TADPOLE Challenge: Prediction of Longitudinal Evolution in Alzheimer's Disease

Răzvan V. Marinescu[a], Neil P. Oxtoby[a], Alexandra L. Young[a], Esther E. Bron[b], Arthur W. Toga[c], Michael W. Weiner[d], Frederik Barkhof[e,f], Nick C. Fox[e], Stefan Klein[b], Daniel C. Alexander[a], the EuroPOND Consortium, for the Alzheimer's Disease Neuroimaging Initiative

[a]*Centre for Medical Image Computing, University College London, Gower Street, London, United Kingdom, WC1E 6BT*
[b]*Biomedical Imaging Group Rotterdam, Erasmus MC, Rotterdam, Netherlands, PO Box 2040 3000 CA*
[c]*Laboratory of Neuro Imaging, Keck School of Medicine, University of Southern California, 2001 N Soto Street, Los Angeles, United States, CA 90032*
[d]*Center for Imaging of Neurodegenerative Diseases, University of California San Francisco, 4150 Clement St. (114M), San Francisco, United States, CA 94121*
[e]*Dementia Research Centre, University College London Institute of Neurology, London, United Kingdom, WC1N 3AX*
[f]*Department of Radiology and Nuclear Medicine, VU University Medical Centre, Amsterdam, Netherlands, 1081 HV*

Abstract

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge compares the performance of algorithms at predicting 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 or any other datasets to which they have access. Participants are then required to make monthly forecasts over a period of 5 years from January 2018, of three key outcomes for ADNI-3 rollover participants: clinical diagnosis, Alzheimer's Disease Assessment Scale Cognitive Subdomain (ADAS-Cog13), and total volume of the ventricles. These individual forecasts are later compared with the corresponding future measurements in ADNI-3 (obtained after the TADPOLE submission deadline). The first submission phase of TADPOLE was open for prize-eligible submissions between 15 June and 15 November 2017. The submission system remains open via the website: https://tadpole.grand-challenge.org, although since 15 November 2017 submissions are not eligible for the first round of prizes. This paper describes the design of the TADPOLE Challenge.

Keywords: Alzheimer's disease, Disease prediction, Community Challenge, Biomarkers

1. Introduction

Alzheimer's disease (AD), and dementia in general, is a key challenge for 21st-century healthcare. The statistics are sobering (Winblad et al., 2016): in 2015, 47 million people worldwide suffer from dementia, of which AD is the most common cause; dementia costs $818 billion worldwide, which is more than 1% of the aggregaste global gross domestic product (GDP); AD might contribute to as many deaths as does heart disease or cancer. There are no available treatments that can cure or even slow the progression of AD – all clinical trials into putative treatments have failed to prove a disease-modifying effect. One key reason for these failures is the difficulty in identifying a group of patients at early stages of the disease, where treatments are most likely to be effective.

While early and accurate diagnosis of dementia can be challenging, this can be aided by quantitative biomarker measurements taken from magnetic resonance imaging (MRI), positron emission tomography (PET), and cerebro-spinal fluid (CSF) samples extracted from lumbar puncture. It has been hypothesized for AD (Jack Jr et al., 2010, 2013; Aisen et al., 2010; Frisoni et al., 2010) that all these biomarkers become abnormal at different intervals before symptom onset, suggesting that together they can be used for accurate prediction of onset and overall disease progression in individuals. In particular, some of the early biomarkers become abnormal decades before symptom onset, and can thus facilitate early diagnosis.

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. Examples include regression of the target variables against clinical diagnosis (Scahill et al., 2002), cognitive test scores (Yang et al., 2011; Sabuncu et al., 2011), rate of cognitive decline (Doody et al., 2010), and retrospectively staging subjects by time to conversion between diagnoses (Guerrero et al., 2016). 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 (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

Email address: tadpole@cs.ucl.ac.uk
URL: http://tadpole.grand-challenge.org

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; Villemagne et al., 2013) biomarker trajectories; richer but less comprehensive models that leverage structure in data such as MR images (Durrleman et al., 2013; Lorenzi et al., 2015; Bilgel et al., 2016); and models of disease mechanisms (Seeley et al., 2009; Zhou et al., 2012; Raj et al., 2012; Iturria-Medina et al., 2016).

These models have shown promise for predicting AD biomarker progression when using existing test data, but few have been tested on truly unseen *future* data. Moreover, different investigators test these models on different datasets (including subsets of a single dataset) and use different processing pipelines. Community challenges have proved effective, in the medical image analysis field and beyond, for providing unbiased comparative evaluations of algorithms and tools designed for a particular task. Previous challenges that focussed on prediction of AD progression include the *CADDementia challenge* (Bron et al., 2015), which aimed to predict clinical diagnosis from MRI scans. A similar challenge, the "*International challenge for automated prediction of MCI from MRI data*" (Sarica et al., 2018) asked participants to predict diagnosis and conversion status from extracted MRI features of subjects from the 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.

