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[7] Polina Golland[1] Stefan Klein[3] 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 NeuroImaging, University of Southern California, USA
5. Center for Imaging of Neurodegenerative Diseases, UCSF, USA
6. Department of Radiology and Nuclear Medicine, VU Medical Centre, Netherlands
7. Dementia Research Centre, UCL Institute of Neurology, UK

Slides available online at http://razvan.csail.mit.edu

What to expect from my presentation

- Which biomarkers can we predict, and which we cannot? (clinical diagnosis, MRI, cognitive tests)

What to expect from my presentation

- Which biomarkers can we predict, and which we cannot? (clinical diagnosis, MRI, cognitive tests)
- What is the state-of-the-art in Alzheimer's prediction?

What to expect from my presentation

- Which biomarkers can we predict, and which we cannot? (clinical diagnosis, MRI, cognitive tests)
- What is the state-of-the-art in Alzheimer's prediction?
- What are the winner algorithms? Should I use deep learning or not?

What to expect from my presentation

- Which biomarkers can we predict, and which we cannot? (clinical diagnosis, MRI, cognitive tests)
- What is the state-of-the-art in Alzheimer's prediction?
- What are the winner algorithms? Should I use deep learning or not?
- Consensus (averaging over predictions of all teams): good or not?

What to expect from my presentation

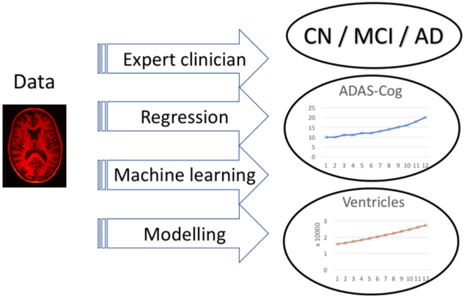
- Which biomarkers can we predict, and which we cannot? (clinical diagnosis, MRI, cognitive tests)
- What is the state-of-the-art in Alzheimer's prediction?
- What are the winner algorithms? Should I use deep learning or not?
- Consensus (averaging over predictions of all teams): good or not?
- Features: which ones are most informative? Do I need to pre-process those DTI scans, are MRIs not enough?

What to expect from my presentation

- Which biomarkers can we predict, and which we cannot? (clinical diagnosis, MRI, cognitive tests)
- What is the state-of-the-art in Alzheimer's prediction?
- What are the winner algorithms? Should I use deep learning or not?
- Consensus (averaging over predictions of all teams): good or not?
- Features: which ones are most informative? Do I need to pre-process those DTI scans, are MRIs not enough?
- How well do algorithms work on "real data", i.e. clinical trials?

TADPOLE is a Challenge to Predict the Progression of Individuals at Risk of AD

- Identify people that will develop Alzheimer's disease (AD) over the next 1-5 years.
    - Predict three target domains: clinical diagnosis, MRI (Ventricle Volume) and cognition (ADAS-Cog 13)
- Evaluation data on 219 subjects acquired by ADNI
- TADPOLE was entirely **prospective** – evaluation data acquired after submission deadline: Nov 2017
- Why predict future evolution of AD?
    - No treatments for AD currently available
    - Select the right subjects for AD clinical trials



