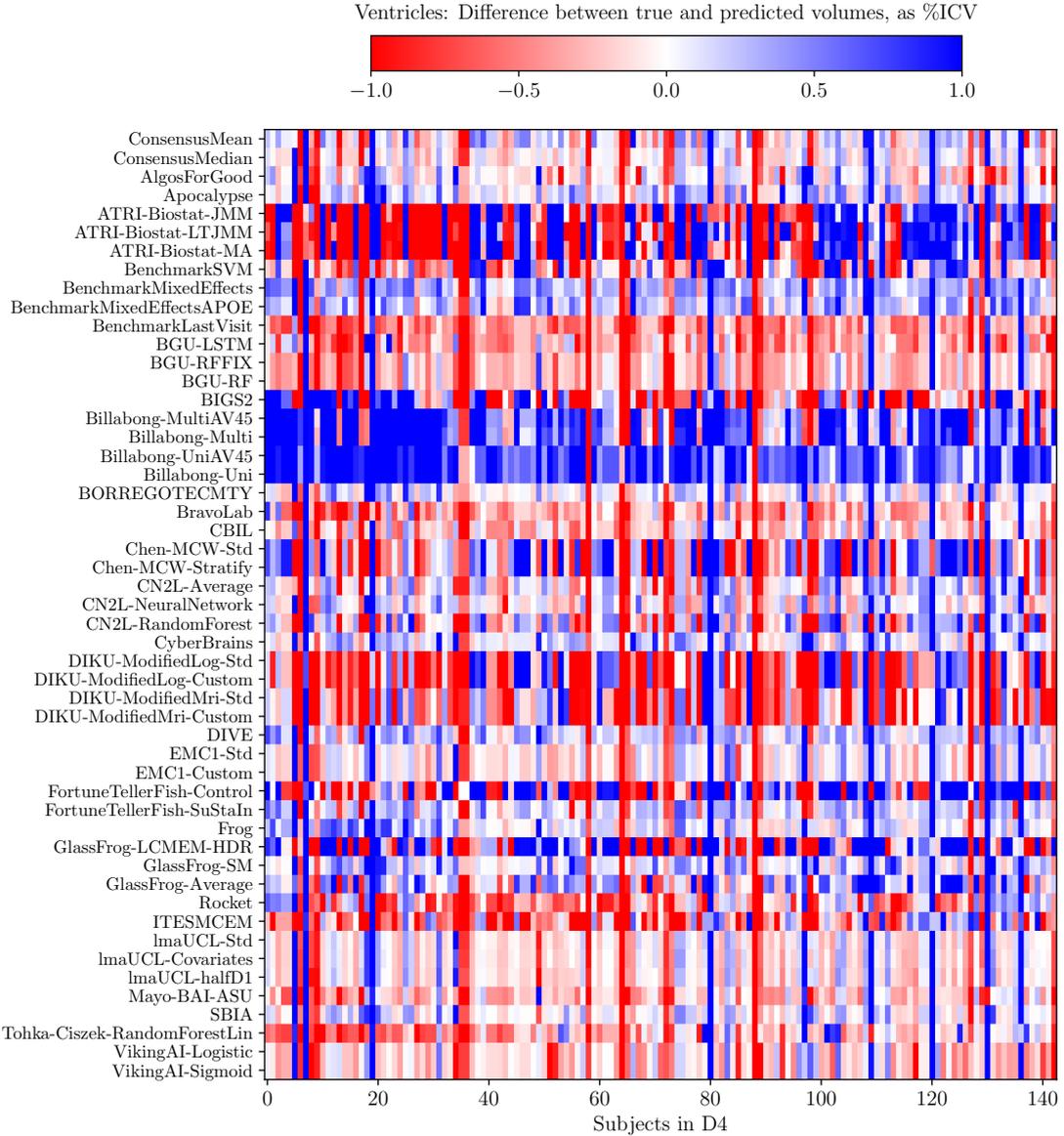


Figure 8: Bias in prediction of ventricle volume. X-axis shows individual subjects with Ventricle volume measurements in D4, while Y-axis shows TADPOLE algorithms. Red represents under-estimates while blue represents over-estimates. Some algorithms systematically underestimate or over-estimate ventricle volume.