The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) Challenge aims to identify the data, features and approaches that are the most predictive of AD progression. In contrast to previous challenges, our motivation is to improve future 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). Identifying such subjects reliably helps cohort selection by focussing on groups that highlight positive treatment effects. The challenge thus focuses on forecasting three key features: clinical status, cognitive decline, and neurodegeneration (brain atrophy), over a five-year timescale. It uses rollover subjects from the ADNI study for whom a history of measurements is available, and who are expected to continue in the study, providing future measurements for testing. Since the test data does not exist at the time of forecast submissions, the challenge provides a completely unbiased basis for performance comparison. TADPOLE goes beyond previous challenges by drawing on a vast set of multimodal measurements from ADNI which support prediction of AD progression.

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 are required to predict future measurements from these individuals and submit their predictions before a given submission deadline. Evaluation of these forecasts occurs post-deadline, after the measurements have been acquired. A diagram of the TADPOLE flow is shown in Fig 1.

3. Forecasts

Since we do not know the exact time of future data acquisitions for any given individual, TADPOLE challenge participants are required to make, for every individual, month-by-month forecasts of three key biomarkers: (1) clinical diagnosis which can be either cognitively normal (CN), mild cognitive impairment (MCI) or probable Alzheimer's disease (AD); (2) ADAS-Cog13 (ADAS13) score; and (3) ventricle volume (divided by intra-cranial volume). Evaluation is performed using forecasts at the months that correspond to data acquisition. TADPOLE forecasts are required to be probabilistic and some evaluation metrics will account for forecast probabilities provided by participants. Methods or algorithms that do not produce probabilistic estimates can still be used, by setting binary probabilities (zero or one) and default confidence intervals.

Participants are required to submit forecasts in a standardised format (see Table 1). For clinical status, relative likelihoods of each option (CN, MCI, and AD) for each individual should be provided. These are normalised at evaluation time; negative likelihoods are set to zero. For ADAS13 and ventricle volume, participants need to provide a best-guess value as well as a 50% confidence interval for each individual. This 50% confidence interval (as opposed to the more standard 95%) was chosen to provide a more symmetric and less noisy evaluation of over- and under-estimation of the confidence interval, because similar sample sizes of data fall inside and outside the interval.

4. Data

We provide participants with a standard ADNI-derived dataset (available via the Laboratory Of NeuroImaging: LONI) which they can use to train their algorithms, removing the need to pre-process the ADNI data themselves or merge different spreadsheets. However, participants are allowed to use a custom training set, by adding any other ADNI data or data from other studies. The software code used to generate the standard dataset is openly available in a Github repository[1] and on the ADNI website, packaged with the standard dataset in the LONI ADNI database.

4.1. ADNI data

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). The ADNI was launched in

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

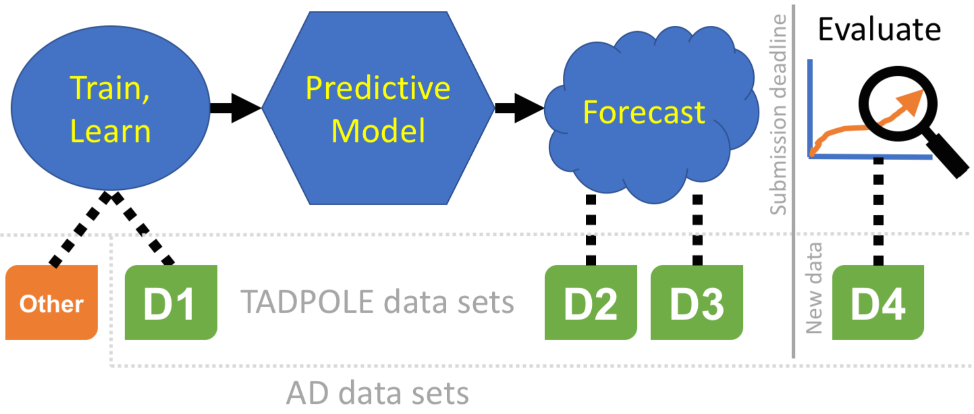

Figure 1: TADPOLE Challenge design. Participants are required to train a predictive model on a training dataset (D1 and/or others) and make forecasts for different datasets (D2, D3) by the submission deadline. Evaluation will be performed on a test dataset (D4) that is acquired after the submission deadline.

