TADPOLE Challenge: Accurate Alzheimer's disease prediction through crowdsourced forecasting of future data

Răzvan V. Marinescu[1,2], Neil P. Oxtoby[2], Alexandra L. Young[2], Esther E. Bron[3], Arthur W. Toga[4], Michael W. Weiner[5], Frederik Barkhof[3,6], Nick C. Fox[6], Polina Golland[1], Stefan Klein[3], and Daniel C. Alexander[2]

[1] Computer Science and Artificial Intelligence Laboratory, MIT, USA
[2] Centre for Medical Image Computing, University College London, UK
[3] Biomedical Imaging Group Rotterdam, Erasmus MC, Netherlands
[4] Laboratory of Neuro Imaging, University of Southern California, USA
[5] Center for Imaging of Neurodegenerative Diseases, UCSF, USA
[6] Dementia Research Centre, UCL Institute of Neurology, UK
[7] Department of Radiology and Nuclear Medicine, VU Medical Centre, Netherlands
tadpole@cs.ucl.ac.uk

Abstract. The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge compares the performance of algorithms at predicting the future evolution of individuals at risk of Alzheimer's disease. TADPOLE Challenge participants train their models and algorithms on historical data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) study. Participants are then required to make forecasts of three key outcomes for ADNI-3 rollover participants: clinical diagnosis, Alzheimer's Disease Assessment Scale Cognitive Subdomain (ADAS-Cog 13), and total volume of the ventricles – which are then compared with future measurements. Strong points of the challenge are that the test data did not exist at the time of forecasting (it was acquired afterwards), and that it focuses on the challenging problem of cohort selection for clinical trials by identifying fast progressors. The submission phase of TADPOLE was open until 15 November 2017; since then data has been acquired until April 2019 from 219 subjects with 223 clinical visits and 150 Magnetic Resonance Imaging (MRI) scans, which was used for the evaluation of the participants' predictions. Thirty-three teams participated with a total of 92 submissions. No single submission was best at predicting all three outcomes. For diagnosis prediction, the best forecast (team *Frog*), which was based on gradient boosting, obtained a multiclass area under the receiver-operating curve (MAUC) of 0.931, while for ventricle prediction the best forecast (team *EMC1*), which was based on disease progression modelling and spline regression, obtained mean absolute error of 0.41% of total intracranial volume (ICV). For ADAS-Cog 13, no forecast was considerably better than the benchmark mixed effects model (*BenchmarkME*), provided to participants before the submission deadline. Further analysis can help understand which input features and algorithms are most suitable for Alzheimer's disease prediction and for

aiding patient stratification in clinical trials. The submission system remains open via the website: https://tadpole.grand-challenge.org/

Keywords: Alzheimer's Disease, Future prediction, Community Challenge

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. Early detection will be critical in the successful administration of disease modifying treatments during presymptomatic phases of the disease prior to widespread brain damage, i.e. when pathological amyloid and tau accumulate [1]. Moreover, accurate prediction of the evolution of subjects at risk of Alzheimer's disease will help to select homogeneous patient groups for clinical trials, thus reducing variability in outcome measures that can obscure positive effects on subgroups of patients who were at the right stage to benefit.

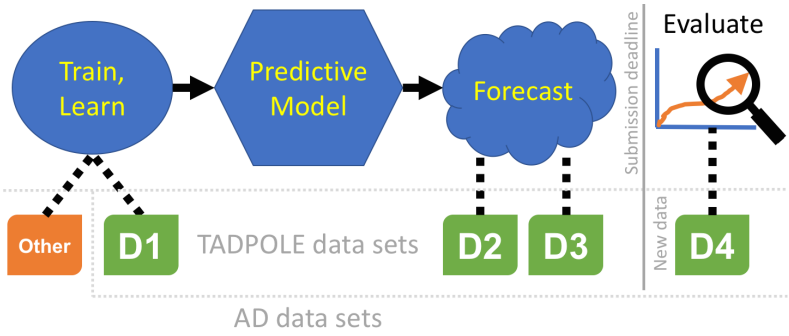

Fig. 1: TADPOLE Challenge design. Participants are required to train a predictive model on a training dataset (D1 and/or others) and make forecasts for different datasets (D2, D3) by the submission deadline. Evaluation will be performed on a test dataset (D4) that is acquired after the submission deadline.