Data

Expert clinician

Regression

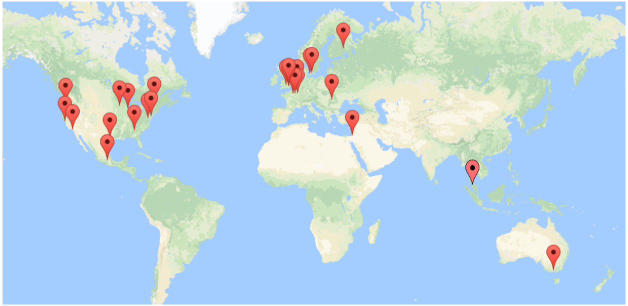
Machine learning

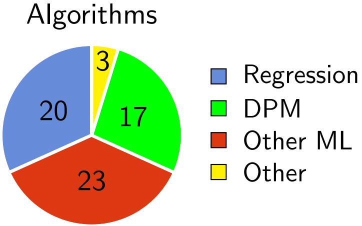
Modelling

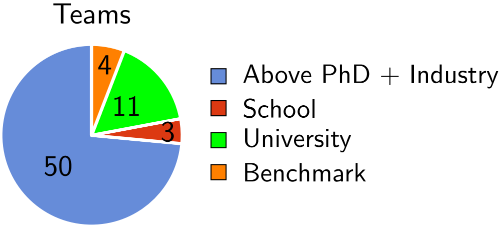
CN / MCI / AD

Submission statistics

33 teams from 12 countries







Prizes

- 30,000 GBP prize fund offered by sponsors:

Alzheimer's Society United Against Dementia

alzheimer's association®

Alzheimer's Research UK
The Power to Defeat Dementia

- Prizes were split according into six categories:

Prize amount	Outcome measure	Eligibility
£5,000	Diagnosis	all
£5,000	Cognition	all
£5,000	Ventricles	all
£5,000	Overall best	all
£5,000	Diagnosis	University teams
£5,000	Diagnosis	High-school teams

Results Outline

- Prediction results:
  - Clinical diagnosis
  - Ventricle volume
  - Cognition
- Overall winners & winning strategy
- Consensus methods
- Results on limited dataset mimicking clinical trial
- Most informative features

Clinical Diagnosis prediction: Winner algorithms achieve considerable gains over best benchmarks and state-of-the-art

- MAUC error reduced by 58% compared to the best benchmark

- Winner (Frog) used a method based on gradient boosting (xgboost)

- TADPOLE algorithms pushed ahead the state-of-the-art:
  - Best/29 algos in CADDementia challenge had a diagnosis MAUC of 0.78
  - Best/15 algos (Morandi, NeuroImage, 2015) obtained AUC of 0.902

- Full results on TADPOLE website:
  https://tadpole.grand-challenge.org/Results

Team Name	**RANK MAUC**	**MAUC**
Frog	1	0.931
Threedays	2	0.921
EMC-EB	3	0.907
GlassFrog-SM	4-6	0.902
GlassFrog-Average	4-6	0.902
GlassFrog-LCMEM-HDR	4-6	0.902
Apocalypse	7	0.902
EMC1-Std	8	0.898
CBIL	9	0.897
CN2L-RandomForest	10	0.896
...	...	...
BenchmarkSVM	30	0.836
...	...	...

- MAUC - multiclass area under the receiver-operator curve

Ventricle prediction: Winner algorithms achieve considerable gains over best benchmarks

- MAE reduced by 58% compared to best benchmark

- Winner (EMC1) used a method based on disease progression models

- No previous state-of-the-art due to lack of studies predicting ventricles

FileName	Rank Ventricles	MAE Ventricles
EMC1-Std	1-2	0.4116
EMC1-Custom	1-2	0.4116
lmaUCL-Covariates	3	0.4155
lmaUCL-Std	4	0.4207
BORREGOTECMTY	5	0.4299
lmaUCL-halfD1	6	0.4402
CN2L-NeuralNetwork	7	0.4409
SBIA	8	0.4410
EMC-EB	9	0.4466
Frog	10	0.4469
VikingAI-Logistic	11-12	0.4534
VikingAI-Sigmoid	11-12	0.4534
CBIL	13	0.4625
...	...	...
BenchmarkMixedEffectsAPOE	23	0.5664
...	...	...

