Contents lists available at ScienceDirect

NeuroImage

journal homepage: www.elsevier.com/locate/neuroimage

Ten years of image analysis and machine learning competitions in dementia

Esther E. Bron [a,*], Stefan Klein [a], Annika Reinke [b], Janne M. Papma [c], Lena Maier-Hein [b], Daniel C. Alexander [d], Neil P. Oxtoby [d]

[a] Department of Radiology and Nuclear Medicine, Erasmus MC, Rotterdam, the Netherlands
[b] Division of Computer Assisted Medical Interventions (CAMI), German Cancer Research Center (DKFZ), Heidelberg 69120, Germany
[c] Department of Neurology, Erasmus MC, Rotterdam, the Netherlands
[d] Centre for Medical Image Computing, Department of Computer Science, University College London, London WC1E 6BT, UK

ABSTRACT

Machine learning methods exploiting multi-parametric biomarkers, especially based on neuroimaging, have huge potential to improve early diagnosis of dementia and to predict which individuals are at-risk of developing dementia. To benchmark algorithms in the field of machine learning and neuroimaging in dementia and assess their potential for use in clinical practice and clinical trials, seven grand challenges have been organized in the last decade: MIRIAD (2012), Alzheimer's Disease Big Data DREAM (2014), CADDementia (2014), Machine Learning Challenge (2014), MCI Neuroimaging (2017), TADPOLE (2017), and the Predictive Analytics Competition (2019). Based on two challenge evaluation frameworks, we analyzed how these grand challenges are complementing each other regarding research questions, datasets, validation approaches, results and impact.

The seven grand challenges addressed questions related to screening, clinical status estimation, prediction and monitoring in (pre-clinical) dementia. There was little overlap in clinical questions, tasks and performance metrics. Whereas this aids providing insight on a broad range of questions, it also limits the validation of results across challenges. The validation process itself was mostly comparable between challenges, using similar methods for ensuring objective comparison, uncertainty estimation and statistical testing. In general, winning algorithms performed rigorous data pre-processing and combined a wide range of input features.

Despite high state-of-the-art performances, most of the methods evaluated by the challenges are not clinically used. To increase impact, future challenges could pay more attention to statistical analysis of which factors (i.e., features, models) relate to higher performance, to clinical questions beyond Alzheimer's disease, and to using testing data beyond the Alzheimer's Disease Neuroimaging Initiative. Grand challenges would be an ideal venue for assessing the generalizability of algorithm performance to unseen data of other cohorts. Key for increasing impact in this way are larger testing data sizes, which could be reached by sharing algorithms rather than data to exploit data that cannot be shared. Given the potential and lessons learned in the past ten years, we are excited by the prospects of grand challenges in machine learning and neuroimaging for the next ten years and beyond.

1. Introduction

Over the past decade there has been a great research interest in artificial intelligence approaches assisting the management of dementia. Especially, machine learning methods exploiting multi-parametric biomarkers from large-scale datasets have shown huge potential to detect dementia at an early stage, to provide insight into etiology and to monitor and predict the development of symptoms. Because of this huge potential and increased availability of data, many articles presenting novel algorithms have been published, most of them using neuroimag-ing as a key biomarker (Ansart et al., 2021; Falahati et al., 2014; Oxtoby and Alexander, 2017; Rathore et al., 2017).

Since 15 years, grand challenges have been organized in the biomedical image analysis research field. These are international benchmarks in competition form that have the goal of objectively comparing algorithms for a specific task on the same clinically representative data using the same evaluation protocol. In such challenges, the organizers supply reference data and evaluation measures on which researchers can evaluate their algorithms. Over the past years, the number and the impact of such grand challenges has increased (Maier-Hein et al., 2018). Also in

the field of machine learning algorithms for dementia, such grand challenges have been organized to gain insight into successful approaches and their potential for use in clinical practice and clinical trials (Allen et al., 2016; Bron et al., 2015; Cash et al., 2015; Fisch et al., 2021; Marinescu et al., 2018; Sabuncu and Konukoglu, 2015; Sarica et al., 2018).

With the trend of grand challenges in the field of biomedical image analysis becoming more important, two frameworks have been proposed in the literature to evaluate and structure the design of grand challenges and assess the effect of the design on the interpretation and impact of challenge results. First, Maier-Hein et al. (2018) performed an evaluation of 150 challenges and identified major problems, which led to best practice advice. Regarding challenge design, they found that in the majority of challenges it was not possible to reproduce results, to adequately interpret them, and compare them across challenges, since relevant information was often missing and since challenge design (e.g. a choice of metrics and methods for rank computation) was highly heterogeneous. In addition, the authors concluded that the ranking of an algorithm in a challenge was not robust, but highly dependent on a number of design choices such as the test data sets used for validation, the observers who annotated the data and the metrics chosen for performance assessment, as well as the methods used for aggregating values. Second, Mendrik and Aylward (2019) categorized two types of challenges: challenges that aim to generate insight (insight challenges) and those that aim to solve a problem (deployment challenges). The main purpose of this categorization was to guide researchers in setting-up challenges of which the objectives match with its design and conclusion. Mixing the insight design with the deployment design will negatively impact the clinical significance of the challenge and generalizability of methods. While the first framework is based on a quantitative approach and the latter framework provides a mostly qualitative view on grand challenges, the identified topics of importance for evaluation of challenge design and results largely overlap.

In this review, we will evaluate the grand challenges in dementia based on the topics defined by the frameworks of Maier-Hein et al. (2018) and Mendrik and Aylward (2019). Our main research question is how the grand challenges on this topic are strengthening and complementing each other. First, we will highlight unanswered clinical questions in dementia and the related tasks that are being addressed by machine learning algorithms. Subsequently, we will summarize the included grand challenges, with an emphasis on the role of neuroimaging. We will summarize design choices, including the alignment of their purpose with the relevant questions in dementia, the data and truth criteria, and performance metrics. In addition, we will outline the results and assess their impact. Finally, we will reflect on the status of the field and the open questions that remain to be addressed by future grand challenges.

2. Questions in clinical and pre-clinical dementia

Worldwide 50 million people are estimated to be living with dementia and Alzheimer's disease is the most prevalent underlying cause (Association, 2020). For Alzheimer's disease, the pre-clinical phase is found to take about 20 years before the onset of clinical symptoms (Gordon et al., 2018). Diagnosis of dementia is highly challenging and usually takes a substantial period of time after the first clinical symptoms arise (Van Vliet et al., 2013). After diagnosis, novel challenges are related to prediction and monitoring of the disease, both in the natural disease course and under treatment with potential disease-modifying drugs in clinical trials. Along different phases of dementia, clinically relevant questions could be identified to which machine learning methods can contribute by solving related tasks exploiting multi-parametric biomarkers.

First, in the pre-clinical stage, i.e. a community-dwelling population without cognitive complaints, a key question is to identify persons at-risk for developing dementia based on biomarker measurements at an individual level. Several approaches for biomarker-based **screen-**ing have been proposed (Cole et al., 2017; Licher et al., 2019; Wang et al., 2019). Regarding neuroimaging, a frequently studied screening biomarker is the brain age gap, i.e. the difference between chronological age and age predicted from MRI using machine learning, which has been shown to identify individuals at risk for dementia in the general population (Wang et al., 2019). While highly important for improving disease understanding, there is currently no clinical motivation for population screening. As currently no disease-modifying medication could be offered, the benefit for individuals at-risk to identify them before symptoms develop may be very limited. However, while the first Alzheimer's disease drug is entering the market and while the search for disease-modifying treatments continues, screening of individuals at a very early stage may aid inclusion for clinical trials (Panegyres et al., 2016).