RID	Forecast Month	Forecast Date	CN relative probability	MCI relative probability	AD relative probability	ADAS	ADAS 50% CI lower	ADAS 50% CI upper	Ventricles	Ventricles 50% CI lower	Ventricles 50% CI upper
A	1	2018-01	0	1	0	30	25	35	0.024	0.021	0.029
B	1	2018-01	3	2	0	25	21	26	0.023	0.021	0.025
C	1	2018-01	0.24	0.38	0.38	40	25	50	0.025	0.023	0.028

Table 1: The format of the forecasts for three example subjects.

2003 by the National Institute on Aging (NIA), the National Institute of Biomedical Imaging and Bioengineering (NIBIB), the Food and Drug Administration (FDA), private pharmaceutical companies and non-profit organizations, as a $60 million, 5-year public-private partnership. The initial goal of ADNI was to recruit 800 subjects but ADNI has been followed by ADNI-GO and ADNI-2. To date these three protocols have recruited over 1500 adults, ages 55 to 90, to participate in the research, consisting of cognitively normal older individuals, people with early or late MCI, and people with early AD. The general ADNI inclusion criteria has been described in Petersen et al. (2010).

The data we 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 alipoprotein 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.

4.2. *Image pre-processing*

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[2]. 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), 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)[3]. Standardised uptake value ratio (SUVR) measures for relevant regions-of-interest were extracted (see Jagust et al. (2010)) after registering the PET images to corresponding MR images using the SPM5 software (Ashburner, 2009). 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 Nir et al. (2013). Diffusion tensor summary measures were estimated based on the Eve white-matter atlas by Oishi et al. (2009).

5. TADPOLE Datasets

The TADPOLE Challenge involves three kinds of data sets: (a) a *training data set*, which is a collection of measurements with associated outcomes that can be used to fit models or train algorithms; (b) a *prediction data set*, which contains only baseline measurements (possibly longitudinal), without associated

[2] see MRI analysis on ADNI website: `http://adni.loni.usc.edu/methods/mri-analysis/mri-pre-processing`

[3] see PET analysis on ADNI website: `http://adni.loni.usc.edu/methods/pet-analysis/pre-processing`

outcomes — this is the data that algorithms, models, or experts use as input to make their forecasts of later patient status and outcome; and (c) *a test data set*, which contains the patient outcomes against which we will evaluate forecasts — in TADPOLE, this data did not exist at the time of submitting forecasts.

In order to evaluate the effect of different methodological choices, we prepared three standard data sets for training and prediction:

- **D1**: The TADPOLE standard training set draws on longitudinal data from the entire ADNI history. The data set contains a set of 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 forecasts. D2 includes all available timepoints for these individuals.
- **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 forecasts from this data set than D2, D3 represents the information typically available when selecting a cohort for a clinical trial.

The forecasts will be evaluated on future data (D4 – test set) from ADNI3 rollovers, acquired after the challenge submission deadline. In addition to the three standard datasets (D1, D2 and D3), challenge participants are allowed to use any other data sets that might serve as useful additional training data.

Fig. 2 shows a diagram highlighting the nested structure of datasets D1–D3. Table 2 shows the proportion of biomarker data available in each dataset. There are a considerable number of entries with missing data, especially for some biomarkers such as tau imaging (AV1451). We also estimated the expected number of subjects and available data for D4, using information from the ADNI3 procedures and using rollovers from previous ADNI studies (Table 2, right-most column) – See Appendix A for more information on D4 estimates. Based on our estimates, we believe the size of D4 (around 330 subjects, 1 visit/subject) should be enough for a reliable evaluation of TADPOLE submissions.

Subject statistics		D1	D2	D3	D4
Nr. of subjects		1667	896	896	*330*
Visits per subject		7.6±3.8	8.5±4.2	1.0±0.0	*1.0±0.0*
Diagnosis* (%)	CN	31	38	45	*39*
	MCI	56	57	39	*49*
	AD	13	5	16	*12*
Data availability**					
Cognitive tests (%)		70	68	84	*62*
MRI (%)		62	56	75	*69*
FDG-PET (%)		16	20	0	*20*
AV45-PET (%)		16	22	0	*19*
AV1451-PET (%)		0.7	1.1	0	*19*
DTI (%)		6	8	0	*15*
CSF (%)		18	19	0	*14*

Table 2: Biomarker summary of TADPOLE datasets D1, D2 and D3. There is considerable amount of missing data in some biomarkers such as AV1451. Numbers for D4 are estimated based on ADNI3 procedures (see ADNI3 procedures manual) and rollovers from previous ADNI studies. (*) Diagnosis at baseline visit. (**) Percentage of all visits (across all subjects) that have measurements for desired biomarker.