D.3 Conclusion

External test set results reaffirm the strong performance of consensus methods in comparison to individual TADPOLE entries and baselines, as well as the particular difficulty in predicting cognitive test scores, such as ADAS-Cog13.

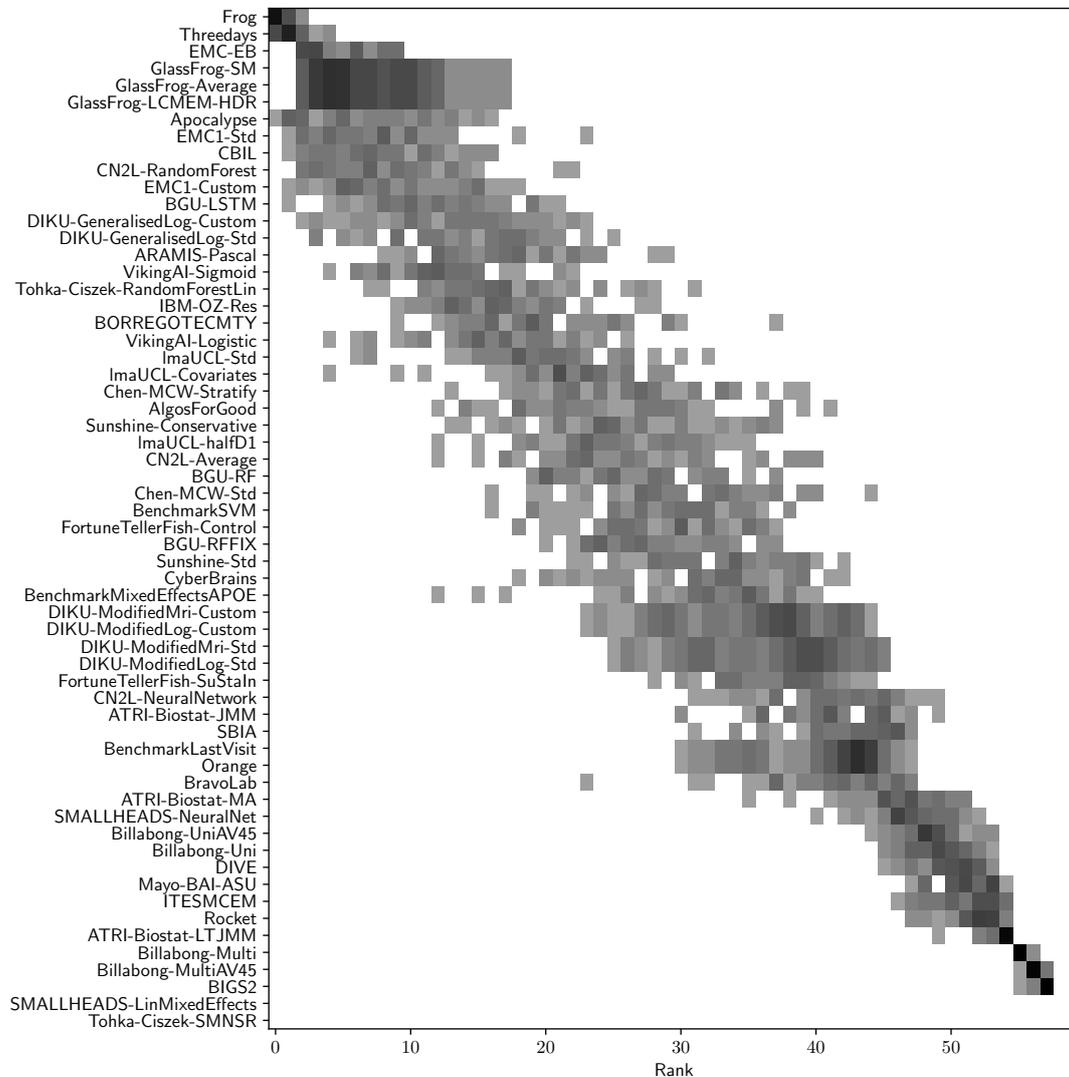


Figure 9: Distribution of ranks in clinical diagnosis MAUC for TADPOLE submissions using the longitudinal prediction set (D2), obtained from $N = 50$ bootstraps of the test set (D4). More precisely, we computed the MAUC ranks given a specific bootstrap of the test set, and then for each TADPOLE submission (Y-axis) we plotted the number of times it achieved a specific rank. Figures 10 – 14 use the same methodology.