Several approaches for predicting AD-related target variables (e.g. clinical diagnosis, cognitive/imaging biomarkers) have been proposed which leverage multimodal biomarker data available in AD. Traditional longitudinal approaches based on statistical regression model the relationship of the target variables with other known variables, such as clinical diagnosis [2], cognitive test scores [3], or time to conversion between diagnoses [4]. Another approach involves supervised machine learning techniques such as support vector machines, random forests, and artificial neural networks, which use pattern recognition to learn the relationship between the values of a set of predictors (biomarkers) and their labels

(diagnoses). These approaches have been used to discriminate AD patients from cognitively normal individuals [5], and for discriminating at-risk individuals who convert to AD in a certain time frame from those who do not [6]. The emerging approach of disease progression modelling [7,8] 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. Such models show promise for predicting AD biomarker progression at group and individual levels. However, previous evaluations within individual publications are not systematic and reliable 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 we lack a comprehensive comparison of the capabilities of these methods on standardised tasks relevant to real-world applications.

Community challenges have consistently proven 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. For Alzheimer's disease prediction in particular, previous challenges include the CADDementia challenge [9] which aimed to identify clinical diagnosis from MRI scans. A similar challenge, the "International challenge for automated prediction of MCI from MRI data" [10] asked participants to predict diagnosis and conversion status from extracted MRI features of subjects from the ADNI study [11]. Yet another challenge, The Alzheimer's Disease Big Data DREAM Challenge [12], asked participants to predict cognitive decline from genetic and MRI data. However, most of these challenges have not evaluated the ability of algorithms to predict clinical diagnosis and other biomarkers at future timepoints and largely used training data from a limited set of modalities. The one challenge that asked participants to estimate a biomarker at future timepoints (cognitive decline in one of the DREAM sub-challenges) used only genetic and cognitive data for training, and aimed to find genetic loci that could predict cognitive decline. Therefore, standardised evaluation of algorithms needs to be done on biomarker prediction at future timepoints, with the aim of improving clinical trials through enhanced patient stratification.

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge aims to identify the data, features and approaches that are most predictive of future progression of subjects at risk of AD. The challenge focuses on forecasting the evolution of three key AD-related domains: clinical diagnosis, cognitive decline, and neurodegeneration (brain atrophy). 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 (defined as 1-5 years). 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 [13] 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 support prediction of AD progression.

This article presents the design of the TADPOLE Challenge and outlines preliminary results.

2 Competition Design

The aim of TADPOLE is to predict future outcome measurements of subjects at-risk of AD, enrolled in the ADNI study. A history of informative measurements from ADNI (imaging, psychology, demographics, genetics, etc.) from each individual is available to inform forecasts. TADPOLE participants were required to predict future measurements from these individuals and submit their predictions before a given submission deadline. Evaluation of these forecasts occurred post-deadline, after the measurements had been acquired. A diagram of the TADPOLE flow is shown in Fig 1.

TADPOLE challenge participants were required to make month-by-month forecasts of three key biomarkers: (1) clinical diagnosis which is either cognitively normal (CN), mild cognitive impairment (MCI) or probable Alzheimer's disease (AD); (2) Alzheimer's Disease Assessment Scale Cognitive Subdomain (ADAS-Cog 13) score; and (3) ventricle volume (divided by intra-cranial volume). TADPOLE forecasts are required to be probabilistic and some evaluation metrics will account for forecast probabilities provided by participants.

3 ADNI data aggregation and processing

TADPOLE Challenge organisers provided participants with a standard ADNI-derived dataset (available via the Laboratory Of NeuroImaging data archive at `adni.loni.usc.edu`) to train algorithms, removing the need for participants to pre-process the ADNI data or merge different spreadsheets. Software code used to generate the standard datasets is openly available on Github[8]. The challenge data includes: (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: FDG (hypometabolism), AV45 (amyloid), AV1451 (tau) as well as diffusion tensor imaging (DTI); (3) cognitive assessments such as ADAS-Cog 13 acquired in the presence of a clinical expert; (4) genetic information such as alipoprotein E4 (APOE4) status extracted from DNA samples; and (5) general demographic information such as age and gender. Extracted features from this data were merged into a final spreadsheet and made available online.