6. Submissions

There are two kinds of submissions that challenge participants can make. A simple entry requires a minimal forecast and a description of methods; it makes participants eligible for the prizes but not co-authorship on the scientific paper documenting the results. A simple entry can use any training data or prediction sets and forecast at least one of the target outcome variables (clinical status, ADAS13 score, or ventricle volume). A full entry entitles participants for consideration as a co-author on the publication documenting the results. Such a full entry requires a complete forecast for all three outcome variables on all subjects from the D2 prediction set, along with a description of the methods. Each individual participant is limited to a maximum of three submissions. This restriction has been introduced to avoid the risk of participants tuning their method on the test set by submitting multiple predictions for a range of algorithm settings. Although not required for a full entry, participants are strongly encouraged to submit predictions also for D3.

Prizes are awarded to the best entries regardless of the choice of training sets (D1/custom) and prediction sets (D2/D3). However, the additional submissions support the key scientific aims of the challenge by allowing us to separate the influence of the choice of training data, post-processing pipelines, and modelling techniques or prediction algorithms. The target variables used for evaluation, in particular ventricle volume, will use the same post-processing pipeline as the standard data sets D1-D3.

Beyond the standard training dataset (D1), participants can include additional forecasts from ”custom” (i.e. constructed by the participant) training data or custom post-processing of the raw data from subjects in the standard training set. The same applies to the prediction sets D2 and D3, which can be customised by the participants if desired, e.g. a prediction set with different features from the same individuals as in D2 and D3. Table 3 shows the twelve possible combinations of subject sets, processing and prediction sets, from which a full-entry submission must contain at least one of the first four (ID 1–4).

7. Forecast Evaluation

7.1. Clinical Status Prediction

For evaluation of clinical status predictions, we will use similar metrics to those that proved effective in the CADDementia

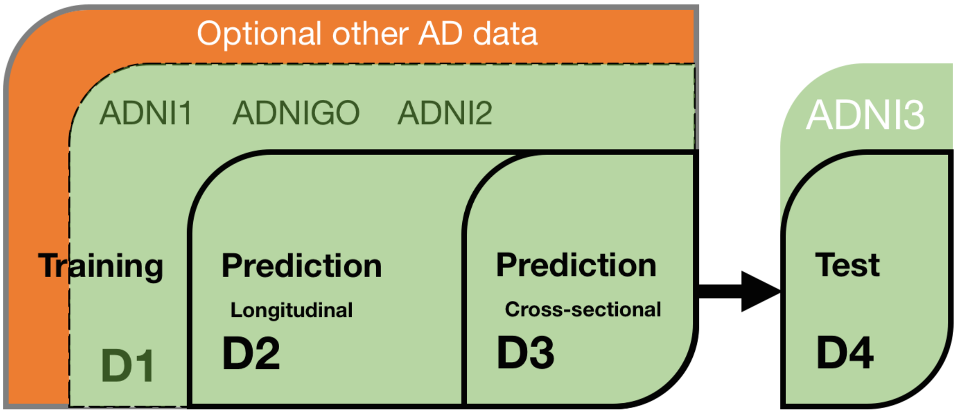

Figure 2: Venn diagram of the ADNI datasets for training (D1), longitudinal prediction (D2), cross-sectional prediction (D3) and the test set (D4). D3 is a subset of D2, which in turn is a subset of D1. Other non-ADNI data can also be used for training.

ID	Training set		Prediction set
	Subject set	Post-processing	
1	D1	standard	D2
2	D1	custom	D2
3	custom	standard	D2
4	custom	custom	D2
5	D1	standard	D3
6	D1	custom	D3
7	custom	standard	D3
8	custom	custom	D3
9	D1	standard	custom
10	D1	custom	custom
11	custom	standard	custom
12	custom	custom	custom

Table 3: Types of submissions that can be made by participants, for different types of training sets, prediction sets and post-processing pipelines.

challenge (Bron et al., 2015): (i) the multiclass area under the receiver operating curve (mAUC); and (ii) the overall balanced classification accuracy (BCA). The mAUC is independent of the group sizes and gives an overall measure of classification ability that accounts for relative likelihoods assigned to each class. The simpler BCA is also independent of group sizes, but does not exploit the probabilistic nature of the forecasts.