- MAE - mean absolute error

Cognition prediction: TADPOLE algorithms **fail to predict** significantly better than random

- RandomisedBest - best out of 100 random guesses
- Likely too much noise in cognitive test (ADAS-Cog 13)
- Methods might be better than random over longer time-windows (> 2 years)

FileName	RANK	MAE
	Cognition	Cognition
RandomisedBest	-	4.52
FortuneTellerFish-Control	1	4.70
BenchmarkMixedEffectsAPOE	2	4.75
FortuneTellerFish-SuStaIn	3	4.81
Frog	4	4.85
Mayo-BAI-ASU	5	4.98
CyberBrains	6	5.16
VikingAI-Sigmoid	7	5.20
GlassFrog-Average	8	5.26
CN2L-Average	9	5.31
CN2L-NeuralNetwork	10	5.36
DIKU-GeneralisedLog-Std	11-12	5.40
DIKU-GeneralisedLog-Custom	11-12	5.40
...	...	...

- MAE - mean absolute error

There was no clear winner method. Deep learning not among top entries.

- Deep Learning

Rank	Diagnosis
1	Gradient boosting
2	Random forest
3	SVM
4-6	Multi state model
4-6	Multi state model
4-6	Multi state model
7	SVM
8	DPM+SVM
9	LSTM
10	Random Forest
11	DPM+SVM
12	feed-forward NN
13-14	Bayesian classifier/LDA + DPM
13-14	Bayesian classifier/LDA + DPM
15	Aalen model
16	DPM + ordered logit model
17	Random forest
...	...

Rank	Ventricles
1-2	DPM + spline regression
1-2	DPM + spline regression
3	Multi-task learning
4	Multi-task learning
5	Ensemble of regression + hazard
6	Multi-task learning
7	RNN
8	Linear mixed effects
9	SVM regressor
10	Gradient boosting
11-12	DPM
11-12	DPM
13	LSTM
14	DPM
15	DPM
16	RNN+RF
17	RF
...	...

Consensus methods achieve top results

- Compared to the best TADPOLE submissions, consensus reduced the error by 11% for Cognition (ADAS) and 8% for Ventricles

- Most methods make systematic errors, either over- or under-estimating the future measurements

Submission	Overall	Diagnosis		Cognition		Ventricles	
	Rank	Rank	MAUC	Rank	MAE	Rank	MAE
ConsensusMedian	-	-	0.925	-	5.12	-	**0.38**
Frog	**1**	**1**	**0.931**	4	4.85	10	0.45
ConsensusMean	-	-	0.920	-	**3.75**	-	0.48
EMC1-Std	2	8	0.898	23-24	6.05	1-2	0.41
VikingAI-Sigmoid	3	16	0.875	7	5.20	11-12	0.45
EMC1-Custom	4	11	0.892	23-24	6.05	1-2	0.41
CBIL	5	9	0.897	15	5.66	13	0.46
Apocalypse	6	7	0.902	14	5.57	20	0.52
...		...		...		...	

Prediction results on limited cross-sectional dataset mimicking a clinical trial are comparable to the full dataset

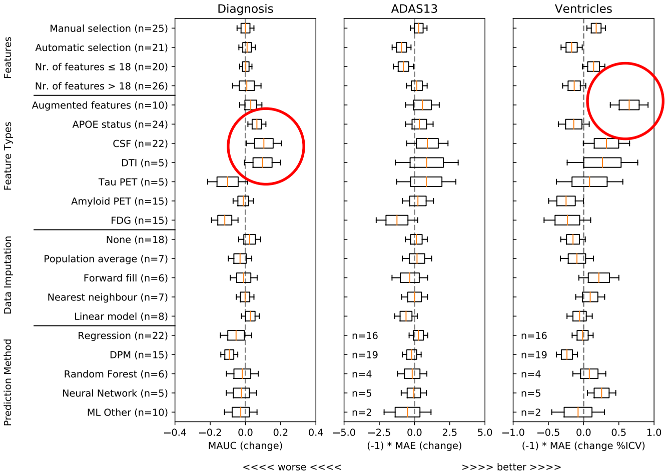
- Little loss of accuracy for the best methods
  - 0.48 vs 0.42 for ventricle MAE
  - 0.917 vs 0.931 for diagnosis MAUC