Second, in an early-stage clinical population, i.e. patients with cognitive complaints that visit a memory clinic, the key question is whether the person is developing dementia. The purpose of **clinical status** estimation is to understand the current 'condition' of the patient and to establish a clinical diagnosis that enables timely decision making regarding care and treatment (Prince et al., 2011). Relevant components of this clinical status are whether there is a cognitive impairment (e.g. quantified by a memory score, or fulfilling diagnostic criteria such as for mild cognitive impairment or dementia (Albert et al., 2011; McKhann et al., 2011)) and what the underlying cause of this cognitive impairment is (e.g. Alzheimer's disease).

In addition to current clinical status, in this early-stage clinical population, i.e. patients with subjective cognitive decline or a diagnosis of mild cognitive impairment, a second key question is whether symptoms will develop into dementia and in what time frame. The purpose of **prediction** is to understand the future 'condition' of the patient, mainly to inform treatment and care decisions, but may also guide inclusion of individuals at risk for dementia in clinical trials. Outcomes of interest for prediction are for example diagnosis and clinical functioning at a future time point or time-to-conversion to dementia.

Lastly, in a later-stage clinical population, after diagnosis has been established, a key question involves **monitoring** the disease development and, if applicable, the patient's response to treatment. In follow-up consultations, biomarkers could be used to assess the disease progression in a quantitative way.

3. Grand challenges in dementia

Here we discuss grand challenges on screening, clinical status estimation, prediction and monitoring in (pre-clinical) dementia with a focus on neuroimaging data. Challenges were found by searching challenges listed by grand-challenge.org and Maier-Hein et al. (2018) as well as searching publications through Pubmed and Google based on key words. Candidate studies were screened by reading the title and if necessary the abstract or website. No criterion on publication year was used. Details regarding the identification and screening can be found in the supplementary information. Inclusion criteria for studies were (1) being organized as a competition, (2) neuroimaging data included as input, (3) challenge tasks were related to screening, clinical status, prediction, or monitoring and (4) medical application area was (pre-clinical) dementia. Hence, we excluded benchmarks that were not set up as a grand challenge (Ansart et al., 2021; Cuingnet et al., 2011; Wen et al., 2020), grand challenges that did not include neuroimaging but used for example clock-drawing test data[1] or speech data (Luz et al., 2020), grand challenges on brain image registration or segmentation (Klein et al., 2009; Mendrik et al., 2015), and grand challenges in diseases other than dementia.

Based on this, we evaluated seven grand challenges:

- Minimal Interval Resonance Imaging in Alzheimer's Disease atrophy challenge (MIRIAD, 2012): the challenge task was to quantify vol-

---

[1] aicrowd.com/challenges/addi-alzheimers-detection-challenge.

umes and atrophy rates of several brain structures based on a pair of images blinded for time points with the objective of monitoring dementia (Cash et al., 2015). A total of 9 teams participated.

- Alzheimer's Disease Big Data DREAM Challenge (DREAM, 2014)[2]: the challenge consisted of three tasks: (1) to predict change in cognitive scores based on genotypes imputed from single-nucleotide polymorphism (SNP) array data, (2) to screen for amyloid positivity in cognitively normals based on genotypes, and (3) to estimate the mini-mental state examination (MMSE) score based on T1-weighted MRI (Allen et al., 2016). A total of 32 teams submitted final results (Task 1: 18 teams, 2: 11 teams, 3: 13 teams; all tasks: 4 teams).
- Computer-aided diagnosis of dementia challenge (CADDementia, 2014)[3]: the challenge task was to classify subjects into the diagnostic categories of Alzheimer's disease, mild cognitive impairment, and cognitively normal controls using a baseline MRI scan (Bron et al., 2015). Fifteen research teams participated with a total of 29 algorithms.
- Machine Learning Challenge (MLC, 2014)[4]: the challenge consisted of several tasks of binary classification and continuous regression of clinical phenotypes based on T1-weighted MRI. Tasks were blinded to participants. One of the tasks, regression of age in healthy subjects, was included in this evaluation; others were out-of-scope (e.g. schizophrenia diagnosis) (Sabuncu and Konukoglu, 2015). A total of 12 teams participated.
- MCI neuroimaging challenge (MCI-NI, 2017)[5]: the challenge task was to classify subjects into four clinical categories (Alzheimer's disease, cognitively normal, mild cognitive impairment (MCI), and MCI subjects that converted to Alzheimer's disease) using a baseline MRI scan and MMSE score, thereby combining current clinical status and prediction into one task (Sarica et al., 2018). 19 teams participated in the challenge with a total of 347 entries.
- The Alzheimer's Disease Prediction Of Longitudinal Evolution Challenge (TADPOLE, 2017)[6]: the challenge consisted of three tasks in predicting future evolution of individuals at risk for Alzheimer's disease (Marinescu et al., 2021; 2018). These tasks were to predict three key outcomes: (1) clinical diagnosis, (2) Alzheimer's Disease Assessment Scale Cognitive Subdomain (ADAS-Cog13), and (3) total volume of the ventricles. A total of 33 teams participated with 92 algorithms.
- Predictive Analytics Competition (PAC, 2019)[7]: the challenge task was to estimate brain age from healthy individuals (age range: 17–90 years) based on T1-weighted MRI (Fisch et al., 2021). As the literature frequently reported a bias between predicted brain age and chronological age, i.e. overpredicting age of young and underpredicting age of elderly individuals (Treder et al., 2021), a second task was to estimate brain age with a minimal bias towards chronological age. A total of 79 teams participated.

4. Challenge design

4.1. Executive summary

We assessed the design choices made by the seven challenges, as these largely determine the reproducibility, interpretability and comparability of the challenges (Maier-Hein et al., 2018). Table 1 summarizes the tasks of all challenges and relates them to the clinical questions. As all challenges had a unique task, different choices have been made regarding data, metrics, and methodology for ranking, which limits comparability. On the other hand, methods for avoiding cheating, uncer-

---

[2] synapse.org/##!Synapse:syn2290704.
[3] caddementia.grand-challenge.org.
[4] codalab.org/competitions/1471.
[5] kaggle.com/c/mci-prediction.
[6] tadpole.grand-challenge.org.
[7] frontiersin.org/research-topics/13501.

Table 1
Overview of the tasks for the seven grand challenges in dementia, relating to the clinical problems of screening, clinical status estimation, prediction and monitoring. AD: Alzheimer's disease, MCI: mild cognitive impairment, CN: cognitively normal.

	Screening	Clinical status	Prediction	Monitoring
MIRIAD (Cash et al., 2015)				Estimation of atrophy from MRI.
DREAM (Allen et al., 2016)	Screening for amyloid pathology from genetics, clinical data and demographics (Task 2).	Estimation of MMSE from MRI (Task 3).	Prediction of MMSE from genetics, clinical data and demographics (Task 1).	
CADDementia (Bron et al., 2015)		Classification of AD/MCI/CN from MRI.		
MLC[4]	Estimation of age from MRI.			
MCI-NI[5]		Classification of AD/MCI convertor/MCI stable/CN from MRI.		
TADPOLE (Marinescu et al., 2018)			Prediction of AD/MCI/CN, ADAS-Cog13 and ventricle size from multi-parametric biomarkers.	
PAC (Fisch et al., 2021)	Estimation of age from MRI.			

Table 2

Overview of data for the seven challenges. Leaderboard data was used to compute algorithm performance for a public leaderboard on the challenge website during the competition phase.