7.1.1. *Multiclass Area Under the Receiver Operating Characteristic (ROC) Curve*

The multiclass Area Under the ROC Curve (mAUC) is a simple generalisation of the area under the ROC curve applicable to problems with more than two classes (Hand and Till, 2001). The AUC $\hat{A}(c_i|c_j)$ for classification of a class c_i against another class c_j, is:

$$\hat{A}(c_i|c_j) = \frac{S_i - n_i(n_i+1)/2}{n_i n_j} \tag{1}$$

where n_i and n_j are the number of points belonging to classes i and j, respectively; while S_i is 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. We further define the average AUC for classes i and j as $\hat{A}(c_i, c_j) = 0.5(\hat{A}(c_i|c_j) + \hat{A}(c_j|c_i))$. The overall mAUC is then obtained by averaging $\hat{A}(c_i, c_j)$ over all pairs of classes:

$$mAUC = \frac{2}{L(L-1)} \sum_{i=2}^{L} \sum_{j=1}^{i} \hat{A}(c_i, c_j) \tag{2}$$

where L is the number of classes. The class probabilities that go into the calculation of S_i in the first equation are p_{CN}, p_{MCI} and p_{AD}, which are derived from the likelihoods of each ADNI subject being assigned to each diagnostic class, by normalising to have unity sum.

7.1.2. *Balanced Classification Accuracy*

The Balanced Classification Accuracy (see Brodersen et al. (2010)) is an extension of the classification accuracy measure that accounts for the imbalance in the numbers of datapoints belonging to each class. However, the measure is not probabilistic, so in TADPOLE the data points need to be assigned a hard classification to the class (CN, MCI, or AD) with the highest likelihood. The balanced accuracy for class i is then:

$$BCA_i = \frac{1}{2}\left[\frac{TP}{TP+FN} + \frac{TN}{TN+FP}\right] \tag{3}$$

where TP, FP, TN, FN represent the number of true positives, false positives, true negatives and false negatives for classification as class i. In this case, true positives are data points with true label i and correctly classified as such, while the false negatives are the data points with true label i and incorrectly classified to a different class $j \neq i$. True negatives and false positives are defined similarly. The overall BCA is given by the mean of all the balanced accuracies for every class.

7.2. *Continuous Feature Predictions*

For ADAS13 and ventricle volume, we will use three metrics: mean absolute error (MAE), weighted error score (WES) and coverage probability accuracy (CPA). The MAE focuses purely on accuracy of the best-guess prediction ignoring the confidence interval, whereas the WES incorporates confidence estimates into the error score. The CPA provides an assessment of the accuracy of the confidence estimates, irrespective of the best-guess prediction accuracy.

7.2.1. *Mean Absolute Error*

The mean absolute error (MAE) is:

$$MAE = \frac{1}{N} \sum_{i=1}^{N} \left| \tilde{M}_i - M_i \right| \tag{4}$$

where N is the number of data points (forecasts) evaluated, M_i is the actual biomarker value in individual i in future data, and $\tilde{M}_i$ is the participant's best prediction for M_i.

7.2.2. *Weighted Error Score*

The weighted error score is defined as:

$$WES = \frac{\sum_{i=1}^{N} \tilde{C}_i \left| \tilde{M}_i - M_i \right|}{\sum_{i=1}^{N} \tilde{C}_i} \tag{5}$$

where the weightings $\tilde{C}_i$ are the participant's relative confidences in their $\tilde{M}_i$. We estimate $\tilde{C}_i$ as the inverse of the width of the 50% confidence interval of their biomarker estimate:

$$\tilde{C}_i = (C_+ - C_-)^{-1} \tag{6}$$

where $[C-, C+]$ is the confidence interval provided by the participant.

7.2.3. *Coverage Probability Accuracy*

The coverage probability accuracy is:

$$CPA = |ACP - NCP| \tag{7}$$

where NCP is the nominal coverage probability, the target for the confidence intervals, and ACP is the actual coverage probability, defined as the proportion of measurements that fall within the corresponding confidence interval. In TADPOLE, we set NCP to be 0.5, which means that ideally only 50% of the measurements would fall inside the confidence interval. The CPA can take values between 0 and 1, and lower scores are better.

8. Prizes

We are extremely grateful to Azheimer's Research UK, The Alzheimer's Society, and The Alzheimer's Association for sponsoring a prize fund of £30,000. At the time of first submission, we proposed six separate prizes, as outlined in Table 4, but reserve the right to reallocate the prize money depending on the numbers of participants eligible for each prize. The first four are general categories (open to all challenge participants) and constitute one prize for the best forecast of each feature as well as one for overall best performance. The last two prizes are for two different student categories.