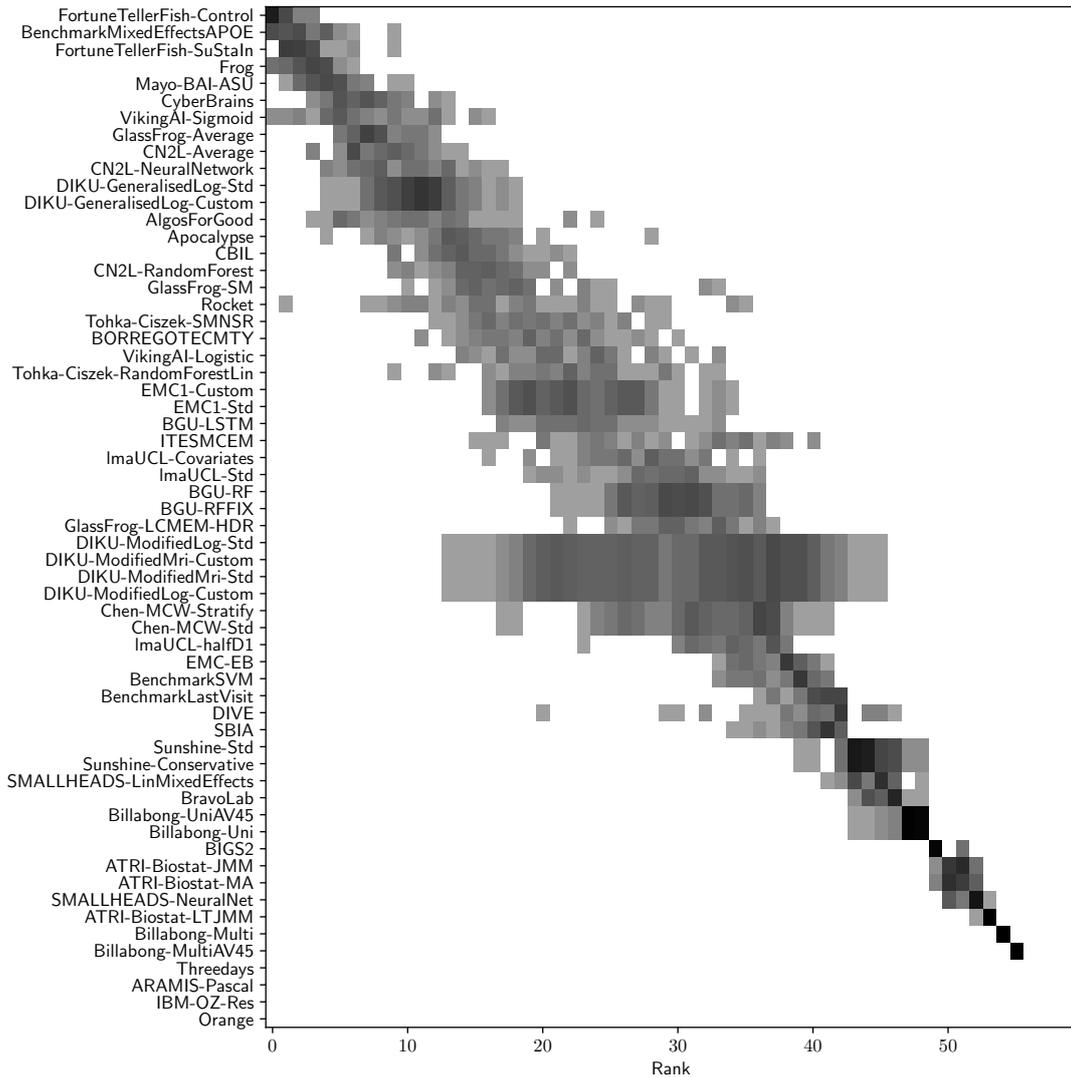


Figure 10: Distribution of ranks in ADAS-Cog13 MAE for TADPOLE submissions using the longitudinal prediction set (D2)