The imaging data was pre-processed with standard ADNI pipelines. For MRI scans, this included correction for gradient non-linearity, B1 non-uniformity correction and peak sharpening[9]. Meaningful regional features such as volume

[8] https://github.com/noxtoby/TADPOLE

[9] see `http://adni.loni.usc.edu/methods/mri-analysis/mri-pre-processing`

and cortical thickness were extracted using Freesurfer. Each PET image (FDG, AV45, AV1451), which consists of a series of dynamic frames, had its frames co-registered, averaged across the dynamic range, standardised with respect to the orientation and voxel size, and smoothed to produce a uniform resolution of 8mm full-width/half-max (FWHM)[10]. Standardised uptake value ratio (SUVR) measures for relevant regions-of-interest were extracted after registering the PET images to corresponding MR images using SPM5. DTI scans were corrected for head motion and eddy-current distortion, skull-stripped, EPI-corrected, and finally aligned to the T1 scans. Diffusion tensor summary measures were estimated based on the Eve white-matter atlas.

3.1 TADPOLE Datasets

In order to evaluate the effect of different methodological choices, we prepared four "standard" data sets: the D1 standard training set contains longitudinal data from the entire ADNI history; the D2 longitudinal prediction set contains all available data from the ADNI rollover individuals, for whom challenge participants are asked to provide forecasts; the D3 cross-sectional prediction set contains a single (most recent) time point and a limited set of variables from each rollover individual – this represents the information typically available in a clinical trial; the D4 test set contains visits from ADNI rollover subjects after 1 Jan 2018, which contain at least one of the following: diagnostic status, ADAS score, or ventricle volume from MRI – this dataset did not exist at the time of submitting forecasts. Full demographics for D1–D4 are given in Table 1.

4 Submissions and evaluation

The challenge had a total of 33 participating teams, who submitted a total of 58 forecasts from D2, 34 forecasts from D3, and 6 forecasts from custom prediction sets. Table 2 summarises the top-3 winner methods in terms of input features used, handling of missing data and predictive models: *Frog* used a gradient boosting method, which combined many weak predictors to build a strong predictor; *EMC1* derived a "disease state" variable aggregating multiple features together and then used an SVM and 2D splines for prediction, while VikingAI used a latent-time parametric model with subject- and feature-specific parameters – see [14] for full method details. We also describe three benchmark models which were provided to participants at the start of the challenge: (i) *BenchmarkLastVisit* uses the measurement at the last available visit, (ii) *BenchmarkME-APOE* uses a mixed effects model with APOE status as covariate and (iii) BenchmarkSVM uses an out-of-the-box support vector machine (SVM) and regressor for forecast.

For evaluation of clinical status predictions, we used similar metrics to those that proved effective in the CADDementia challenge [9]: (i) the multiclass area under the receiver operating curve (MAUC); and (ii) the overall balanced classification accuracy (BCA). For ADAS and ventricle volume, we used three metrics:

[10] see http://adni.loni.usc.edu/methods/pet-analysis/pre-processing

Measure	**D1**	**D2**	**D3**	**D4**
Subjects	1667	896	896	219
			Cognitively Normal	
Number (% total)	508 (30%)	369 (41%)	299 (33%)	94 (42%)
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%	47%	43%	47%
MMSE	29.1 ± 1.1	29.0 ± 1.2	28.9 ± 1.4	29.1 ± 1.1
Converters (% total CN)	18 (3.5%)	9 (2.4%)	-	-
			Mild Cognitive Impairment	
Number (% total)	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 (% total MCI)	117 (13.9%)	37 (8.1%)	-	9 (10.0%)
			Alzheimer's Disease	
Number (% total)	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 (% total AD)	-	-	-	9 (31.0%)

Table 1: Summary of TADPOLE datasets D1–D4. Each subject has been allocated to either Cognitively Normal, MCI or AD group based on diagnosis at the first available visit within each dataset.

Submission	**Extra† Features**	**Nr. of features**	**Missing data imputation**	**Diagnosis prediction**	**ADAS/Vent. prediction**
Frog	most features	70+420*	none	gradient boosting	gradient boosting
EMC1-Std	MRI, ASL, cognitive	250	nearest neighbour	DPM SVM 2D-spline	DPM 2D-spline
VikingAI-Sigmoid	MRI, cognitive, tau	10	none	DPM + ordered logit	DPM
BenchmarkLastVisit	-	3	none	constant model	constant model
BenchmarkME-APOE	APOE	4	none	Gaussian model	linear mixed effects model
BenchmarkSVM	age, APOE	6	mean of previous values	SVM	support vector regressor

Table 2: Summary of benchmarks and top-3 methods used in the TADPOLE submissions. DPM – disease progression model. (†) Aside from the three target biomarkers (*) Augmented features: e.g. min/max, trends, moments.