Prize amount	Outcome measure	Performance Metric	Eligibility
£5,000	Clinical status	mAUC	all
£5,000	ADAS13	MAE	all
£5,000	Ventricle volume	MAE	all
£5,000	Overall best	Lowest sum of ranks*	all
£5,000	Clinical status	mAUC	University teams
£5,000	Clinical status	mAUC	High-school teams

Table 4: Prize allocation scheme using funds from Azheimer's Research UK, The Alzheimer's Society and The Alzheimer's Association. There are 6 prizes awarded to different outcome measures, the last two of which are eligible only for university and high-school teams. (*) The overall best team will be the team that obtains the lowest sum of ranks in the clinical status, ADAS13 and ventricle volume categories.

9. Discussion

We have outlined the design of the TADPOLE Challenge, which aims to identify algorithms and features that can best predict the evolution of Alzheimer's disease. Challenge participants use historical data from ADNI in order to predict three key outcomes: clinical diagnosis, ADAS-Cog13 and ventricle volume. Determining which features and algorithms best predict AD evolution can aid refinement of cohorts and endpoint assessment for clinical trials, and can provide accurate prognostic information in clinical settings.

The TADPOLE Challenge was designed to be transparent and accessible. To this end, all of our scripts are available in an open repository[4]. We also created a public forum[5] where we answer participant questions. Finally, in order to enable participants to share algorithm performance results throughout the competition, we created a leaderboard system[6] that evaluates submissions on an existing test dataset and publishes the results live on our website.

Going forward, we hope that by November 2018 sufficient data will be available from ADNI3 rollovers for a first meaningful evaluation of the forecasts. We plan to publish the results on the website in January 2019, and then submit a publication of the results soon after. However, we reserve the right to delay evaluation until sufficient data is available. At that time, we will also 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.

10. Acknowledgements

TADPOLE Challenge has been organised by the European Progression Of Neurological Disease (EuroPOND) consortium,

[4] TADPOLE repository: https://github.com/noxtoby/TADPOLE

[5] TADPOLE forum: https://groups.google.com/forum/#!forum/tadpolechallenge

[6] Leaderboard: https://tadpole.grand-challenge.org/leaderboard/

in collaboration with the ADNI. We thank all the participants and advisors, in particular Clifford R. Jack Jr., Mayo Clinic, Rochester, United States and Bruno M. Jedynak, Portland State University, Portland, United States for useful input and feedback.

The organisers are extremely grateful to Azheimer's Research UK, The Alzheimer's Society, and The Alzheimer's Association 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 is supported by the EPSRC Centre For Doctoral Training in Medical Imaging with grant EP/L016478/1. NPO, FB, SK, and DCA are supported by EuroPOND, which is an EU Horizon 2020 project. ALY is currently supported by an EPSRC Doctoral Prize fellowship and was previously supported by EPSRC grant EP/J020990/01. DCA is supported by EPSRC grants J020990, M006093 and M020533. 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 is 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.

Appendix A. Expected number of subjects and available data for D4

We estimated the number of subjects and available data in D4 (Table 2, last column) using information from the ADNI procedures manual and previous ADNI rollovers. For estimating the total number of subjects (first row) expected in D4, we computed the dropout rate (0.36) based on ADNI1 rollovers to ADNI2, then multiplied it by the total number of subjects in D2 (896). For estimating the proportions of each diagnostic category (third row), we used the proportion of diagnostic rates in D2 and multiplied them with conversion rates within 1 year from ADNI1/GO/2 (see website FAQ). For estimating the average number of visits per subject (mean ± std.) in D4 (second row), we used the proportions for each diagnostic group and considered one visit per subject (ADNI procedures). We set the standard deviation to be zero, although in practice this won't be the case.

For estimating the available biomarker data (lower half of table), we used a 1-year time-frame from start of ADNI2 (July 2012 – July 2013) and computed the proportion of available data in that time frame. For AV1451, we used the same estimate as for AV45, due to the fact that the scan was introduced later on in ADNI2, and we expect more subjects to undergo AV1451 scans in ADNI3. A Python script that computes all the data from Table 2 is given in the TADPOLE repository: https://github.com/noxtoby/TADPOLE/blob/master/statistics/tadpoleStats.py.

References

References

Aisen, P. S., Petersen, R. C., Donohue, M. C., Gamst, A., Raman, R., Thomas, R. G., Walter, S., Trojanowski, J. Q., Shaw, L. M., Beckett, L. A., et al., 2010. Clinical core of the Alzheimer's Disease Neuroimaging Initiative: progress and plans. Alzheimer's & dementia: the journal of the Alzheimer's Association 6 (3), 239–246.

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., 2016. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. Alzheimer's & dementia: the journal of the Alzheimer's Association 12 (6), 645–653.