	Training data	Extra training data	Leaderboard data	Test data
MIRIAD (Cash et al., 2015)	69 subjects (46 Alzheimer's disease, 23 controls), 1–12 scans in 9 visits during 1–2 years. 708 images. (Malone et al., 2013)	N.A.	N.A.	Same as training data.
DREAM (Allen et al., 2016)	ADNI data: clinical data, imputed genotypes, (processed) MRI data (Task 1: N = 767, 2: N = 176, 3: N = 628).	AddNeuroMed: clinical data, imputed genotypes (Task 1: N = 409)	A subset of the test data (Task 1: N = 588, 2: N = 129, 3: N = 94). Leaderboard usage was limited (Task 1: 100x, Task 2 and 3: 50x).	ROS/MAP: clinical data, imputed genotypes (Task 1: N = 1762, 2: N = 257) and AddNeuroMed: clinical data, imputed genotypes, (processed) MRI data (Task 3: N = 182)
CADDementia (Bron et al., 2015)	CADDementia data from Erasmus MC, VU University Medical Center and University of Porto: MRI + diagnostic labels (N = 30)	Any training data permitted.	N.A.	CADDementia data: MRI (N = 354). Data was roughly balanced over the classes.
MLC[4]	OASIS data (Sabuncu and Konukoglu, 2015): pre-processed MRI + label (N = 315, split into a training and testing set).	N.A.	All test data.	OASIS data: pre-processed MRI, split from training data, size unknown.
MCI-NI[5]	ADNI data: randomly selected, pre-processed MRI, MMSE score and classification label (N = 240).	N.A.	Approximately half of the test data (including dummy data).	ADNI (N = 160), dummy data (N = 340). Data was balanced over the classes. Dummy data was excluded for the final evaluation.
TADPOLE (Marinescu et al., 2018)	ADNI data: multi-parametric including (pre-processed) structural MRI, diffusion MRI, PET imaging, cerebrospinal fluid biomarkers and cognitive tests (N = 1737)	Any training data permitted.	A subset of the training data.	ADNI data: follow-up diagnosis, ADAS-Cog13 score and ventricle volume (N = 219).
PAC (Fisch et al., 2021)	Data from Imperial College London (Cole et al., 2017) and the Institute of Translational Psychiatry Münster: MRI + age (N = 2640)	Any training data permitted.	N.A.	Data from Imperial College London (Cole et al., 2017) and the Institute of Translational Psychiatry Münster: MRI (N = 660)

tainty estimation and statistical testing are quite similar. An overview of data provided for training and testing can be found in Table 2. Table 3 provides an overview of truth criteria, anti-cheating strategies, used performance metrics, uncertainty estimates and statistical tests. These design choices, including approaches for challenge and result dissemination, are discussed in more detail in the remainder of this section.

4.2. Research questions and tasks

The research questions of the seven challenges largely correspond to the four identified clinical questions regarding screening, clinical status, prediction and monitoring (Table 1). PAC, MLC and one of the tasks in DREAM (Task 2) address a screening question. PAC and MLC aim to find the most accurate model (and for PAC additionally the most accurate model under a small bias) for estimating brain age in a healthy population, DREAM aims to detect cognitive normals with abnormal amyloid pathology. Clinical status estimation is addressed by three challenges in a slightly different way. DREAM estimated the MMSE score from structural MRI (Task 3), while the other two challenges classified different diagnostic groups based on structural MRI (CADDementia: 3 groups, MCI-NI: 4 groups). Prediction is addressed by three challenges that predicted future outcomes in diagnosis (MCI-NI, TADPOLE), cognitive score (DREAM Task 1: MMSE score, TADPOLE: ADAS-Cog13) and MRI measures (TADPOLE: Ventricle volume). In contrary to the clinical status challenges, more diverse inputs were used for the prediction tasks (DREAM: genetics, MCI-NI: MRI, TADPOLE: MRI, PET, cognition and cerebrospinal fluid markers). Finally, one challenge addressed monitoring: MIRIAD estimated volume and volume changes from longitudinal structural MRI after diagnosis. All challenges were insight challenges (Mendrik and Aylward, 2019) that aim to provide insight into methodological choices for algorithms in the tasks of screening, clinical status, prediction and monitoring.

4.3. Data

Data set sizes differed substantially ranging from 30 to 2640 subjects for training and 69 to 1762 subjects for testing (Table 2). Multi-center data was used in all challenges but MIRIAD and MLC. No specific attention was given to diversity of the data. Whereas most challenges used same population training and testing data, DREAM explicitly used data sets of different cohorts. Four challenges allowed participants to use additional training data (DREAM, CADDementia, TADPOLE, PAC). The Alzheimer's Disease Neuroimaging Initiative (ADNI)[8] was used in four challenges: either as (optional) training data (DREAM, CADDementia) or both for training and testing (MCI-NI, TADPOLE). In addition, four challenges had a public leaderboard on their website showing performances on a data subset during the competition phase (DREAM, MLC, MCI-NI, TADPOLE).

All challenges used a fixed test data set, aiming to compare methods for one specific setting. Labels of the test data were blinded and submission of prediction results was requested. The test set was in most cases defined in a natural way, e.g. based on visit date (e.g. MIRIAD, TADPOLE) or using a separate cohort (e.g., DREAM, CADDementia). Other challenges used a random split into a training and testing (MLC, MCI-NI, PAC). The challenges could have provided a more complete view of performance in different settings by addressing variance in the challenge design, which can be done by randomizing as many sources of variation as possible and using multiple data splits (Bouthillier et al., 2021).

Whereas most of the challenges focused on brain MRI as only type of input, TADPOLE and DREAM specifically targetted multi-modal data including cognition, clinical data and genetics. The neuroimaging data was used in different ways; some challenges provided raw imaging data (CADDementia, PAC), whereas other provided pre-computed imaging measurements (MLC, MCI-NI) or allowed for both (DREAM, TADPOLE).

_____
[8] adni.loni.usc.edu.

Table 3

Overview of performance metrics and truth criteria for the seven challenges. AD: Alzheimer's disease, MCI: mild cognitive impairment, MMSE: mini-mental state examination, PET: positron emission tomography.

	Task	Truth criteria	Anti-cheating	Performance metrics	Uncertainty est.	Statistical tests
MIRIAD (Cash et al., 2015)	Estimation of atrophy.	N.A., instead reliability and variance of atrophy measurements were assessed (Fox et al., 2011).	Timepoint labels of scans blinded (only one baseline labeled)	Sample size to detect a 25% atrophy rate reduction at 80% power. Secondary: repeatability, consistency, inter- and intra-subject variance.	Confidence intervals from 2000 bootstrap samples	Non-parametric tests between pairs of methods (2000 bootstrap samples, $\alpha < 0.05$)
DREAM (Allen et al., 2016)	Prediction of cognition (Task 1)	MMSE score established after 2 years.	Test data outcomes were not released. A maximum of 2 entries per task per team.	Pearson's correlation, Spearman's correlation (using clinical input only and a combination of clinical and genetics input)	Confidence intervals from 100,000 bootstrap samples.	Non-parametric tests (10,000 permutations) to assess if methods performed better than random ($\alpha < 0.05$).
	Screening of amyloid pathology (Task 2)	Amyloid positivity from PET or cerebrospinal fluid.	”	Balanced accuracy, area-under-the-curve	”	”
	Estimation of cognition (Task 3).	MMSE score.	”	Pearson's correlation, Lin's concordance correlation coefficient	”	”
CADDementia (Bron et al., 2015)	Classification of AD, MCI, and cognitively normals.	Clinical diagnosis by multi-disciplinary consensus.	Test data labels not released. A maximum of 5 entries per team.	Accuracy. Secondary: Area-under-the-curve, True positive fraction	Confidence intervals from 1000 bootstrap samples	McNemar's test between pairs of methods ($\alpha < 0.05$).
MLC[4]	Estimation of age.	Chronological age.	Tasks and labels blinded, test labels not released. A maximum of 15 entries per team.	Root-mean-square error, Pearson's correlation. Secondary: Generalization of performance w.r.t. 5-fold cross-validation.	Confidence intervals from 10,000 bootstrap samples.	N.A.
MCI-NI[5]	Classification of AD, MCI converting to AD, stable MCI, and cognitively normals.	Diagnosis using clinical criteria and follow-up diagnosis after 48 months for MCI patients.	Public and private leaderboard, but not final test data, were inflated by dummy data. A maximum of 1 entry per team.	Accuracy	N.A.	N.A.
TADPOLE (Marinescu et al., 2021)	Prediction of a future status of AD, MCI, and cognitively normal (Task 1)	Diagnosis established at a follow-up time point.	Evaluation data acquired after challenge deadline. A maximum of 3 entries per team.	Area-under-the-curve. Secondary: Balanced classification accuracy	Confidence intervals from 50 bootstrap samples	Non-parametric test between pairs of methods (50 bootstrap samples of the test set, $\alpha < 0.05$).
	Prediction of ADAS-Cog13 and ventricle size (Task 2&3)	Variables established at a follow-up time point.	”	Mean absolute error. Secondary: weighted error score, coverage probability accuracy	”	Wilcoxon signed-rank test between pairs of methods ($\alpha < 0.05$).
PAC (Fisch et al., 2021)	Estimation of age	Chronological age.	Age for testing data not released. Limited time (1 week) between test data release and submission deadline.	Mean absolute error. Secondary: Bias using Spearman's correlation of predicted age difference and chronological age.	N.A.	N.A.