(i) mean absolute error (MAE), (ii) weighted error score (WES) and (iii) coverage probability accuracy (CPA). BCA and MAE focus purely on prediction accuracy ignoring confidence, MAUC and WES include confidence, while CPA provides an assessment of the confidence interval only. Complete formulations for these can be found in Table 3, with detailed explanations in the TADPOLE design paper [13]. To compute an overall rank, we first calculated the sum of ranks from MAUC, ADAS MAE and Ventricle MAE for each submission, and the overall ranking was derived from these sums of ranks.

Formula	**Definitions**
$mAUC = \frac{2}{L(L-1)} \sum_{i=2}^{L} \sum_{j=1}^{i} \hat{A}(c_i, c_j)$	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 data points
$BCA = \frac{1}{2L} \sum_{i=1}^{L} \left[\frac{TP}{TP+FN} + \frac{TN}{TN+FP} \right]$	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 data points
$MAE = \frac{1}{N} \sum_{i=1}^{N} \left\lvert \tilde{M}_i - M_i \right\rvert$	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
$WES = \frac{\sum_{i=1}^{N} \tilde{C}_i \lvert \tilde{M}_i - M_i \rvert}{\sum_{i=1}^{N} \tilde{C}_i}$	M_i, $\tilde{M}_i$ and N defined as above. $\tilde{C}_i = (C_+ - C_-)^{-1}$, where $[C_-, C_+]$ is the 50% confidence interval
$CPA = \lvert ACP - 0.5 \rvert$	actual coverage probability (ACP) - the proportion of measurements that fall within the 50% confidence interval.

Table 3: TADPOLE performance metric formulas and definitions for the terms.

5 Results

While full results can be found on the TADPOLE website [14], here we only include the top-3 winners. Table 4 compiles all metrics for top-3 TADPOLE forecasts from the D2 prediction set. The best overall performance was obtained by team *Frog*, with a clinical diagnosis MAUC of 0.931, ADAS MAE of 4.85 and Ventricle MAE of 0.45. Among the benchmark methods, *BenchmarkME-APOE* had the best overall rank of 18, obtaining an MAUC of 0.82, ADAS MAE of 4.75 and Ventricle MAE of 0.57. In terms of diagnosis predictions, *Frog* had an overall MAUC score of 0.931. For ADAS prediction, *BenchmarkME-APOE* had the best MAE of 4.75. For Ventricle prediction, *EMC1-Std* had the best MAE of 0.41 and WES of 0.29. In terms of the most accurate confidence interval estimates, *VikingAI* achieved the best CPA scores of 0.02 for ADAS and 0.2 for Ventricles.

Submission	Overall Rank	Diagnosis		ADAS			Ventricles (% ICV)		
		MAUC	BCA	MAE	WES	CPA	MAE	WES	CPA
Frog	1	0.931	0.849	4.85	4.74	0.44	0.45	0.33	0.47
EMC1-Std	2	0.898	0.811	6.05	5.40	0.45	0.41	0.29	0.43
VikingAI-Sigmoid	3	0.875	0.760	5.20	5.11	0.02	0.45	0.35	0.20
BenchmarkME-APOE	18	0.822	0.749	4.75	4.75	0.36	0.57	0.57	0.40
BenchmarkSVM	34	0.836	0.764	6.82	6.82	0.42	0.86	0.84	0.50
BenchmarkLastVisit	40	0.774	0.792	7.05	7.05	0.45	0.63	0.61	0.47

Table 4: Ranked forecasting scores for benchmark models and top-3 TADPOLE submissions.

6 Discussion

In the current work we have outlined the design and key results of TADPOLE Challenge, which aims to identify algorithms and features that can best predict the evolution of Alzheimer's disease. Despite the small number of converters in the training set, the methods were able to accurately forecast the clinical diagnosis and ventricle volume, although they found it harder to forecast cognitive test scores. Compared to the benchmark models, the best submissions had considerably smaller errors that represented only a small fraction of the errors obtained by benchmark models (0.42 for clinical diagnosis MAUC and 0.71 for ventricle volume MAE). For clinical diagnosis, this suggests that more than half of the subjects originally misdiagnosed by the best benchmark model (*BenchmarkSVM*) are now correctly diagnosed with the new methods. Moreover, the results suggest that we do not have a clear winner on all categories. While team Frog had the best overall submission with the lowest sum of ranks, for each performance metric individually we had different winners.

Additional work currently in progress [14] suggests that consensus methods based on averaging predictions from all participants perform better than any single individual method. This demonstrates the power of TADPOLE in achieving state-of-the-art prediction accuracy through crowd-sourcing prediction models.