Ashburner, J., 2009. Computational anatomy with the SPM software. Magnetic resonance imaging 27 (8), 1163–1174.

Bilgel, M., Prince, J. L., Wong, D. F., Resnick, S. M., Jedynak, B. M., 2016. A multivariate nonlinear mixed effects model for longitudinal image analysis: Application to amyloid imaging. Neuroimage 134, 658–670.

Brodersen, K. H., Ong, C. S., Stephan, K. E., Buhmann, J. M., 2010. The balanced accuracy and its posterior distribution. In: Pattern recognition (ICPR), 2010 20th international conference on. IEEE, pp. 3121–3124.

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., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage 111, 562–579.

Donohue, M. C., Jacqmin-Gadda, H., Le Goff, M., Thomas, R. G., Raman, R., Gamst, A. C., Beckett, L. A., Jack, C. R., Weiner, M. W., Dartigues, J.-F., et al., 2014. Estimating long-term multivariate progression from short-term data. Alzheimer's & dementia: the journal of the Alzheimer's Association 10 (5), S400–S410.

Doody, R. S., Pavlik, V., Massman, P., Rountree, S., Darby, E., Chan, W., 2010. Predicting progression of alzheimer's disease. Alzheimer's research & therapy 2 (1), 2.

Durrleman, S., Pennec, X., Trouvé, A., Braga, J., Gerig, G., Ayache, N., 2013. Toward a comprehensive framework for the spatiotemporal statistical analysis of longitudinal shape data. International journal of computer vision 103 (1), 22–59.

Fonteijn, H. M., Modat, M., Clarkson, M. J., Barnes, J., Lehmann, M., Hobbs, N. Z., Scahill, R. I., Tabrizi, S. J., Ourselin, S., Fox, N. C., Alexander, D. C., 2012. An event-based model for disease progression and its application in familial Alzheimer's disease and Huntington's disease. NeuroImage 60 (3), 1880–1889.

Frisoni, G. B., Fox, N. C., Jack Jr, C. R., Scheltens, P., Thompson, P. M., 2010. The clinical use of structural MRI in Alzheimer disease. Nature Reviews Neurology 6 (2), 67.

Guerrero, R., Schmidt-Richberg, A., Ledig, C., Tong, T., Wolz, R., Rueckert, D., ADNI, et al., 2016. Instantiated mixed effects modeling of Alzheimer's disease markers. NeuroImage 142, 113–125.

Hand, D. J., Till, R. J., 2001. A simple generalisation of the area under the ROC curve for multiple class classification problems. Machine learning 45 (2), 171–186.

Iturria-Medina, Y., Sotero, R. C., Toussaint, P. J., Mateos-Pérez, J. M., Evans, A. C., Weiner, M. W., Aisen, P., Petersen, R., Jack, C. R., Jagust, W., et al., 2016. Early role of vascular dysregulation on late-onset Alzheimers disease based on multifactorial data-driven analysis. Nature communications 7, 11934.

Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Petersen, R. C., Weiner, M. W., Aisen, P. S., Shaw, L. M., Vemuri, P., Wiste, H. J., Weigand, S. D., et al., 2013. Update on hypothetical model of Alzheimers disease biomarkers. Lancet neurology 12 (2), 207.

Jack Jr, C. R., Knopman, D. S., Jagust, W. J., Shaw, L. M., Aisen, P. S., Weiner, M. W., Petersen, R. C., Trojanowski, J. Q., 2010. Hypothetical model of dynamic biomarkers of the Alzheimer's pathological cascade. The Lancet Neurology 9 (1), 119–128.

Jagust, W. J., Bandy, D., Chen, K., Foster, N. L., Landau, S. M., Mathis, C. A., Price, J. C., Reiman, E. M., Skovronsky, D., Koeppe, R. A., 2010. The Alzheimer's Disease Neuroimaging Initiative positron emission tomography core. Alzheimer's & dementia: the journal of the Alzheimer's Association 6 (3), 221–229.

Jedynak, B. M., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B. T., Raunig, D., Jedynak, C. P., Caffo, B., Prince, J. L., et al., 2012. A computational neurodegenerative disease progression score: method and results with the Alzheimer's Disease Neuroimaging Initiative cohort. Neuroimage 63 (3), 1478–1486.

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., 2008. Automatic classification of MR scans in Alzheimer's disease. Brain 131 (3), 681–689.

Lorenzi, M., Pennec, X., Frisoni, G. B., Ayache, N., 2015. Disentangling normal aging from Alzheimer's disease in structural magnetic resonance images. Neurobiology of aging 36, S42–S52.