Pre-computing image-based features is an advantage for interpretation (differences between methods could be attributed to machine learning methodology only) and participation (more teams can participate, also those without expertise in MRI analysis), but may be a disadvantage for assessing state-of-the-art (variation is limited, promising algorithms may be excluded).

4.4. Truth criteria and anti-cheating strategies

To objectify the output of tasks, a reference standard, or truth criterion, should be in place. The used truth criteria by the challenges were clinical diagnosis, follow-up measurements and independent biomarker measurements (Table 3). These criteria do not provide a ground-truth and are all subject to some measurement error. Especially when using clinical diagnosis, it is possible that some of the patients are misdiagnosed as the clinical diagnosis does not take account of post-mortem histopathology or amyloid and tau measurements. None of the challenges reported an analysis of these specific measurement errors or an inter-rater analysis of clinical diagnosis. An alternative approach was

used by MIRIAD, which instead of using a direct truth criterion for atrophy quantification used multiple indirect truth criteria focused on variance (i.e., power computation, repeatability, consistency, intra- and intersubject variance).

To ensure objective comparison, all challenges put some 'anti-cheating' strategies in place to promote objective comparison. The most common strategy was to not release test data labels. Using ADNI data for testing however requires an alternative approach as all data is public. TADPOLE therefore required predictions to be submitted before acquisition of the test set by ADNI, and MCI-NI provided fake test data. In this last solution, fake data was not used for the final evaluation but was used for the public leaderboard during the competition phase. As results that include these fake data may have influenced classifier parameters and model selection, final classification performances could have been suboptimal (Castiglioni et al., 2018; Donnelly-Kehoe et al., 2018). Another complementary approach, used by all challenges, was to limit the numbers of submissions per team or limit the time between test data release and the submission deadline, in order to reduce the opportunity for model tuning and selection using test set knowledge.

4.5. Performance metrics

For classification tasks, common performance metrics were (balanced) accuracy and area-under-the-ROC-curve (AUC). Accuracy (i.e. correct classification rate) was chosen as primary metric in tasks with balanced data across classes (CADDementia, MCI-NI), while challenges with class imbalance used balanced accuracy or AUC as primary metrics (DREAM, TADPOLE). For regression tasks, common metrics were measures of the differences between values (mean absolute error, root-mean-square error; MLC, TADPOLE, PAC) and correlation (Pearson's or Spearman's; DREAM, MLC). In addition, TADPOLE included the weighted error score and coverage probability accuracy metrics that take account of confidence of the predictions given by the individual methods. MLC asked participants to also submit their 5-fold cross-validation results on the training set in order to analyze the generalization error. The MIRIAD challenge, which had no access to ground truth atrophy measurements, used as a primary metric the required sample size to detect atrophy rate reduction of 25% with 80% statistical power, related to potential application of such methods for patient monitoring in clinical trials.

All grand challenges came up with a final ranking of methods and announced a winner. Ranks were computed on one primary performance metric for each task. Where most challenges ranked methods based a single prediction, DREAM used the median ranking among 100,000 bootstrap resamplings of the test data. TADPOLE summed ranks for their three tasks to make one overall ranking. While such ranking and the possibility to 'win' a challenge is highly motivating, a risk of this approach may be in parameter tuning. Instead, collaborative challenges without winners may lead to less parameter tuning and could be more successful in learning about strengths and weaknesses of algorithms (Maier-Hein et al., 2018).

Most challenges used bootstrapping to estimate confidence intervals on the performance metrics, which is useful for detecting statistically meaningful improvements (Bouthillier et al., 2021). The number of bootstrap samples varied from 50 to 100,000. TADPOLE also evaluated confidence estimations of individual predictions provided by the participants, which is important in clinical application for knowing when estimates are unreliable. In addition, most challenges performed statistical tests, either to test for significance differences between pairs of methods or to test whether performance was significantly better than random guessing. Mostly non-parametric tests were used (permutation testing, Wilcoxon signed-rank), except for CADDementia (McNemar's test). MCI-NI and PAC did not provide strategies for uncertainty handling and statistical testing.

4.6. Dissemination

The methods for dissemination differed per challenge. All challenges had a challenge website, mostly connected to a challenge platform (Grand-challenge.org: 2, Kaggle: 1, Codalab: 1, Synapse: 1). Two challenges are still open for new submissions (CADDementia, TADPOLE). For MLC and PAC, websites were not available anymore during our evaluation, so organizers were contacted to provide missing information. Most challenges published an overview paper to present the challenge set-up and discuss results (MIRIAD, CADDementia, TADPOLE, DREAM; TADPOLE additionally published a design paper; CADDementia and DREAM additionally published workshop proceedings with short papers describing individual algorithms). PAC and MCI-NI instead published a special issue with individual papers for the algorithms and an editorial explaining and interpreting the results, and MLC did not publish. Whereas a special issue might be most useful for in-depth understanding and reuse of algorithms, an overview paper may ease results to be interpreted and conclusions to be implemented by the community. We successfully obtained all information required for the overview, which means that aspects determined relevant by Maier-Hein et al. (2018) were reported.

4.7. Limitations

All challenges made specific design choices that may have negatively affected their impact, such as using single center data (e.g. MIRIAD), not including the results on all tasks in the publication (e.g. DREAM), using clinical diagnosis as a reference standard (e.g. CADDementia), limiting the use of domain knowledge by blinding the context (e.g. MLC), mixing the clinical questions of current clinical status and prediction into one task (e.g. MCI-NI), using ADNI data for both training and testing (e.g. TADPOLE), or not providing strategies for uncertainty handling and statistical testing (e.g. PAC).

5. Challenge findings

5.1. Main findings

An overview of the main findings and best performing methods can be found in Table 4. All challenges aimed to obtain insight into the best strategy for solving the problem and the current state-of-the-art performance. They reported results as performance-based rankings of participating algorithms, compared to the literature and drew conclusions about the contribution of specific modeling choices (i.e. method, features).

There were methodological commonalities and differences across winning methods in the challenges. Winning algorithms generally had special attention for pre-processing of the data and combined a wide range of input features. Also, including more training data or pre-training on other data sets showed advantage. No single modeling methodology stood out among the winners across these challenges, which is perhaps unsurprising given the different objectives of each challenge and the time span involved (10 years) and the plethora of available methods in this field. There was a large variety in types of modeling, varying from Gaussian processes to gradient boosting and from random forests to convolutional neural networks. To illustrate the changes of the field over time: whereas in CADDementia (2014) only one participant used a neural network achieving a relatively low rank, PAC (2019) was dominated by convolutional neural networks achieving high performance.