The TADPOLE Challenge and its preliminary results presented here are of importance for clinical trials and general clinical settings. The best algorithms identified here could be used for a-priori stratification in clinical trials, for e.g. separating fast progressors from slow progressors, with the hope that a larger drug effect would be observed between fast-progressors and the placebo group. In order to make these models applicable to clinical settings, they need to be further validated on a subject population with post-mortem confirmation, as clinical diagnosis of probable AD only has moderate agreement with gold-standard neuropathological post-mortem diagnosis (70.9% – 87.3% sensitivity and 44.3% – 70.8% specificity, according to [15]). We hope such a validation will be possible in the future, with the advent of neuropathological confirmation in large, longitudinal, multimodal datasets such as ADNI.

In future work, we plan to analyse which features and methods were most useful for predicting AD progression, and assess if the results are sufficient to improve stratification for AD clinical trials. We also plan to evaluate the impact and interest of the first phase of TADPOLE within the community, to guide decisions on whether to organise further submission and evaluation phases.

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 Azheimer's Research UK for sponsoring the challenge by providing the 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. The UCL-affiliated researchers received support from the NIHR UCLH Biomedical Research Centre. 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). FB was supported by the NIHR UCLH Biomedical Research Centre and the AMYPAD project, which has received support from the EU-EFPIA Innovative Medicines Initiatives 2 Joint Undertaking (AMYPAD project, grant 115952). This project has received funding from the EU Horizon 2020 research and innovation programme under grant agreement No 666992.

References

1. Mehta, D., Jackson, R., Paul, G., Shi, J., Sabbagh, M.: Why do trials for Alzheimer's disease drugs keep failing? A discontinued drug perspective for 2010–2015. Expert opinion on investigational drugs **26**(6) (2017) 735
2. Scahill, R.I., Schott, J.M., Stevens, J.M., Rossor, M.N., Fox, N.C.: 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) (2002) 4703–4707

3. Yang, E., Farnum, M., Lobanov, V., Schultz, T., Raghavan, N., Samtani, M.N., Novak, G., Narayan, V., DiBernardo, A.: Quantifying the pathophysiological timeline of Alzheimer's disease. Journal of Alzheimer's Disease **26**(4) (2011) 745–753
4. Guerrero, R., Schmidt-Richberg, A., Ledig, C., Tong, T., Wolz, R., Rueckert, D., ADNI, et al.: Instantiated mixed effects modeling of Alzheimer's disease markers. NeuroImage **142** (2016) 113–125
5. Klöppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr, C.R., Ashburner, J., Frackowiak, R.S.: Automatic classification of MR scans in Alzheimer's disease. Brain **131**(3) (2008) 681–689
6. Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S., ADNI, et al.: Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical **2** (2013) 735–745
7. Young, A.L., Oxtoby, N.P., Daga, P., Cash, D.M., Fox, N.C., Ourselin, S., Schott, J.M., Alexander, D.C.: A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain **137**(9) (2014) 2564–2577
8. Lorenzi, M., Filippone, M., Frisoni, G.B., Alexander, D.C., Ourselin, S., Initiative, A.D.N., et al.: Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in Alzheimer's disease. NeuroImage **190** (2017) 56–68
9. Bron, E.E., Smits, M., Van Der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Orellana, C.M., Meijboom, R., et al.: Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage **111** (2015) 562–579
10. Sarica, A., Cerasa, A., Quattrone, A., Calhoun, V.: Editorial on special issue: Machine learning on MCI. Journal of neuroscience methods **302** (2018) 1
11. Weiner, M.W., Veitch, D.P., Aisen, P.S., Beckett, L.A., Cairns, N.J., Green, R.C., Harvey, D., Jack Jr, C.R., Jagust, W., Morris, J.C., et al.: Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. Alzheimer's & Dementia **13**(4) (2017) e1–e85
12. Allen, G.I., Amoroso, N., Anghel, C., Balagurusamy, V., Bare, C.J., Beaton, D., Bellotti, R., Bennett, D.A., Boehme, K.L., Boutros, P.C., et al.: Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. Alzheimer's & dementia: the journal of the Alzheimer's Association **12**(6) (2016) 645–653
13. Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Klein, S., Alexander, D.C., et al.: Tadpole challenge: Prediction of longitudinal evolution in Alzheimer's disease. arXiv preprint arXiv:1805.03909 (2018)
14. https://tadpole.grand-challenge.org/Results/
15. Beach, T.G., Monsell, S.E., Phillips, L.E., Kukull, W.: 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) (2012) 266–273