- Results suggest TADPOLE methods could be applied to clinical trial settings

	Overall	Diagnosis		Cognition		Ventricles	
Submission	Rank	Rank	MAUC	Rank	MAE	Rank	MAE
ConsensusMean	-	-	**0.917**	-	4.58	-	0.73
ConsensusMedian	-	-	0.905	-	5.44	-	0.71
GlassFrog-Average	**1**	2-4	0.897	5	5.86	3	0.68
GlassFrog-LCMEM-HDR	2	2-4	0.897	9	6.57	**1**	**0.48**
GlassFrog-SM	3	2-4	0.897	4	5.77	9	0.82
Tohka-Ciszek-RandomForestLin	4	11	0.865	2	4.92	10	0.83
RandomisedBest	-	-	0.811	-	4.54	-	0.92
...		...		...		...	

What matters for good predictions?

- DTI and CSF features for clinical diagnosis prediction
- Augmented features for ventricle prediction
- However, further analysis needs to be done to make clear conclusions



Conclusions

- Which biomarkers can we predict, and which we cannot?
  - YES: diagnosis, ventricles
  - NO: cognition (ADAS-Cog 13)

Conclusions

- Which biomarkers can we predict, and which we cannot?
    - YES: diagnosis, ventricles
    - NO: cognition (ADAS-Cog 13)

- What is the state-of-the-art in Alzheimer's prediction

Diagnosis MAUC	Cognition MAE	Ventricles MAE
0.931	-	0.41

Conclusions

- Which biomarkers can we predict, and which we cannot?
  - YES: diagnosis, ventricles
  - NO: cognition (ADAS-Cog 13)

- What is the state-of-the-art in Alzheimer's prediction

Diagnosis MAUC	Cognition MAE	Ventricles MAE
0.931	-	0.41

- What are the winner algorithms? Should I use deep learning or not?
  - No clear winner
  - Clinical diagnosis: gradient boosting
  - Ventricle MAE: disease progression model
  - Best deep learning algo: 5th place

Conclusions

- Which biomarkers can we predict, and which we cannot?
  - YES: diagnosis, ventricles
  - NO: cognition (ADAS-Cog 13)

- What is the state-of-the-art in Alzheimer's prediction

Diagnosis MAUC	Cognition MAE	Ventricles MAE
0.931	-	0.41

- What are the winner algorithms? Should I use deep learning or not?
  - No clear winner
  - Clinical diagnosis: gradient boosting
  - Ventricle MAE: disease progression model
  - Best deep learning algo: 5th place

- Consensus (averaging over teams' predictions): good or not?
  - Consensus achieves top results
  - Diagnosis: 11% better than TADPOLE best for cognition
  - Ventricles: 8% better than TADPOLE best

Conclusions

- Which biomarkers can we predict, and which we cannot?
  - YES: diagnosis, ventricles
  - NO: cognition (ADAS-Cog 13)

- What is the state-of-the-art in Alzheimer's prediction

Diagnosis MAUC	Cognition MAE	Ventricles MAE
0.931	-	0.41

- What are the winner algorithms? Should I use deep learning or not?
  - No clear winner
  - Clinical diagnosis: gradient boosting
  - Ventricle MAE: disease progression model
  - Best deep learning algo: 5th place

- Consensus (averaging over teams' predictions): good or not?
  - Consensus achieves top results
  - Diagnosis: 11% better than TADPOLE best for cognition
  - Ventricles: 8% better than TADPOLE best

- Features: which ones are most informative? Do I need to pre-process those DTI scans, are MRIs not enough?
  - Diagnosis: CSF and DTI
  - Ventricles: Augmented features

Conclusions

- Which biomarkers can we predict, and which we cannot?
  - YES: diagnosis, ventricles
  - NO: cognition (ADAS-Cog 13)
- What is the state-of-the-art in Alzheimer's prediction