Winners were announced for all except for two tasks in which algorithms did not outperform random guessing. In the second task of the DREAM challenge, which focused on prediction of cognition based on genetics, participants were unable to develop algorithms with predictive performances significantly better than random. According to Allen et al. (2016), this might be due to low sample size and trait heterogeneity, but also indicates the difficulty of the task, i.e. information about cognitive resilience is not easily discoverable from genetic analysis. Similarly in the TADPOLE challenge, the ADAS-Cog13 cognitive score was more difficult to forecast than clinical diagnosis or ventricle volume. The only method that performed better than random guessing was a simple mixed effects model. According to Marinescu et al. (2021), the difficulty could be due to variability in administering the cognitive tests, practice effects and a short follow-up time.

Most challenges looked at qualitative trends in factors –such as data types, features or models– leading to higher performance. This evaluation is usually not performed as a statistical analysis. Instead, as earlier concluded by Roß et al. (2021) for image analysis challenges in general, results analysis is often restricted to pure ranking tables, leaving important questions unanswered. For example, quantitative statistical analyses are desirable to properly understand which factors (i.e. features, models) contributed to the ranking, and to find failure modes for any given algorithm, data or objective. Such analysis were not performed by any of the challenges as the challenge design and number of submissions are insufficient.

5.2. Clinical impact

Most challenges generated novel insights into the best current strategy for a specific clinical problem related to screening, clinical status,

Table 4
Overview of the main findings of the seven challenges.

	Main research question	Main findings	Best methods
MIRIAD (Cash et al., 2015)	What is the best current strategy for atrophy measurement from MRI and what is the required sample size for hypothetical clinical trials using such methods?	Measures of volume change, particularly ventricle and whole brain, were consistent and robust, leading to stable sample size estimates. Hippocampal measures were more variable, likely due to the differing definitions of the structure. Direct measures of change had smaller variance than indirect measures.	The methods achieving the smallest required sample sizes were the Boundary Shift Integral (for whole brain atrophy measurement) and the combination of Demons-LCC registration and regional flux analysis (for ventricle and hippocampus atrophy measurement).
DREAM (Allen et al., 2016)	What are the current capabilities for estimation of cognition and prediction of cognitive decline using genetic and imaging data from public data resources?	Predictions of cognitive performance developed from genetic or structural imaging data were modest across a diverse set of models, with most algorithms performing roughly equivalently. The data was probably inadequate to support these tasks. There was no clinical value in prediction of cognitive decline based on genetic data.	The algorithm that generated the best absolute mean combined rank in task 1 and 3 used Gaussian process regression. For task 2, no method performed better than random.
CADDementia (Bron et al., 2015)	What is the best current strategy for classification of Alzheimer's disease and mild cognitive impairment based on MRI and how does it perform in a clinically representative multi-center data set?	The best performing algorithms incorporated features describing different properties of the scans such as shape and intensity features in addition to volume and thickness. Performance was additionally influenced by the classification model and training data size. While all methods outperformed random guessing, performance was too low for clinical application.	The method that achieved best accuracy used linear discriminant analysis of a combination of features measuring volume, thickness, shape and intensity relations of brain regions.
MLC[4]	What is the current state of the art in the field of neuroimage-based prediction?	*An interpretation of the findings has not been published.*	*Descriptions of individual methods are published.*
MCI-NI[5]	What is the best current strategy for prediction based on structural MRI features of early diagnosis of Alzheimer's disease, differential diagnosis of mild cognitive impairment and conversion into Alzheimer's disease?	*Articles on individual algorithms were published,[a] but an overall interpretation of the findings has not been published.*	The method that achieved best accuracy used an ensemble of random forest classifiers and morphological MRI features.
TADPOLE (Marinescu et al., 2021)	What are the data, features and approaches that are the most predictive of future progression of subjects at risk of Alzheimer's disease to aid identification of patients for inclusion in clinical trials who are in early stages of disease and are likely to progress over the short-to-medium term (1–5 years)?	ADAS-Cog13 scores were more difficult to forecast than clinical diagnosis or ventricle volume. No single method performed well on all prediction tasks, while ensemble methods yielded consistently strong results. Diffusion MRI and cerebrospinal fluid biomarkers are most associated with high prediction performance of clinical diagnosis.	The method that achieved the overall best rank used a gradient boosting regression method based on a combination cognitive, clinical and imaging data.
PAC (Fisch et al., 2021)	What is the best current strategy for accurate prediction of age from brain MRI in an healthy population and what is the best strategy while avoiding bias of age estimation towards the mean age of the training dataset?	Deep learning models performed better than classic machine learning algorithms.	An ensemble of lightweight convolutional neural networks pretrained on UK Biobank data achieved the best performance (MAE=2.90 years, which slightly increased after bias correction to MAE=2.95 years.)

[a] sciencedirect.com/journal/journal-of-neuroscience-methods/vol/302.

prediction or monitoring in dementia. Next to conclusions on methodology, unanswered questions were identified (e.g. limited performance in predicting cognitive scores) and some advice for use in practice was given (e.g. use a direct measure of ventricular volume change to assess atrophy in clinical trials). In general, challenges to date in dementia are all insight challenges; their results are not directly applicable in clinical practice but they are rather pushing the state-of-the-art towards answering clinical questions.

Most challenges attempted to distinguish the best performing submissions using statistical significance testing. However, the considerable challenge of determining the practical significance of such statistical differences was generally not tackled (Mendrik and Aylward, 2019). For example, according to the non-parametric test in TADPOLE, the third-ranked algorithm (multiclass AUC = 0.907) performed significantly worse than winning algorithm (multiclass AUC = 0.931). It could be argued that such a difference in performance is not clinically relevant.

Indeed, other factors might be more important in such cases, such as the ease of applying the algorithm and the possibilities for understanding its outputs (Maier-Hein et al., 2018).

Confounding may also impact the results of grand challenges. Awareness of possible confounding is essential to ensure that actual clinical outcomes rather than confounders are modeled. Age is for example a well-known confounder of dementia that should be taken account of (Dukart et al., 2011). This also holds for biases more generally, the presence of biases such as scanner type (Kruggel et al., 2010) and comorbidities (Ramirez et al., 2016) in challenge data may impact challenge results. For example, despite age-matched classes in the CADDementia test set, challenge outcomes may have been influenced by age as test subjects were younger than the subjects whose data was used by most algorithms for training.

Key to clinical impact is that the performance of the algorithms reported by the challenges generalizes to real-world clinical data. The dominance of ADNI data in the challenges may limit this generalizability. We observed that all challenges on clinical status and prediction used ADNI data (2 for training; 2 for training and testing). While the challenges using ADNI for testing took measures to limit the advantage of researchers being more familiar to the data, the challenge results may still be biased towards its specific data characteristics. Data from the ADNI plays a major role in the research field of machine learning in Alzheimer's disease and is used for optimization and validation in 60–90% of the published articles (Grueso and Viejo-Sobera, 2021; Rathore et al., 2017). Therefore, state-of-the-art performance in ADNI is likely to be an overestimation of performance in clinical practice. Grand challenges would be an ideal venue for assessing the generalizability of algorithm performance beyond ADNI to unseen data of other cohorts, ideally real-world clinical data. This was done by the two challenges using ADNI for training resulting in insight in the performance on external test data. However, future challenges and studies should provide insight into which dataset and algorithm characteristics are key to the generalizability of performance. Since in clinical practice there is generally limited training data available that represents new patients, there would be high relevance for future grand challenges to evaluate not only the performance drop but also factors influencing such generalization.

The code of algorithms that participated in challenges are mostly not available to the community for future research. While some individual methods are available through participants' public repositories, none of challenges released algorithms. For the TADPOLE challenge, algorithms were collected post-hoc, resulting in a collection of the code of seven algorithms in a central repository.[9] In general, availability of the algorithms, especially those that performed well, would increase the challenge's impact on further research and eventually clinical practice.