Algorithm	Diagnosis		ADAS-Cog13		Ventricles, % ICV	
	MAUC (D7)	MAUC (D4)	MAE (D6)	MAE (D4)	MAE (D6)	MAE (D4)
ConsensusMedian	0.864	0.925	6.39	5.12	0.35	0.38
ConsensusMean	0.863	0.920	7.45	3.75	0.39	0.48
Frog	0.835	0.931	9.18	4.85	0.92	0.45
BenchmarkLastVisit	0.787	0.774	8.88	7.05	0.70	0.63
BenchmarkSVM	0.773	0.836	9.77	6.82	0.57	0.86
BenchmarkMixedEffects	0.740	0.846	10.09	4.19	0.42	0.56
BenchmarkMixedEffectsAPOE	0.740	0.822	9.61	4.75	0.42	0.57

Table 7: External validation results using consensus models and benchmark methods on D6 (ADAS-Cog13 and Ventricles) and D7 (Diagnosis), together with internal test results on D4 (as in Figure 3.1). The increase in performance of consensus methods over individual methods and benchmarks remains similar. Cognitive test scores remain difficult to predict.

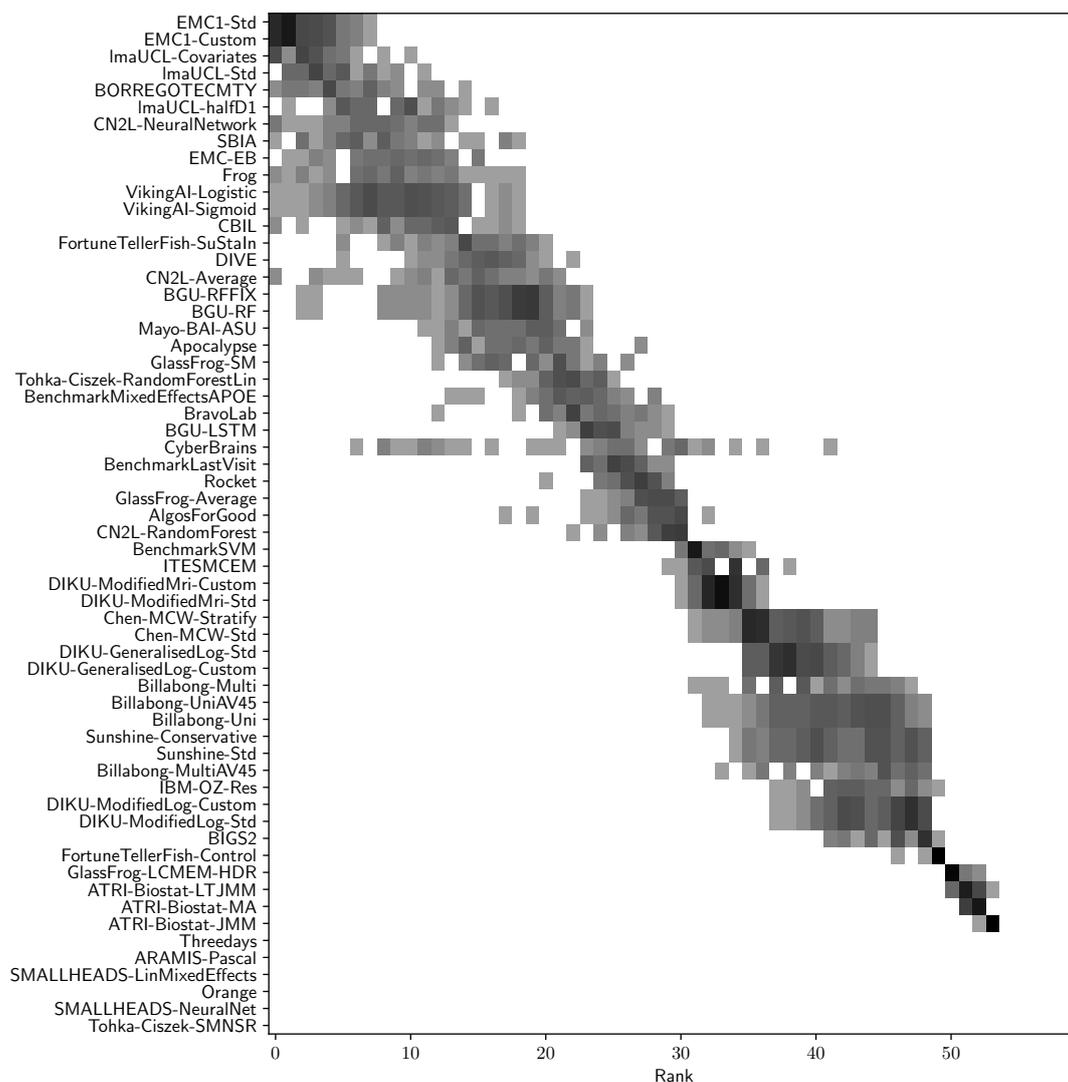


Figure 11: Distribution of ranks in Ventricle Volume MAE for TADPOLE submissions using the longitudinal prediction set (D2).

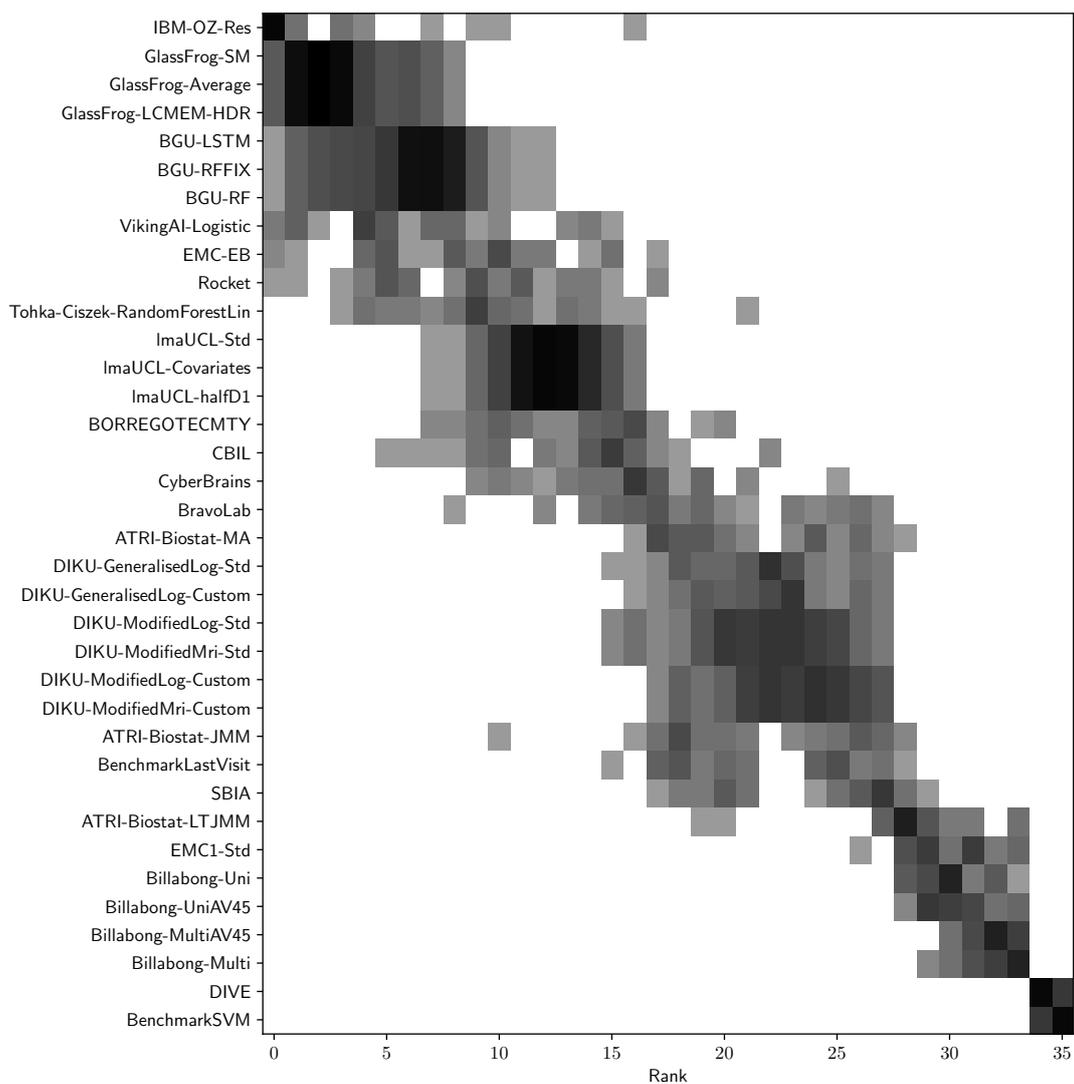


Figure 12: Distribution of ranks in clinical diagnosis MAUC for TADPOLE submissions using the longitudinal prediction set (D3).