Diagnosis MAUC	Cognition MAE	Ventricles MAE
0.931	-	0.41

- What are the winner algorithms? Should I use deep learning or not?
  - No clear winner
  - Clinical diagnosis: gradient boosting
  - Ventricle MAE: disease progression model
  - Best deep learning algo: 5th place
- Consensus (averaging over teams' predictions): good or not?
  - Consensus achieves top results
  - Diagnosis: 11% better than TADPOLE best for cognition
  - Ventricles: 8% better than TADPOLE best
- Features: which ones are most informative? Do I need to pre-process those DTI scans, are MRIs not enough?
  - Diagnosis: CSF and DTI
  - Ventricles: Augmented features
- How well do algorithms work on "real data"? i.e. clinical trials
  - minor loss in prediction performance
  - 0.917 vs 0.931 on diagnosis prediction

Next steps

- Manuscript in preparation
- TADPOLE SHARE
  - share methods for validation and further development
  - 11 teams already sharing
  - Lead by Esther Bron: e.bron@erasmusmc.nl
- AAIC 2020 special symposium
- Follow-on evaluations as more ADNI data becomes available
- Challenge still ongoing, D4 leaderboard now live

netherlands eScience center

Acknowledgements

- **Challenge Participants**
- **Sponsors**

Alzheimer's Society
United Against Dementia

Alzheimer's Research UK
The Power to Defeat Dementia

alzheimer's association®

- **Funders**

SEVENTH FRAMEWORK PROGRAMME

EPSRC
Pioneering research and skills

NAC

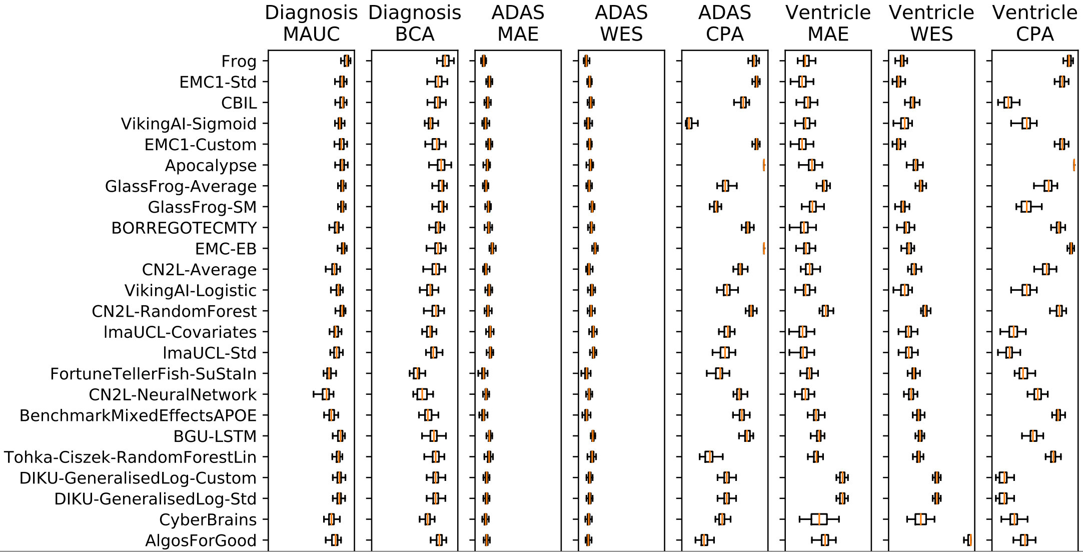
Submissions

Submission	Extra Features	Nr. of features	Missing data imputation	Diagnosis prediction	ADAS/Vent. prediction
Submission	Feature selection	Number of features	Missing data imputation	Diagnosis prediction	ADAS/Vent. Prediction
AlgosForGood	Manual	16+5*	forward-filling	Aalen model	linear regression
Apocalypse	Manual	16	population average	SVM	linear regression
ARAMIS-Pascal	Manual	20	population average	Aalen model	-
ATRI-Biostat-JMM	automatic	15	random forest	random forest	linear mixed effects model
ATRI-Biostat-LTJMM	automatic	15	random forest	random forest	DPM
ATRI-Biostat-MA	automatic	15	random forest	random forest	DPM + linear mixed effects model
BGU-LSTM	automatic	67	none	feed-forward NN	LSTM
BGU-RF/ BGU-RFFIX	automatic	67+1340*	none	semi-temporal RF	semi-temporal RF
BIGS2	automatic	all	Iterative Soft-Thresholded SVD	RF	linear regression
Billabong (all)	Manual	15-16	linear regression	linear scale	non-parametric SM
BORREGOSTECMTY	automatic	100 + 400*	nearest-neighbour	regression ensemble	ensemble of regression + hazard models