6. Recommendations for future challenges

Most of the methods evaluated by the challenges are not clinically used, which indicates that the field is some way off from making real clinical impact. Based on this review, we identified several key aspects relating to clinical impact in dementia that are not or only partly addressed by current challenges but should, in our opinion, be a primary focus in developing future challenges to increase their contribution to clinical science.

First, regarding clinical impact, it is key that the tasks being addressed by the algorithms contribute to solving questions relevant to clinicians and also that the algorithms and their output can be understood and used by clinicians. Therefore, future challenges should aim to maximize clinical relevance of the challenge findings for example by maximizing engagement of clinicians in the challenge design. In addition, future challenge should strive for a challenge design that includes

evaluation of the understandability of the output by clinicians and evaluation of the interpretability of the algorithms, e.g. using machine learning interpretability techniques (Dyrba et al., 2021).

Second, it is key to consider clinical impact beyond Alzheimer's disease and ADNI. Challenges with a wider focus could place the application of algorithms in a broader clinical context by assessing generalizability of performance to real-life applications using clinical or population-based data or by assessing clinical questions such as differential diagnosis, prediction or monitoring in other diseases underlying dementia, e.g. frontotemporal dementia or dementia with Lewy bodies.

Third, extended statistical evaluation of factors related to higher performance -such as data types, features or models- will contribute to evidence-based development of future methods increasing performance and impact on clinical implementation (Roß et al., 2021). However, such a statistical analysis requires a test set of sufficient sample size and a sufficient number of challenge submissions.

Finally, a key factor complicating the organization of challenges with either large sample sizes or with clinical questions beyond Alzheimer's disease and ADNI is the collection of clinical test data that can be shared with participants. Developments in research software and infrastructure may provide a solution: sharing algorithms rather than the data. While already done by some challenges,[10,11] dementia challenges have not yet used this approach. Such an approach would only share a small training dataset, but would keep test data private. Wrapping an algorithm in a container (e.g. Docker,[12] Singularity (Kurtzer et al., 2017)) and applying the algorithms locally to the data (at one site or multiple sites in a federated approach) enables challenges to benchmark algorithms on large sets of data that cannot leave their respective institutes. This could improve both data set size and data set diversity. Such an approach could be also used for enabling training on larger datasets (i.e. federated learning). The nature of dementia neuroimaging data needing a lot of data-preprocessing may make this more difficult. Platforms such as grand-challenge.org provide an infrastructure that can be used by challenges without having to implement a custom evaluation system.

Future challenges should strive to have real impact on dementia science by addressing these recommendations. In short, tomorrow's ideal dementia challenge will: (1) encourage explainability and interpretability; (2) address key clinical questions such as early detection, non-AD dementia, and differential diagnosis; (3) include a thorough statistical evaluation to map out the performance landscape; and (4) strive for unbiased generalizability such as by using privacy-preserving technologies.

7. Conclusion

We evaluated grand challenges benchmarking algorithms in the field of screening, clinical status, prediction and monitoring of dementia based on neuroimaging and additional data. We used the frameworks of Maier-Hein et al. (2018) and Mendrik and Aylward (2019) to identify topics for this evaluation.

Our main research question was how the grand challenges in dementia strengthened and complemented each other. Although the number of challenges was small ($n = 7$), they had little overlap: they addressed different clinical questions, had unique tasks and made different choices regarding data, metrics and methodology for ranking. Methods used for avoiding cheating, uncertainty estimation and statistical testing were similar. Together the challenges provide valuable insight into the state-of-the-art and identified where key limitations currently exist. In general, winning algorithms made an effort regarding data pre-processing and combined a wide range of input features. Whereas complementarity is a strength providing insight on a broad range of questions, it also limits the validation of results between challenges.

---

[9] tadpole-share.github.io.

[10] mrbrains18.isi.uu.nl.

[11] node21.grand-challenge.org.

[12] www.docker.com.

As dementia continues to become increasingly important for 21st-century medicine, we are encouraged by the lessons learned from the past ten years of machine learning and neuroimaging grand challenges. This review of current challenge frameworks in dementia showed that challenges have been highly complementary and provided new insights. Future challenges could increase impact by evaluation of interpretability and understandability by clinicians, by more detailed evaluation of factors related to performance, and by addressing the performance of algorithms for a wider range of clinical questions and for a larger variety of test data. We are excited by the prospects for the next ten years and beyond.

Credit authorship contribution statement

Esther E. Bron: Conceptualization, Methodology, Validation, Formal analysis, Investigation, Writing – original draft, Supervision, Project administration, Funding acquisition. **Stefan Klein:** Conceptualization, Writing – review & editing. **Annika Reinke:** Writing – review & editing. **Janne M. Papma:** Writing – review & editing. **Lena Maier-Hein:** Writing – review & editing. **Daniel C. Alexander:** Conceptualization, Funding acquisition, Writing – review & editing. **Neil P. Oxtoby:** Conceptualization, Validation, Writing – review & editing.

Acknowledgments

Supplementary material

Supplementary material associated with this article can be found, in the online version, at 10.1016/j.neuroimage.2022.119083.

References

Albert, M.S., DeKosky, S.T., Dickson, D., Dubois, B., Feldman, H.H., Fox, N.C., Gamst, A., Holtzman, D.M., Jagust, W.J., Petersen, R.C., Snyder, P.J., Carrillo, M.C., Thies, B., Phelps, C.H., 2011. The diagnosis of mild cognitive impairment due to Alzheimer's disease: recommendations from the national institute on aging and Alzheimer's association workgroup. Alzheimer's Dement. 7, 270–279. doi:10.1016/j.jalz.2011.03.008.

Allen, G.I., Amoroso, N., Anghel, C., Balagurusamy, V., Bare, C.J., Beaton, D., Bellotti, R., Bennett, D.A., Boehme, K., Boutros, P.C., 2016. Crowdsourced estimation of cognitive decline and resilience in Alzheimer's disease. Alzheimer's Dement. 12, 645–653. doi:10.1016/j.jalz.2016.02.006.

Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronne, R., Faouzi, J., Koval, I., Louis, M., Couronné, R., Thibeau-Sutre, E., Wen, J., Wild, A., Burgos, N., Dormont, D., Colliot, O., Durrleman, S., 2021. Predicting the progression of mild cognitive impairment using machine learning: a systematic and quantitative review. Med. Image Anal. 67, 101848. doi:10.1016/j.media.2020.101848.

Association, A., 2020. 2020 Alzheimer's disease facts and figures. Alzheimer's Dement. 16, 391–460. doi:10.1002/alz.12068.

Bouthillier, X., Delaunay, P., Bronzi, M., Trofimov, A., Nichyporuk, B., Szeto, J., Sepah, N., Raff, E., Madan, K., Voleti, V., Kahou, S. E., Michalski, V., Serdyuk, D., Arbel, T., Pal, C., Varoquaux, G., Vincent, P., 2021. Accounting for variance in machine learning benchmarks. arXiv:2103.03098

Bron, E.E., Smits, M., van der Flier, W.M., Vrenken, H., Barkhof, F., Scheltens, P., Papma, J.M., Steketee, R.M., Méndez Orellana, C., Meijboom, R., Pinto, M., Meireles, J.J.R., Garrett, C., Bastos-Leite, A.A.J., Abdulkadir, A., Ronneberger, O., Amoroso, N., Bellotti, R., Cárdenas-Peña, D., Álvarez-Meza, A.M., Dolph, C., Iftekharuddin, K., Eskildsen, S.F., Coupé, P., Fonov, V.S., Franke, K., Gaser, C., Ledig, C., Guerrero, R., Tong, T., Gray, K., Moradi, E., Tohka, J., Routier, A., Durrleman, S., Sarica, A., Di Fatta, G., Sensi, F., Chincarini, A., Smith, G., Stoyanov, Z.V., Sørensen, L., Nielsen, M., Tangaro, S., Inglese, P., Wachinger, C., Reuter, M., van Swieten, J., Niessen, W.J., Klein, S., 2015. Standardized evaluation of algorithms for computer-aided diagnosis of dementia based on structural MRI: the CADDementia challenge. NeuroImage 111, 562–579. doi:10.1016/j.neuroimage.2015.01.048.

