Crowdsourced estimation of cognitive decline and resilience in Alzheimer’s disease


"Department of Statistics and Electrical and Computer Engineering, Rice University, Houston, TX, USA

bDipartimento di Fisica “M. Merlin”, Università degli studi di Bari “A. Moro”, Bari, Italy

cSezione di Bari, Istituto Nazionale di Fisica Nucleare, Bari, Italy

dOntario Institute for Cancer Research, Informatics and Bio-computing Program, MaRS Centre, Toronto, ON, Canada

IBM Computational Biology Center, IBM Research, NY, USA

fSage Bionetworks, Seattle, WA, USA

Data used in preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wp-content/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

http://dx.doi.org/10.1016/j.jalz.2016.02.006
1552-5260 © 2016 The Authors. Published by Elsevier Inc. on behalf of the Alzheimer’s Association. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).
Abstract

Identifying accurate biomarkers of cognitive decline is essential for advancing early diagnosis and prevention therapies in Alzheimer’s disease. The Alzheimer’s disease DREAM Challenge was designed as a computational crowdsourced project to benchmark the current state-of-the-art in predicting cognitive outcomes in Alzheimer’s disease based on high dimensional, publicly available genetic and structural imaging data. This meta-analysis failed to identify a meaningful predictor developed from either data modality, suggesting that alternate approaches should be considered for prediction of cognitive performance.

© 2016 The Authors. Published by Elsevier Inc. on behalf of the Alzheimer’s Association. This is an open access article under the CC BY-NC-ND license (http://creativecommons.org/licenses/by-nc-nd/4.0/).

Keywords:
Alzheimer’s disease; Biomarkers; Crowdsourcing; Big data; Bioinformatics; Cognitive decline; Imaging; Genetics
1. Introduction

The Alzheimer’s disease DREAM challenge (http://dx.doi.org/10.7303/syn2290704) was designed to provide an unbiased assessment of current capabilities for estimation of cognition and prediction of cognitive decline using genetic and imaging data from public data resources using a crowdsourced approach. The ability to predict rate of cognitive decline—both before and after diagnosis—is essential to effective trial design for the development of therapies for Alzheimer’s disease (AD) prevention and treatment. Major collaborative efforts in the field are assessing the association of genetic loci with AD diagnosis and the application of structural imaging for development of early biomarkers of diagnosis, but the utility of these approaches to estimate cognition or predict cognitive decline is not well established. This project was designed under the advisement of a panel of experts in the field to evaluate whether these questions could be meaningfully addressed with current methods given existing public data sources. To ensure that these questions were tested across a broad spectrum of the latest analytical approaches, the study was designed as a crowdsourced, community-based challenge in which participants were invited to address one or more of the following three questions [1]: The prediction of cognitive decline over time based on genetic data [2]. The prediction of resilience to cognitive decline in individuals with elevated amyloid burden based on genetic data [3]. The estimation of cognitive state based on structural magnetic resonance (MR) imaging data.

2. Results

2.1. Study design and data harmonization

To ensure that predictors were detecting true biological variation rather than study-specific technical variation, this project required inclusion of data from multiple study sources. Although genetic and imaging data have been generated within many rich longitudinal cohorts across the field, the procurement and harmonization of these data sets were a nontrivial problem that required solutions to overcome political, ethical, and technical barriers. For example, the generation of whole genome sequencing data across multiple AD cohorts within the NIH-funded AD sequencing project has resulted in a powerful resource for genetic analysis in the field but longitudinal information on cognitive traits have not succeeded in providing robust and replicable models [6–8]. Genetic variation has been demonstrated to influence AD status: rare genetic mutations at several loci are implicated in familial forms of early-onset disease [9], whereas common variation contributes 33% to variance in sporadic AD, and 22 loci have been implicated by large-scale genetic association analyses [10,11]. However, with the exception of the APOE e4 haplotype, there has been little success in transforming these genetic associations into meaningful clinical predictions of cognitive decline. For this purpose, participants were challenged to predict 2-year changes in MMSE scores based on genotypes imputed from SNP array data. Participants trained their algorithms with 767 ADNI samples, and the algorithms’ predictions were evaluated on a test set of 1175 ROS/MAP samples with blinded outcome measures. The algorithm with the best predictive performance at the midpoint of the challenge did not contain any genetic features beyond APOE haplotype. As the goal of this question was to assess genetic contribution to prediction of cognitive decline, this top-ranked algorithm was openly shared across teams as an interim baseline on which to incorporate additional genetic predictors (http://dx.doi.org/10.7303/syn2838779). Eighteen teams submitted final predictions. Most methods performed significantly better than a permutation-based random model prediction (Fig. 2A). A cluster of six methods performed significantly better than the others (including the interim baseline model) but were...
statistically indistinguishable among themselves (Fig. 2D).