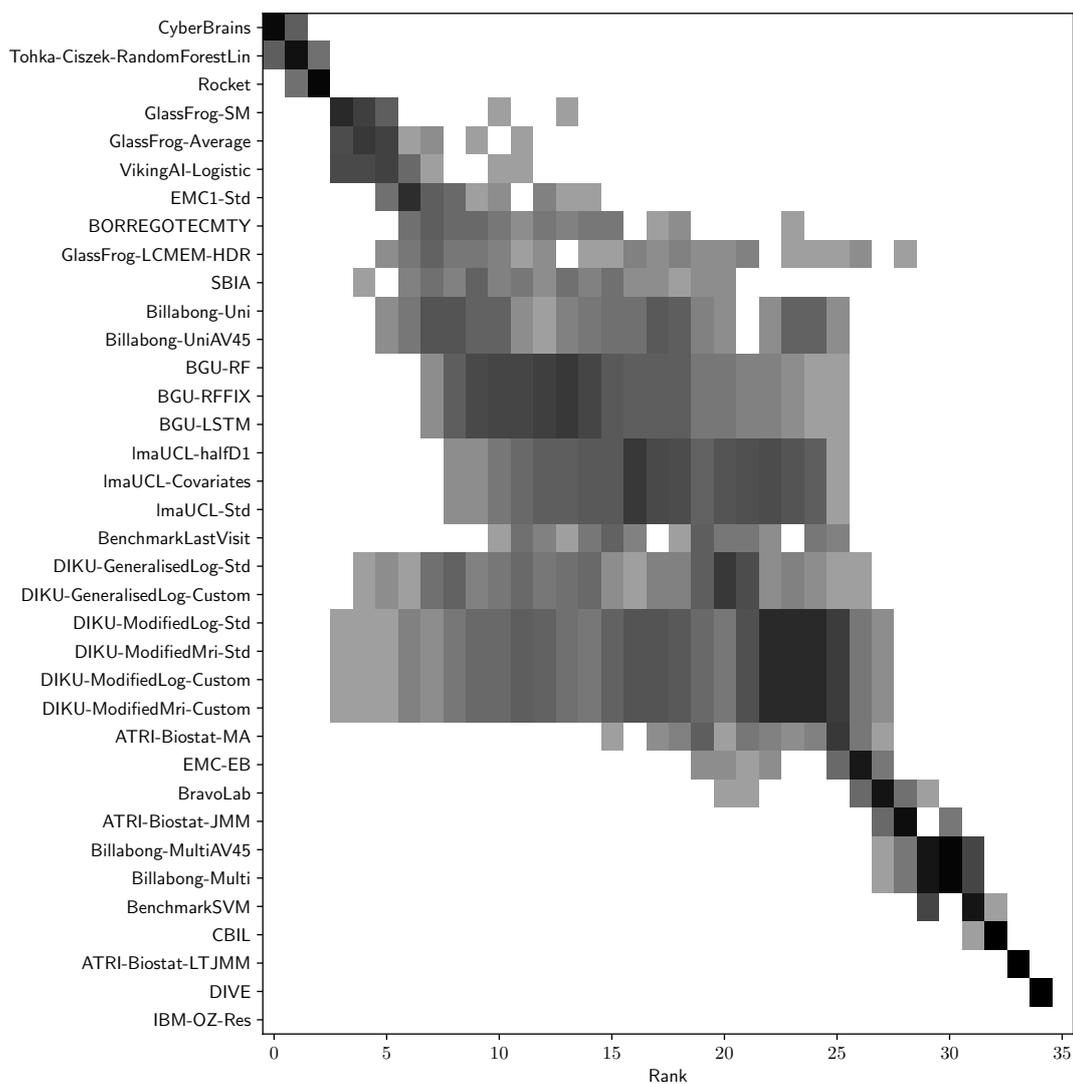


Figure 13: Distribution of ranks in ADAS-Cog13 MAE for TADPOLE submissions using the longitudinal prediction set (D3).

