Title: Prediction of conversion to Alzheimer’s disease using deep survival analysis of MRI images

Authors: Tomonori Nakagawa, Manabu Ishida, Junpei Naito, Atsushi Nagai, Shuhei Yamaguchi, Keiichi Onoda, Alzheimer’s Disease Neuroimaging Initiative.

Supplementary Document
Subjects

**ADNI:** This project was launched in 2003 by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, the Food and Drug Administration, private pharmaceutical companies, and non-profit organisations. The primary goals of the project were to develop standardised imaging techniques and biomarker procedures for AD research and to test whether imaging findings, other biological markers and clinical and neuropsychological assessments could be used in combination to predict the progression of early dementia (Mueller et al., 2005). ADNI is the result of efforts by many investigators from a broad range of academic institutions and private corporations. Subjects for ADNI have been recruited from more than 50 sites across the U.S. and Canada (see www.adni-info.org/ for the latest information).

**AIBL:** This project was launched in 2006 by the Australian Commonwealth Scientific and Industrial Research Organization (CSIRO) and several other leading Australian research organisations. The project was designed to discover biomarkers, cognitive characteristics and health and lifestyle factors that determine the subsequent development of symptomatic AD. AIBL adopted the MRI protocol of ADNI, and neuropsychological tests have been designed to permit comparisons and combinations with the results of ADNI (Ellis et al., 2009). Diagnostic and prognostic datasets have been defined by AIBL and are similar to the ADNI dataset.

**J-ADNI:** The Japanese version of ADNI was launched in 2007 by the Ministry of Health, Labour and Welfare and the New Energy and Industrial Technology Development Organization. J-ADNI protocols are designed to achieve the highest degree of compatibility with ADNI in North America. The MRI data can be downloaded at the National Bioscience Database Centre (NBDC, http://human dbs.biosciencedbc.jp/en/) of the Japan Science and Technology Agency. Similar to AIBL, J-ADNI has performed follow-up studies at 18 and 36 months after the baseline.

We also selected patients and healthy elderly people at our hospital (Shimane set). Among outpatients of the memory clinic, patients who met the clinical criteria of MCI, had undergone MPRAGE imaging and had been followed-
up at least once were selected. The clinical criteria of MCI were based on those developed at a workshop convened by the National Institute on Aging and the Alzheimer’s Association (Albert et al., 2011). We selected participants from our brain dry-dock (check-up) system when selecting age-matched control subjects. The elderly participants (≥65 years) participated in check-ups at least twice and consistently had no objective or subjective memory impairment and no history of neurological or psychiatric disorders. The absence of memory impairment was confirmed using the Mini-Mental State Examination (score cut-off ≥26) and Wechsler memory scale short version (cut-off ≥−1 standard deviation, adjusted for age and years of education). Although a determinate diagnosis as a cognitively healthy elderly person requires a clinical dementia rating, the NCs in Shimane set did not undergo this examination. Accordingly, this set might include individuals with MCI or who converted to MCI during the follow-up period. However, the group did not include any possible patients with AD because all were screened to ensure a lack of objective memory impairment. Moreover, patients with dementia should have been identified by a trained neurologist during the check-up interviews. Finally, 57 patients with MCI and 43 healthy elderly were registered as the test set for this study.

The subjects in the current study were selected from a group diagnosed as having MCI and NC at baseline. This selection was also based on the availability of MPRAGE images. The follow-up interval was estimated as the difference between the date of the baseline MRI measurement and the date of diagnosis at follow-up.

The conversion from MCI or NC to AD was defined as an event, whereas conversion from NC to MCI was not considered as an event in the analysis. In the case of conversion from NC to AD via MCI, the timing of the conversion from MCI to AD was treated as an event.

**Structural image pre-processing**

A computational anatomy toolbox for statistical parametric mapping (CAT12) uses two noise reduction methods to perform data processing and tissue segmentation. The first method involves a spatial-adaptive non-local means-
denoising filter during the pre-processing step. The second method involves a classical Markov random field approach, in which spatial information from adjacent voxels is included in the estimation of segmentation. The segmentation step is based on a local adaptive segmentation and adaptive maximum a posterior technique, without requiring a priori information about the tissue probabilities. Spatial normalisation was based on DARTEL (Ashburner, 2007) and Geodesic Shooting (Ashburner and Friston, 2011).

The noise contrast ratio measures the local standard deviation in optimised white matter segments and is scaled using the minimum tissue contrast. The NCR value of the images in this study was 0.13 ± 0.03, which corresponds to a good grade (B) in CAT12. Images with a sufficient grade of D or less (NCR >0.35) were not included in our analysis. Other images included in the deep survival analysis were those that received a good grade of B or a satisfactory grade of C.