Of these, the prediction with the best overall score (team GuanLab_umich from the University of Michigan) achieved a Pearson correlation of 0.382 and a Spearman correlation of 0.433 (the overall score was a rank-based combination of these two measures of performance; see online Supplement and Supplementary Methods: http://dx.doi.org/10.7303/syn3383106). However, no significant contribution of genetics beyond APOE haplotype to predictive performance was observed across any of the submissions. Given the small sample size, no conclusions can be inferred from this analysis regarding the existence of genetic loci associated with cognitive decline. Rather, these observations suggest that predictors of cognitive decline developed based on genetic data will not be useful within the clinical setting.

### 2.3. Genetic prediction of cognitive resilience

The second question challenged participants to identify genetic predictors that could distinguish individuals who exhibit resilience to AD pathology as defined by minimal change in cognitive function despite evidence of amyloid deposition [12,13]. Identification of genetic signatures predictive of cognitive resilience would aid in the elucidation of mechanisms that may confer resilience, providing a powerful tool to help advance AD prevention.

---

**Table 1: Challenge Overview**

<table>
<thead>
<tr>
<th>Challenge question</th>
<th>Training data + outcomes</th>
<th>Test data</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. predict cognition (MMSE)</td>
<td><img src="image1.png" alt="DNA" /> n=757</td>
<td><img src="image2.png" alt="DNA" /> n=1,762</td>
</tr>
<tr>
<td>2. predict resilience (amyloid)</td>
<td><img src="image3.png" alt="ADNI" /> n=628</td>
<td><img src="image4.png" alt="ANM" /> n=182</td>
</tr>
<tr>
<td>3. estimate cognition (MMSE)</td>
<td><img src="image5.png" alt="ADNI" /> n=628</td>
<td><img src="image6.png" alt="ANM" /> n=182</td>
</tr>
</tbody>
</table>

---

**Figure 1:** Challenge overview. The top schematic summarizes the three challenge questions on the left column, the training data in the middle, and the test data on the right, including numbers of subjects. The symbols represent sources of data (demographic, ROS/MAP genetic, and ADNI or ANM brain images and shape information). The bottom panel provides example brain image labels and shape information derived from the Mindboggle software (http://mindboggle.info) provided to the participants for question 3. Anatomic labels for left cortical regions are shown on the left and just a couple of the cortical surface shape measures are shown on the right (travel depth on top and mean curvature below), for both uninflated and inflated surfaces (top and bottom rows, respectively).
Fig. 2. Performance evaluation results. (A), (B), and (C) report the $P$ values (in negative log 10 scale) for intersection union tests investigating which teams performed better than random for questions 1, 2, and 3, respectively. Explicitly, for question 1 (A), we tested the null hypothesis that at least one of the four correlation coefficients (namely Pearson/clinical, Pearson/clinical $1^{\text{genetics}}$, Spearman/clinical, and Spearman/clinical $1^{\text{genetics}}$) is equal to zero, against the alternative that all four correlation coefficients are larger than zero. Adopting a 0.05 significance level, 26 of the 32 submissions were statistically better than random, after Bonferroni multiple testing correction for 32 tests (submissions crossing the black vertical line). For question 2 (B), we tested the null hypothesis that balanced accuracy $\geq 0.5$ or AUC $\geq 0.5$, against the alternative that balanced accuracy $< 0.5$ and AUC $< 0.5$. In this case, no model performed significantly better than random, and, therefore, no best performer was declared. For question 3 (C), we tested the null hypothesis that Pearson’s correlation (COR) or Lin’s concordance correlation coefficient (CCC) are equal to zero, against the alternative that both COR and CCC are larger than zero. Adopting a 0.05 significance level, all 23 submissions were statistically better than random, after Bonferroni correction. For all three questions, the $P$ values were computed from an empirical null distribution based on 10,000 permutations. (D) and (E) report the bootstrapped assessment of ranks for questions 1 and 3, respectively. Samples were resampled with replacement from the original data (true outcome and team’s predictions), and the ranks of the different teams were reassessed in each of 100,000 resamplings. Submissions were sorted according to the median of their bootstrapped average ranking distributions. The black horizontal line represents the posterior odds cutoff from the Bayesian analysis. Teams above the black line are statistically tied to the top-ranked model, according to a posterior odds threshold of 3.
strategies and treatment development. Eleven teams submitted predictions of resilience based on a training set derived from 176 ADNI subjects. Evaluations were made using data derived from 257 individuals from the ROS/MAP data. Despite using the largest such public data set assembled to date, participants were unable to develop algorithms with predictive performances significantly better than random (see Fig. 2B, online Supplement and Supplementary Methods in Synapse: http://dx.doi.org/10.7303/syn3383106). Although it is likely that the study was underpowered due to small sample size and trait heterogeneity, this result suggests that information about cognitive resilience is not easily discoverable from SNP analysis.