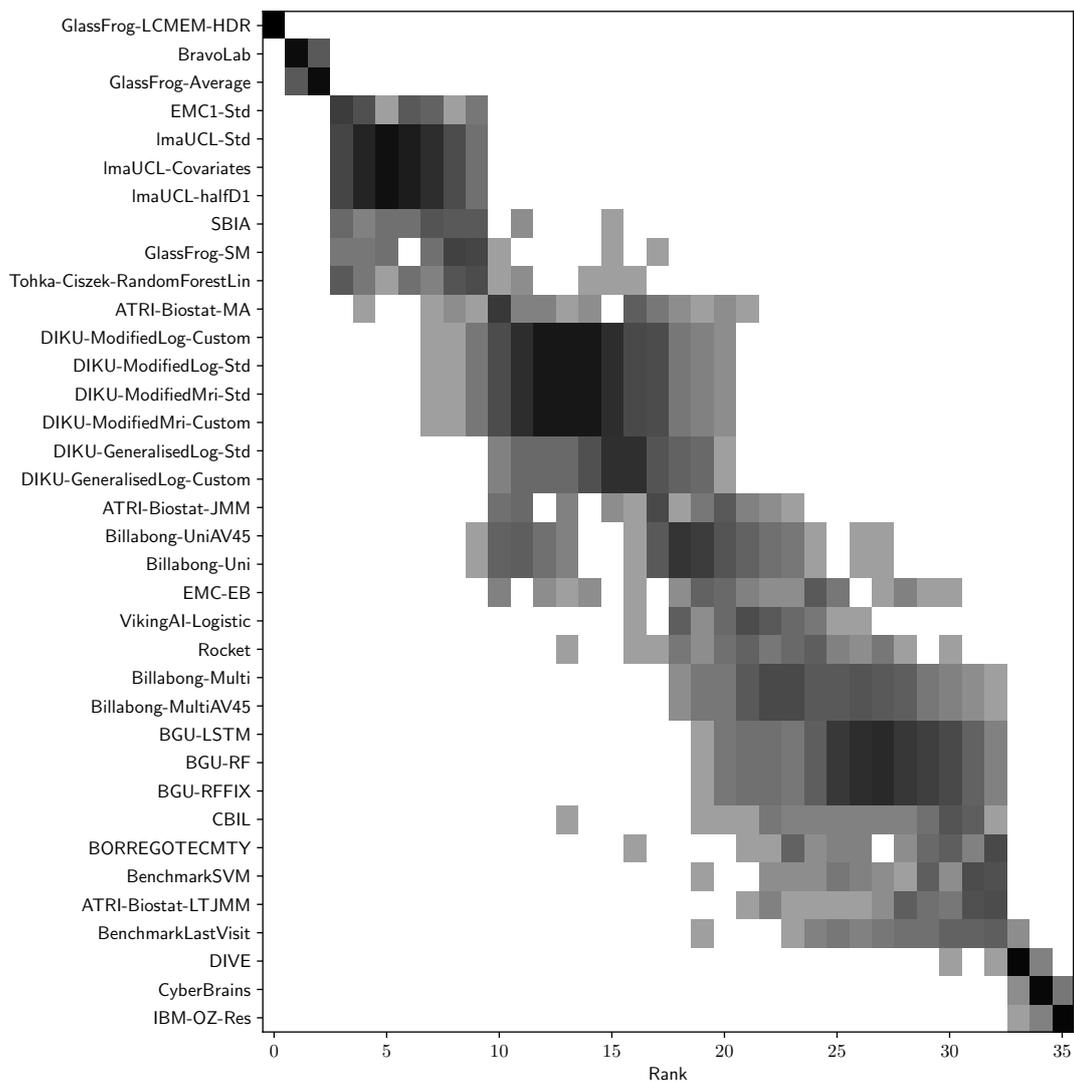


Figure 14: Distribution of ranks in Ventricle Volume MAE for TADPOLE submissions using the longitudinal prediction set (D3).