Mattila, J., Koikkalainen, J., Virkki, A., Simonsen, A., van Gils, M., Waldemar, G., Soininen, H., Lötjönen, J., 2011. A disease state fingerprint for evaluation of Alzheimer's disease. Journal of Alzheimer's Disease 27 (1), 163–176.

Nir, T. M., Jahanshad, N., Villalon-Reina, J. E., Toga, A. W., Jack, C. R., Weiner, M. W., Thompson, P. M., ADNI, et al., 2013. Effectiveness of regional DTI measures in distinguishing Alzheimer's disease, MCI, and normal aging. NeuroImage: clinical 3, 180–195.

Oishi, K., Faria, A., Jiang, H., Li, X., Akhter, K., Zhang, J., Hsu, J. T., Miller, M. I., van Zijl, P. C., Albert, M., et al., 2009. 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.

Petersen, R. C., Aisen, P., Beckett, L., Donohue, M., Gamst, A., Harvey, D., Jack, C., Jagust, W., Shaw, L., Toga, A., et al., 2010. Alzheimer's Disease Neuroimaging Initiative (adni) clinical characterization. Neurology 74 (3), 201–209.

Raj, A., Kuceyeski, A., Weiner, M., 2012. A network diffusion model of disease progression in dementia. Neuron 73 (6), 1204–1215.

Reuter, M., Schmansky, N. J., Rosas, H. D., Fischl, B., 2012. Within-subject template estimation for unbiased longitudinal image analysis. Neuroimage 61 (4), 1402–1418.

Sabuncu, M. R., Desikan, R. S., Sepulcre, J., Yeo, B. T. T., Liu, H., Schmansky, N. J., Reuter, M., Weiner, M. W., Buckner, R. L., Sperling, R. A., et al., 2011. The dynamics of cortical and hippocampal atrophy in Alzheimer disease. Archives of Neurology 68 (8), 1040–1048.

Sarica, A., Antonio, C., Aldo, Q., Vince, C., 2018. A machine learning neuroimaging challenge for automated diagnosis of mild cognitive impairment. in press.

Scahill, R. I., Schott, J. M., Stevens, J. M., Rossor, M. N., Fox, N. C., 2002. 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.

Seeley, W. W., Crawford, R. K., Zhou, J., Miller, B. L., Greicius, M. D., 2009. Neurodegenerative diseases target large-scale human brain networks. Neuron 62 (1), 42–52.

Villemagne, V. L., Burnham, S., Bourgeat, P., Brown, B., Ellis, K. A., Salvado, O., Szoeke, C., Macaulay, S. L., Martins, R., Maruff, P., et al., 2013. Amyloid β deposition, neurodegeneration, and cognitive decline in sporadic Alzheimer's disease: a prospective cohort study. The Lancet Neurology 12 (4), 357–367.

Weiner, M. W., Veitch, D. P., Aisen, P. S., Beckett, L. A., Cairns, N. J., Green, R. C., Harvey, D., Jack, C. R., Jagust, W., Morris, J. C., et al., 2017. Recent publications from the Alzheimer's Disease Neuroimaging Initiative: Reviewing progress toward improved AD clinical trials. Alzheimer's & dementia: the journal of the Alzheimer's Association 13 (4), e1–e85.

Winblad, B., Amouyel, P., Andrieu, S., Ballard, C., Brayne, C., Brodaty, H., Cedazo-Minguez, A., Dubois, B., Edvardsson, D., Feldman, H., et al., 2016. Defeating alzheimer's disease and other dementias: a priority for european science and society. The Lancet Neurology 15 (5), 455–532.

Yang, E., Farnum, M., Lobanov, V., Schultz, T., Raghavan, N., Samtani, M. N., Novak, G., Narayan, V., DiBernardo, A., 2011. Quantifying the pathophysiological timeline of Alzheimer's disease. Journal of Alzheimer's Disease 26 (4), 745–753.

Young, A. L., Oxtoby, N. P., Daga, P., Cash, D. M., Fox, N. C., Ourselin, S., Schott, J. M., Alexander, D. C., 2014. A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain 137 (9), 2564–2577.

Young, J., Modat, M., Cardoso, M. J., Mendelson, A., Cash, D., Ourselin, S., ADNI, et al., 2013. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clinical 2, 735–745.

Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D., ADNI, et al., 2011. Multimodal classification of Alzheimer's disease and mild cognitive impairment. Neuroimage 55 (3), 856–867.

Zhou, J., Gennatas, E. D., Kramer, J. H., Miller, B. L., Seeley, W. W., 2012. Predicting regional neurodegeneration from the healthy brain functional connectome. Neuron 73 (6), 1216–1227.