2.4. Structural imaging-based estimation of cognition

The third question challenged participants to estimate cognitive state using structural brain image data (Fig. 1, lower panel). Brain imaging has emerged as a powerful method for monitoring neurodegeneration, and there is a great enthusiasm in the field to make use of images for diagnosis and prediction. There have been many attempts in the past to correlate changes in brain shape with disease progression and/or diagnosis, conventionally using measures of volume for a given brain region [14,15]. More detailed shape measures of image features including cortical thickness, curvature, and depth have also been found to be relevant to a variety of neurologic conditions [16]. Participants were challenged to estimate MMSE scores based on structural brain images, or shape information derived from these images. Participants trained algorithms using ADNI data (N = 628) and were evaluated using AddNeuromed data (N = 182) for which they were blind to outcome measures. To engage as many participants as possible from both within and beyond the neuroimaging community, the data were provided both as raw MR images and as tables containing shape measures (volume, thickness, area, curvature, depth, and so forth) for every labeled brain region. Thirteen teams submitted estimates for final evaluation, and all teams performed better than a random model (see online Supplement and Supplementary Methods in Synapse: http://dx.doi.org/10.7303/syn3383106). Three teams performed significantly better than the others (teams GuanLab_umich from the University of Michigan, ADDT from the Karolinska Institute and Pythia from the University of Pennsylvania; Fig. 2C) but were statistically indistinguishable from one another and tied for top average rank (Fig. 2E). The algorithm that generated the best absolute mean combined rank (Team GuanLab_umich) achieved a concordance correlation coefficient of 0.569 and Pearson’s correlation of 0.573 (the overall score was a rank-based combination of these two measures of performance). The most common features that contributed heavily to the MMSE estimates across the algorithms were hippocampal volume and entorhinal thickness, corroborating prior work [17–19]. The top three teams also found that inclusion of shape measures of the entorhinal cortex (volume, curvature, surface area, travel, and geodesic depth) improved overall estimation. Other features that contributed to predictions within the top three teams’ results included volume of inferior lateral ventricle and amygdala (see online Supplement and Supplementary Methods in Synapse: http://dx.doi.org/10.7303/syn3383106). These results validate an established relationship between structural imaging data and cognition. However, the correlative performance of these estimators was low suggesting that their application in the clinical setting may not be sufficient to inform patient care.

3. Discussion

The AD DREAM challenge provided a formalized assessment of the ability to develop meaningful predictions of cognitive performance from public genetic or imaging data using contemporary state-of-the-art predictive algorithms. Predictive performance across all three of the questions was modest, and most methods performed roughly equivalently. Given this uniform performance, we do not expect that the presented results are a failure of current modeling methods. A more likely explanation is that the data used to address these questions were inadequate to support these tasks. We also note that most research teams that participated in this challenge did not have expertise in the field of AD. Although the few teams that did possess this knowledge did not do better than the others, there remains the possibility that performance would have been improved by the inclusion of more domain experts.