BravoLab	automatic	25	hot deck	LSTM	LSTM
CBIL	Manual	21	linear interpolation	LSTM	LSTM
Chen-MCW	Manual	9	none	linear regression	DPM
CN2L-NeuralNetwork	automatic	all	forward-filling	RNN	RNN
CN2L-RandomForest	Manual	$>$200	forward-filling	RF	RF
CN2L-Average	automatic	all	forward-filling	RNN/RF	RNN/RF
CyberBrains	Manual	5	population average	linear regression	linear regression
DIKU (all)	semi-automatic	18	none	Bayesian classifier/LDA + DPM	DPM
DIVE	Manual	13	none	KDE+DPM	DPM
EMC1	automatic	250	nearest neighbour	DPM + 2D spline + SVM	DPM + 2D spline
EMC-EB	automatic	200-338	nearest-neighbour	SVM classifier	SVM regressor
FortuneTellerFish-Control	Manual	19	nearest neighbour	multiclass ECOC SVM	linear mixed effects model
FortuneTellerFish-SuStaIn	Manual	19	nearest neighbour	multiclass ECOC SVM + DPM	linear mixed effects model + DPM
Frog	automatic	70+420*	none	gradient boosting	gradient boosting
GlassFrog-LCMEM-HDR	semi-automatic	all	forward-fill	multi-state model	DPM + regression
GlassFrog-SM	Manual	7	linear model	multi-state model	parametric SM
GlassFrog-Average	semi-automatic	all	forward-fill/linear	multi-state model	DPM + SM + regression
IBM-OZ-Res	Manual	Oct-15	filled with zero	stochastic gradient boosting	stochastic gradient boosting
ITESMCEM	Manual	48	mean of previous values	RF	LASSO + Bayesian ridge

Performance metrics

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.

Confidence intervals on longitudinal dataset D2



Prize winners

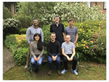
Frog: overall **TADPOLE champions** & clinical status winners

Apocalypse: Uni. Student winners

EMC1: ventricle volume winners

Chen: high school student winners

CyberBrains: high school student runners-up



GlassFrog: cross-sectional prediction winners

Category	Team	Members	Institution	Country	Prize
Overall best	Frog	Keli Liu, Paul Manser, Christina Rabe	Genentech	USA	£5000
Clinical Diagnosis	Frog	Keli Liu, Paul Manser, Christina Rabe	Genentech	USA	£5000
Ventricle volume	EMC1	Vikram Venkatraghavan, Esther Bron, Stefan Klein	Erasmus MC	Netherlands	£5000
Best university team	Apocalypse	Manon Ansart	ICM, INRIA	France	£5000
High-School (best)	Chen-MCW	Gang Chen	Medical College Wisconsin	USA	£5000
High-School (runner up)	CyberBrains	Ionut Buciuman, Alex Kelner, Raluca Pop, Denisa Rimocea, Kruk Zsolt	Vasile Lucaciu College	Romania	£2500
Overall best D3 prediction	GlassFrog	Steven Hill, Brian Tom, Anais Rouanst, Zhiyue Huang, James Howlett, Steven Kiddle, Simon R. White, Sach Mukherjee, Bernd Taschler	Cambridge University	UK	£2500