In a voxel-level analysis, different magnetic field strengths yield distinct regional sensitivity patterns to morphometric change (Tardif et al., 2010), which affect the estimation of AD-related atrophy (Marchewka et al., 2013). In contrast, studies of patients with epilepsy that are based on atlas-based analyses are less sensitive to the specific effect of the magnetic field strength (Briellmann et al., 2001; Scorzin et al., 2008). These latter studies have reported agreement between the volume estimates of age- or disease-dependent changes at different field strengths (Briellmann et al., 2001; Goodro et al., 2012). Accordingly, an atlas-based analysis appears to reduce the effects of differences in magnetic field strengths. Therefore, we used the mean values of the atlas-based analysis as the inputs in this study, which included data obtained using both 1.5- and 3.0-T MRI systems.

**Model evaluation**

A concordance index was used to evaluate the performances of the models. This index is a standard measure used in survival analyses to estimate the efficiency of a model for ranking survival times. The index calculates the probability that a model will correctly rank the event times of pairs of causes. The concordance index can be considered a generalisation of the area under the curve
AD prediction via deep survival analysis

(AUC) in a receiver operating characteristic (ROC) analysis, which is capable of handling censored data. Here, let $T_i$ denote the interval for the individual, $i$, and let $E_i$ be the conversion, censored or uncensored. The range of $(T_1, E_1), \ldots, (T_n, E_n)$ represents all the events in our dataset. By considering all possible pairs of $(T_i, E_i), (T_j, E_j)$ for $i \leq j$, the concordance index is then calculated by considering the number of pairs correctly ordered by the model, divided by the total number of admissible pairs. A pair cannot be ordered if both events are censored or if the earlier time in the pair is censored. A tied pair is counted as half corrected. A concordance index equal to 1 indicates a perfect prediction, whereas a concordance index equal to 0.5 indicates a random prediction. The concordance index was averaged over time (1–10 years).

**Contribution of each region**

Deep learning models are like black boxes in the sense that it is not obvious how and why the models arrive at particular predictions. This feature was previously considered seriously unfavourable, as it precluded verification, interpretation and understanding the reasoning of the models by a human expert. However, new methods have recently been proposed to understand the predictions of deep learning models (Bach et al., 2015; Kindermans et al., 2017; Montavon et al., 2017). These techniques are available to the public as the "iNNvestigate neural networks!" toolbox (github.com/albermax/inninvestigate) (Alber et al., 2018). We used a deep Taylor decomposition (Montavon et al., 2017), which decomposes a model prediction into the contributions of its input elements, to assess the importance of single ROIs in the prediction of AD conversion. A deep Taylor decomposition efficiently utilises the structure of the model by backpropagating explanations from the output layer and exploits the property in which a function learned by a deep network is decomposed into a set of simpler subfunctions. Based on this idea, the deep Taylor decomposition calculates a relevance score that indicates to what extent the element (i.e., each ROI) contributes to an explanation of the prediction model for an input. The relevance of each element can be stored in the same dimensions and can be visualised as an image. We computed and averaged the relevance scores for the ROI data of each individual. The ROI-level
significance was assessed using a permutation test that created a null distribution by repeating the prediction with a randomised conversion label and interval of 100 times.

**Additional analysis**

ADNI includes amyloid PET measurement. We extracted subjects who had a standardised uptake value (SUV) relative to cerebellar reference on [18F]-AV45 PET data at baseline, and conducted sub-group analysis for the subjects (N = 555). We added the SUV as an optional input feature. Cox and DeepHit models were not available for this dataset, because the number of subjects was too small. The number of independent variables for the Cox model was relatively larger than that of subjects. Further, the DeepHit model could not estimate the conversion probability for a longer range (about 5 year over) due to the small sample size. Therefore, we showed the results of our deep survival model.