3.1. Use of genetic information for cognitive prediction

The modest performance observed in the 3 questions focused on genetic analysis demonstrated that contemporary algorithms were not able to leverage genetic signal to make useful predictions for cognition. These results support the prevailing expectation that genetic variants of moderate to high frequency will not support viable biomarker development in AD [9–11]. Although heritability estimates and linkage studies have demonstrated that there is a considerable estimated genetic contribution to AD onset and progression [11,20,21], evidence both within the AD field and across other complex disease [22] traits has indicated that this overall genetic contribution is the aggregated compilation of a large number of loci with small—independent or epistatic—effects. Historically, this type of signal is difficult to capture in predictive models and unlikely to be useful in a diagnostic setting [23]. Furthermore, cognition is highly influenced by a host of nongenetic factors relating to lifestyle choices and accumulated exposures that were not represented across all these data sets and, in fact, are not fully captured in most cohorts [24–27]. Nongenetic contributions to cognitive performance may themselves provide an important base for successful predictions.
Within the context of genetic analysis, the absence of these factors from models confounds the ability to detect real genetic signal and impacts the ability to accurately model state-specific genetic contributions. As such, future inquiry into the use of genetic testing for prediction of cognitive performance and AD risk assessment may be better served by focusing on the contribution of rare genetic variation. Recently discovered disease-associated rare variants have larger effect sizes than common variants and confer 2- to 5-fold greater risk or protection in carriers relative to the general population [28–30]. Ongoing large-scale sequencing analyses will identify additional associated rare risk variants. In sufficient numbers, the aggregate prevalence would support the development of a genetic diagnostic containing a library of rare variants.

3.2. Use of structural imaging data for cognitive estimation

Although the inexpensive and noninvasive nature of genetic testing make this approach amenable to population-level disease screening, the resource-intensive nature of image-based testing is better positioned for careful evaluation of high-risk individuals. As such, these approaches are needed to provide a higher confidence estimate of cognitive performance. Although a variety of methods developed within the context of this challenge were able to successfully estimate cognition, none of these methods were sufficiently accurate to merit clinical consideration. These observations support previous work in the field [17,19] and highlight the imperfect relationship between brain structure and function. Newer imaging modalities that focus on brain function and/or pathology—such as FDG-PET [31] or tau imaging [32]—may prove more successful for assessing cognitive dysfunction.

3.3. Effective performance of meta-analysis across diverse cohorts

A major consideration for any meta-analysis is the issue of appropriate harmonization of data across disparate sources. Despite leveraging several of the most deeply phenotyped cohorts in the field, this challenge limited analysis to those traits that were in common across cohorts. Although this approach to data harmonization is standard practice for meta-analyses [10], it greatly reduced the depth of the information available for modeling and influenced the selection of cognitive measures for use as prediction outcomes. Because each cohort had performed a battery of study-specific tests, this greatly limited the ability for finer grained assessment across cognitive processes. A more sensible approach for future analyses may be to focus effort on more sophisticated methods to calibrate disparate cognitive phenotypes across cohorts [33]. Another undesirable consequence of the focus on traits measured in common was the inability to incorporate into model development the full spectrum of nongenetic and nonimaging factors that are known to influence cognitive performance [24–27]. This suggests the need for development of different approaches for integrating heterogeneous data and/or assessing replication across cohorts. Alternatively, smaller scale analyses that prioritize phenotypic depth over sample size may afford a more refined view of disease.

In summary, this challenge demonstrated that predictions of cognitive performance developed from genetic or structural imaging data were modest across a diverse set of contemporary modeling methods. Future efforts to identify clinically relevant predictors of cognition will benefit from a focus on alternative data sources and methods that work to incorporate greater phenotypic complexity.

Acknowledgments

European AddNeuroMed study—The AddNeuroMed data are from a public-private partnership supported by EFPIA companies, SMEs and the EU under the FP6 programme. Clinical leads responsible for data collection are Iwona Klozesweska (Lodz), Simon Lovestone (London), Patrizia Meocci (Perugia), Hilkka Soininen (Kuopio), Magda Tsolaki (Thessaloniki), and Bruno Vellas (Toulouse).

**References**