Cash, D.M., Frost, C., Iheme, L.O., Ünay, D., Kandemir, M., Fripp, J., Salvado, O., Bourgeat, P., Reuter, M., Fischl, B., Lorenzi, M., Frisoni, G.B., Pennec, X., Pierson, R.K., Gunter, J.L., Senjem, M.L., Jack, C.R., Guizard, N., Fonov, V.S., Collins, D.L., Modat, M., Cardoso, M.J., Leung, K.K., Wang, H., Das, S.R., Yushkevich, P.a., Malone, I.B., Fox, N.C., Schott, J.M., Ourselin, S., 2015. Assessing atrophy measurement techniques in dementia: results from the MIRIAD atrophy challenge. NeuroImage 123, 149–164. doi:10.1016/j.neuroimage.2015.07.087.

Castiglioni, I., Salvatore, C., Ramirez, J., Górriz Sáez, J.M., 2018. Machine-learning neuroimaging challenge for automated diagnosis of mild cognitive impairment: lessons learnt. J. Neurosci. Methods doi:10.1016/j.jneumeth.2017.12.019.

Cole, J.H., Poudel, R.P., Tsagkrasoulis, D., Caan, M.W., Steves, C., Spector, T.D., Montana, G., 2017. Predicting brain age with deep learning from raw imaging data results in a reliable and heritable biomarker. NeuroImage 163, 115–124. doi:10.1016/j.neuroimage.2017.07.059.

Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehéricy, S., Habert, M.O.O., Chupin, M., Benali, H., Colliot, O., 2011. Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. NeuroImage 56, 766–781. doi:10.1016/j.neuroimage.2010.06.013.

Donnelly-Kehoe, P.A., Pascariello, G.O., Gómez, J.C., 2018. Looking for Alzheimer's disease morphometric signatures using machine learning techniques. J. Neurosci. Methods 302, 24–34. doi:10.1016/j.jneumeth.2017.11.013.

Dukart, J., Schroeter, M.L., Mueller, K., 2011. Age correction in dementia–matching to a healthy brain. PLoS One 6, e22193. doi:10.1371/journal.pone.0022193.

Dyrba, M., Hanzig, M., Altenstein, S., Bader, S., Ballarini, T., Brosseron, F., Buerger, K., Cantré, D., Dechent, P., Dobisch, L., Düzel, E., Ewers, M., Fliessbach, K., Glanz, W., Haynes, J.D., Heneka, M.T., Janowitz, D., Keles, D., Kilimann, I., Laske, C., Maier, F., Metzger, C.D., Munk, M.H., Perneczky, R., Peters, O., Preis, L., Priller, J., Rauchmann, B., Roy, N., Scheffler, K., Schneider, A., Schott, B.H., Spottke, A., Spruth, E.J., Weber, M.A., Ertl-Wagner, B., Wagner, M., Wiltfang, J., Jessen, F., Teipel, S.J., 2021. Improving 3D convolutional neural network comprehensibility via interactive visualization of relevance maps: evaluation in Alzheimer's disease. Alzheimer's Res. Therapy 13, 1–18. doi:10.1186/s13195-021-00924-2.

Falahati, F., Westman, E., Simmons, A., 2014. Multivariate data analysis and machine learning in Alzheimer's disease with a focus on structural magnetic resonance imaging. J Alzheimer Disease 41, 685–708. doi:10.3233/JAD-131923.

Fisch, L., Leenings, R., Winter, N.R., Dannlowski, U., Gaser, C., Cole, J.H., Hahn, T., 2021. Editorial: predicting chronological age from structural neuroimaging: the predictive analytics competition 2019. Front. Psychiatry 12, 2019–2021. doi:10.3389/fpsyt.2021.710932.

Fox, N.C., Ridgway, G.R., Schott, J.M., 2011. Algorithms, atrophy and Alzheimer's disease: cautionary tales for clinical trials. NeuroImage 57, 15–18. doi:10.1016/j.neuroimage.2011.01.077.

Gordon, B.A., Blazey, T.M., Su, Y., Hari-Raj, A., Dincer, A., Flores, S., Christensen, J., McDade, E., Wang, G., Xiong, C., Cairns, N.J., Hassenstab, J., Marcus, D.S., Fagan, A.M., Jack, C.R., Hornbeck, R.C., Paumier, K.L., Ances, B.M., Berman, S.B., Brickman, A.M., Cash, D.M., Chhatwal, J.P., Correia, S., Förster, S., Fox, N.C., Graff-Radford, N.R., la Fougère, C., Levin, J., Masters, C.L., Rossor, M.N., Salloway, S., Saykin, A.J., Schofield, P.R., Thompson, P.M., Weiner, M.M., Holtzman, D.M., Raichle, M.E., Morris, J.C., Bateman, R.J., Benzinger, T.L., 2018. Spatial patterns of neuroimaging biomarker change in individuals from families with autosomal dominant Alzheimer's disease: a longitudinal study. Lancet Neurol. 17, 241–250. doi:10.1016/S1474-4422(18)30028-0.

Grueso, S., Viejo-Sobera, R., 2021. Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. Alzheimer's Res. Therapy 13. doi:10.1186/s13195-021-00900-w.

Klein, A., Andersson, J., Ardekani, B.A., Ashburner, J., Avants, B., Chiang, M.C., Christensen, G.E., Collins, D.L., Gee, J., Hellier, P., Song, J.H., Jenkinson, M., Lepage, C., Rueckert, D., Thompson, P., Vercauteren, T., Woods, R.P., Mann, J.J., Parsey, R.V., 2009. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. NeuroImage 46, 786–802. doi:10.1016/j.neuroimage.2008.12.037.

Kruggel, F., Turner, J., Muftuler, L.T., 2010. Impact of scanner hardware and imaging protocol on image quality and compartment volume precision in the ADNI cohort. NeuroImage 49, 2123–2133. doi:10.1016/j.neuroimage.2009.11.006.

Kurtzer, G.M., Sochat, V., Bauer, M.W., 2017. Singularity: scientific containers for mobility of compute. PLoS One 12, 1–20. doi:10.1371/journal.pone.0177459.

Licher, S., Leening, M.J., Yilmaz, P., Wolters, F.J., Heeringa, J., Bindels, P.J., Vernooij, M.W., Stephan, B.C., Steyerberg, E.W., Ikram, M.K., Ikram, M.A., 2019. Development and validation of a dementia risk prediction model in the general population: an analysis of three longitudinal studies. Am. J. Psychiatry 176, 543–551. doi:10.1176/appi.ajp.2018.18050566.

Luz, S., Haider, F., de la Fuente, S., Fromm, D., MacWhinney, B., 2020. Alzheimer's dementia recognition through spontaneous speech: The ADReSS challenge. In: Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH, 2020-Octob, pp. 2172–2176. doi:10.21437/Interspeech.2020-2571.