**Supplemental Table 1. Concordance index of the deep survival model**

<table>
<thead>
<tr>
<th>Atlas</th>
<th>Input</th>
<th>Concordance index for MCI and HC</th>
<th>Concordance index for MCI</th>
</tr>
</thead>
<tbody>
<tr>
<td>AAL</td>
<td>GMV</td>
<td>0.766 (0.751–0.781)</td>
<td>0.720 (0.712–0.738)</td>
</tr>
<tr>
<td>AAL</td>
<td>GMV + Age</td>
<td>0.773 (0.759–0.788)</td>
<td>0.720 (0.701–0.740)</td>
</tr>
<tr>
<td>AAL</td>
<td>GMV + Age + PET</td>
<td>0.822 (0.809–0.835)*</td>
<td>0.775 (0.760–0.790)*</td>
</tr>
<tr>
<td>AAL</td>
<td>GMV + Age + MMSE</td>
<td>0.780 (0.767–0.793)</td>
<td>0.729 (0.712–0.746)</td>
</tr>
<tr>
<td>AAL</td>
<td>GMV + Age + PET + MMSE</td>
<td>0.804 (0.790–0.819)</td>
<td>0.769 (0.753–0.785)*</td>
</tr>
<tr>
<td>BNA</td>
<td>GMV</td>
<td>0.788 (0.773–0.804)</td>
<td>0.744 (0.727–0.761)</td>
</tr>
<tr>
<td>BNA</td>
<td>GMV + Age</td>
<td>0.805 (0.789–0.821)</td>
<td>0.749 (0.732–0.766)</td>
</tr>
<tr>
<td>BNA</td>
<td>GMV + Age + PET</td>
<td>0.811 (0.796–0.826)</td>
<td>0.787 (0.769–0.806)*</td>
</tr>
<tr>
<td>BNA</td>
<td>GMV + Age + MMSE</td>
<td>0.798 (0.784–0.813)</td>
<td>0.751 (0.733–0.769)</td>
</tr>
<tr>
<td>BNA</td>
<td>GMV + Age + PET + MMSE</td>
<td>0.831 (0.819–0.842)*</td>
<td>0.775 (0.757–0.793)</td>
</tr>
</tbody>
</table>

* indicates a significantly higher concordance index for the model with PET.
relative to the model without PET.

**Additional discussion**

Our study revealed that the prediction performance did not depend on the selection of brain atlas used to distinguish the features. In contrast, a previous study used the BNA and AAL to classify individuals with MCI and normal cognition via SVM, using the GMVs based on different atlases as features (Long et al., 2018). Those authors observed a higher level of accuracy with the BNA than with the AAL atlas. Therefore, it remains unclear which atlas is superior for the classification and prediction of subjects with cognitive impairment. The AAL divides the human brain into 116 regions according to the gyrus and sulcus of a single-subject brain (Tzourio-Mazoyer et al., 2002), and does not precisely correspond with the functional division. The BNA was constructed based on the anatomical and functional connectivity architecture, which were estimated using diffusion tensor imaging and resting-state functional MRI data (Fan et al., 2016). Future neuroimaging studies should more closely evaluate the choice of brain parcellation in atlas-based studies.

The GMVs of the ROIs were used as the classification feature in this study. Other anatomical measurements, including the thickness, three-dimensional (3D) shape and texture of the hippocampus might yield improved predictions. For example, Eskildsen et al. (2013) reported that the thickness of the entorhinal cortex changed 3 years prior to AD conversion, and was followed by hippocampal changes (Eskildsen et al., 2013). When using these changes, the authors achieved a predictive accuracy of 0.76. Costafreda et al. (2011) used an automated procedure to extract the 3D hippocampal shape morphology (Costafreda et al., 2011). Their study, which relied only on hippocampal changes, achieved a predictive accuracy of 0.80, which was comparable or superior to other studies using a multi-region approach. In addition, Sørensen et al. (2016) evaluated the capacity of the hippocampal texture and reported that this variable yielded a higher predictive accuracy than the GMV when predicting the MCI to AD conversion within 2 years.
AD prediction via deep survival analysis (Sørensen et al., 2016). This evidence indicates that the combined use of image indices corresponding to multiple structures might improve the predictive accuracy of a deep survival analysis.

Combination with other imaging techniques is also useful for predicting the conversion to AD. We found that quantification of amyloid beta load improved the prediction performance of AD conversion. A previous study using a Cox model reported that patients with MCI with amyloid positivity were more likely to progress to AD than those with MCI with amyloid negativity, and the risk reached a ceiling at higher values of brain amyloid beta load (Jack et al., 2010b). However, there are few studies that investigated amyloid PET with MCI apart from the ones by Wang et al. (2016) and Ben Bouallègue et al. (2019). Many combined studies using MRI and other modalities have used fluorodeoxyglucose positron emission tomography (FDG-PET) (see Liu et al., 2018), which measures the metabolic activity in the brain. FDG-PET has been reported to capture changes related to AD that happen prior to structural changes (Jack et al., 2010a). Lu et al. (2018) proposed a deep-learning framework using FDG-PET images to identify subjects at the MCI stage with presymptomatic AD and differentiate them from other subjects with MCI (non-AD/non-progressive) (Lu et al., 2018). They obtained an 82.51% classification accuracy by using FDG-PET measures. The reliability of their evidence was high because of the sizeable FDG-PET database of 1,051 subjects.

References