Maier-Hein, L., Eisenmann, M., Reinke, A., Onogur, S., Stankovic, M., Scholz, P., Arbel, T., Bogunovic, H., Bradley, A.P., Carass, A., Feldmann, C., Frangi, A.F., Full, P.M.,

---

[13] demondementia.com.

van Ginneken, B., Hanbury, A., Honauer, K., Kozubek, M., Landman, B.A., März, K., Maier, O., Maier-Hein, K., Menze, B.H., Müller, H., Neher, P.F., Niessen, W., Rajpoot, N., Sharp, G.C., Sirinukunwattana, K., Speidel, S., Stock, C., Stoyanov, D., Taha, A.A., van der Sommen, F., Wang, C.W., Weber, M.A., Zheng, G., Jannin, P., Kopp-Schneider, A., 2018. Why rankings of biomedical image analysis competitions should be interpre ted with care. Nat. Commun. 9, 5217. arXiv:1806.02051.

Malone, I.B., Cash, D., Ridgway, G.R., MacManus, D.G., Ourselin, S., Fox, N.C., Schott, J.M., 2013. MIRIAD-public release of a multiple time point Alzheimer's MR imaging dataset. NeuroImage 70, 33–36. doi:10.1016/j.neuroimage.2012.12.044.

Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Eshaghi, A., Toni, T., Salaterski, M., Lunina, V., Ansart, M., Durrleman, S., Lu, P., Iddi, S., Li, D., Thompson, W.K., Donohue, M.C., Nahon, A., Levy, Y., Halbersberg, D., Cohen, M., Liao, H., Li, T., Yu, K., Zhu, H., Tamez-Peña, J.G., Ismail, A., Wood, T., Bravo, H.C., Nguyen, M., Sun, N., Feng, J., Thomas Yeo, B.T., Chen, G., Qi, K., Chen, S., Qiu, D., Buciuman, I., Kelner, A., Pop, R., Rimocea, D., Ghazi, M.M., Nielsen, M., Ourselin, S., Sørensen, L., Venkatraghavan, V., Liu, K., Rabe, C., Manser, P., Hill, S.M., Howlett, J., Huang, Z., Kiddle, S., Mukherjee, S., Rouanet, A., Taschler, B., Tom, B.D., White, S.R., Faux, N., Sedai, S., de Velasco Oriol, J., Clemente, E.E., Estrada, K., Aksman, L., Altmann, A., Stonnington, C.M., Wang, Y., Wu, J., Devadas, V., Fourrier, C., Raket, L.L., Sotiras, A., Erus, G., Doshi, J., Davatzikos, C., Vogel, J., Doyle, A., Tam, A., Diaz-Papkovich, A., Jammeh, E., Koval, I., Moore, P., Lyons, T.J., Gallacher, J., Tohka, J., Ciszek, R., Jedynak, B., Pandya, K., Bilgel, M., Engels, W., Cole, J., Golland, P., Klein, S., Alexander, D.C., 2021. The Alzheimer's disease prediction of longitudinal evolution (TADPOLE) challenge: results after 1 year follow-up. Mach. Learn. Biomed. Imaging 1. arXiv:2002.03419.

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., Consortium, t. E., 2018TADPOLE challenge: prediction of longitudinal evolution in Alzheimer's disease. arXiv1–8. arXiv:1805.03909.

McKhann, G.M., Knopman, D.S., Chertkow, H., Hyman, B.T., Jack Jr, C.R., Kawas, C.H., Klunk, W.E., Koroshetz, W.J., Manly, J.J., Mayeux, R., Mohs, R.C., Morris, J.C., Rossor, M.N., Scheltens, P., Carillo, M.C., Thies, B., Weintraub, S., Phelps, C.H., 2011. The diagnosis of dementia due to Alzheimer's disease: recommendations from the national institute on aging-Alzheimer's association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's Dement. 7, 263–269. doi:10.1016/j.jalz.2011.03.005.

Mendrik, A., Vincken, K., Kuijf, H., Breeuwer, M., Bouvy, W., de Bresser, J., Alansary, A., de Bruijne, M., Carass, A., El-Baz, A., Jog, A., Katyal, R., Khan, A., van der Lijn, F., Mahmood, Q., Mukherjee, R., van Opbroek, A., Paneri, S., Pereira, S., Persson, M., Rajchl, M., Sarikaya, D., Smedby, O., Silva, C., Vrooman, H., Vyas, S., Wang, C., Zhao, L., Biessels, G., Viergever, M., 2015. MRBrains challenge: online evaluation framework for brain image segmentation in 3T MRI scans. Comput. Intell. Neurosci. doi:10.1155/2015/813696.

Mendrik, A. M., Aylward, S. R., 2019A framework for challenge design: insight and deployment challenges to address medical image analysis problems. arXiv:1911.08531.

Oxtoby, N.P., Alexander, D.C., 2017. Imaging plus x: multimodal models of neurodegenerative disease. Curr. Opin. Neurol. 30, 371–379. doi:10.1097/WCO.0000000000000460.

Panegyres, P.K., Berry, R., Burchell, J., 2016. Early dementia screening. Diagnostics 6, 1–13. doi:10.3390/diagnostics6010006.

Prince, M., Bryce, R., Ferri, C., 2011. World Alzheimer Report 2011, The Benefits of Early Diagnosis and Intervention. Alzheimer's Disease International.

Ramirez, J., McNeely, A.A., Scott, C.J., Masellis, M., Black, S.E., 2016. White matter hyperintensity burden in elderly cohort studies: the sunnybrook dementia study, Alzheimer's disease neuroimaging initiative, and three-city study. Alzheimer's Dement. 12, 203–210. doi:10.1016/j.jalz.2015.06.1886.

Rathore, S., Habes, M., Aksam Iftikhar, M., Shacklett, A., Davatzikos, C., 2017. A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. NeuroImage 155, 530–548. doi:10.1016/j.neuroimage.2017.03.057.

Roß, T., Bruno, P., Reinke, A., Wiesenfarth, M., Koeppel, L., Full, P. M., Pekdemir, B., Godau, P., Trofimova, D., Isensee, F., Moccia, S., Calimeri, F., Müller-Stich, B. P., Kopp-Schneider, A., Maier-Hein, L., 2021How can we learn (more) from challenges? Astatistical approach to driving future algorithm development. ArXivarXiv:2106.09302

Sabuncu, M.R., Konukoglu, E., 2015. Clinical prediction from structural brain MRI scans: a large-scale empirical study. Neuroinform 13, 31–46. doi:10.1007/s12021-014-9238-1.

Sarica, A., Cerasa, A., Quattrone, A., Calhoun, V., 2018. Editorial on special issue: machine learning on MCI. J. Neurosci. Methods 302, 1–2. doi:10.1016/j.jneumeth.2018.03.011.

Treder, M.S., Shock, J.P., Stein, D.J., Du Plessis, S., Seedat, S., Tsvetanov, K.A., 2021. Correlation constraints for regression models: controlling bias in brain age prediction. Front. Psychiatry 12, 615754. doi:10.3389/fpsyt.2021.615754. www.frontiersin.org

Van Vliet, D., De Vugt, M.E., Bakker, C., Pijnenburg, Y.A., Vernooij-Dassen, M.J., Koopmans, R.T., Verhey, F.R., 2013. Time to diagnosis in young-onset dementia as compared with late-onset dementia. Psychol. Med. 43, 423–432. doi:10.1017/S0033291712001122.

Wang, J., Knol, M.J., Tiulpin, A., Dubost, F., De Bruijne, M., Vernooij, M.W., Adams, H.H., Ikram, M.A., Niessen, W.J., Roshchupkin, G.V., 2019. Gray matter age prediction as a biomarker for risk of dementia. Proc. Natl. Acad. Sci. USA 116, 21213–21218. doi:10.1073/pnas.1902376116.

Wen, J., Thibeau-Sutre, E., Samper-Gonzalez, J., Routier, A., Bottani, S., Durrleman, S., Burgos, N., Colliot, O., 2020. Convolutional neural networks for classification of Alzheimer's disease: overview and reproducible evaluation. Med. Image Anal. 63. doi:10.1016/j.media.2020.101694.