MRI field strength predicts Alzheimer's disease: a case example of bias in the ADNI data set

Elina Thibeau-Sutre, Baptiste Couvy-Duchesne, Didier Dormont, Olivier Colliot, Ninon Burgos

MRI FIELD STRENGTH PREDICTS ALZHEIMER'S DISEASE: A CASE EXAMPLE OF BIAS IN THE ADNI DATA SET

Elina Thibeau-Sutre[1] *Baptiste Couvy-Duchesne*[1,2] *Didier Dormont*[1,3]
Olivier Colliot[1] *Ninon Burgos*[1]

[1]Sorbonne Université, Institut du Cerveau - Paris Brain Institute, Inserm, CNRS, AP-HP, Hôpital Pitié Salpêtrière, Inria, Aramis project-team, Paris, France
[2]Institute for Molecular Biosciences, the University of Queensland, Brisbane, Australia
[3]AP-HP, Hôpital de la Pitié Salpêtrière, Department of Neuroradiology, Paris, France

ABSTRACT

The Alzheimer's Disease Neuroimaging Initiative (ADNI) data set has been extensively used for the prediction of the progression of prodromal patients to Alzheimer's disease dementia. However, the deep learning community is not always aware of the biases that may contaminate neuroimaging data sets, which may lead to flawed results. In this case example, we demonstrated how ignoring the magnetic resonance (MR) field strength can bias performance of deep learning prediction when using MR images as input. Finally, we discussed options to overcome this problem.

Index Terms— deep learning, neuroimaging, Alzheimer, ADNI, bias

1. INTRODUCTION

Alzheimer's disease (AD) is the most common form of dementia worldwide: in 2016, it affected 43.8 millions people [1]. It causes a diversity of symptoms in patients, which substantially deteriorate their living conditions. One of the prodromal symptom of AD is mild cognitive impairment (MCI), although it is difficult to predict whether a patient with MCI will develop AD. Understanding the risk of AD progression would help clinicians organize more relevant clinical trials, for example by selecting patients who are prone to convert rapidly to AD.

This is one of the objectives of the Alzheimer's Disease Neuroimaging Initiative (ADNI), which is the most used database for the study of AD. Indeed, many machine learning studies relied on this data set to differentiate MCI patients who will progress to dementia in a given time (pMCI), from MCI patients who will stay stable during the same period (sMCI) [2]. As deep learning methods have shown a high performance potential for medical image analysis [3], particularly classification for computer-aided diagnosis and prognosis, they became also used for this task.

ADNI is not a homogeneous cohort, being composed of several waves/phases, ADNI-1, ADNI-GO, ADNI-2 and ADNI-3, which used slightly different protocols. In particular, the 1.5 T magnetic resonance imaging (MRI) machines were progressively phased out (replaced by 3 T MRI), and recruitment targeted different MCI groups.

Moreover, several deep learning studies (including our previous work) evaluating the evolution of MCI status mixed the different cohorts of ADNI to create their population [4, 5, 6] or did not mention this issue and used ADNI as a homogeneous data set [7, 8, 9, 10]. Here, we attempted to quantify the bias that field strength could cause in deep learning studies, with two different experiments. First we trained a CNN to predict the MRI field strength and showed it could significantly predict sMCI vs pMCI. In a second part, we reused the networks trained in our previous study [6] to show the existence of bias in some of our previously published results. The main objective of this study is to raise awareness in the community about the necessity to identify and control for known confounders, in order to report robust results.

2. MATERIALS

We included all recruitment phases of ADNI: ADNI-1, ADNI-GO, ADNI-2 and ADNI-3 (data released before January 26, 2021). Some participants may be followed across several phases, then they are not independent. Two diagnosis groups were considered:

- pMCI: sessions of subjects who were diagnosed as MCI, and progressed to AD during the 36 months following the current visit;
- sMCI: sessions of subjects who were diagnosed as MCI, and neither progress nor regress to AD during the 36 months following the current visit.

Table 1 summarizes the demographics, clinical scores, MRI field strength and distribution in ADNI cohorts of the participants. We observe in this table that there is a difference between sMCI and pMCI field strength distributions. The p-value computed with a chi-square test between the classifica-

							ADNI cohorts			
Label	Subjects	Sessions	Age	% Female	MMSE	% 1.5T	1	GO	2	3
sMCI	266	1 105	72.4 (7.3)	40.0%	27.9 (1.7)	33.1%	88	61	114	3
pMCI	328	918	74.4 (7.1)	41.4 %	26.7 (1.9)	61.6%	193	11	111	13

Table 1: Summary of participant demographics, mini-mental state examination (MMSE) score, field strength and number of sessions of ADNI cohorts at baseline. Values are presented as mean (standard deviation).

tion label and the field strength is significant (8.6×10^{-12}). Then an algorithm learning the sMCI vs pMCI classification task may take advantage of the field strength distribution.

3. METHODS

3.1. Field strength classification task

3.1.1. MRI preprocessing

We used the N4ITK method [11] for bias field correction. Then, T1-weighted MR images were linearly registered [12] to the MNI space (ICBM 2009c nonlinear symmetric template) and cropped to remove all rows and columns containing background voxels only. Finally we rescaled intensity between 0 and 1. All data management and preprocessing was carried out using the Clinica software (github.com/aramis-lab/clinica) [13].

3.1.2. Data split

We considered all the ADNI phases (ADNI1, GO, 2 and 3), which we split into training/validation and test sets. Our test set consisted of 100 subjects chosen to be a representative subset (according to age, sex and field strength distributions) of each diagnostic class. We used the rest of the ADNI data set as training/validation set. We trained the models using the training/validation data set. Training and validation sets were generated with a 5-fold cross-validation stratified according to the field strength value (to ensure that the field strength distribution is equivalent in all folds), which resulted in one fold (20%) of the data for validation and the rest for training. As we used longitudinal data, all splits were performed at the participant level to ensure no data leakage between the training, validation or test sets.

3.1.3. CNN architecture & training - 1.5 T vs 3 T

We trained the network to differentiate 1.5 T from 3 T MRI by optimizing the cross-entropy loss during ten epochs. We used the same architecture as in [6]. This CNN consists of five convolutional blocks and three fully-connected layers. Each convolutional block is sequentially made of one convolutional layer, one batch normalization layer, one ReLU and one max pooling layer.

The final model was the one that obtained the highest validation balanced accuracy during training. The balanced accuracy of the model was evaluated at the end of each epoch. Network training and inference was performed with ClinicaDL (github.com/aramis-lab/clinicaDL/) [14].

3.1.4. Evaluation procedure

After training to differentiate 1.5 T from 3 T MRI, the network was applied to two binary classification tasks: 1.5 T vs 3 T and sMCI vs pMCI. We present in Section 4.1 the mean balanced accuracy of the models applied to the test set, followed by the five mean balanced accuracies of each model obtained on a fold of the 5-fold cross-validation between squared brackets.

3.2. Quantifying bias in previously published results

In a previous article [6], we trained networks to differentiate AD patients from cognitively normal (CN) participants and sMCI from pMCI patients on the baseline sessions of the first three phases of ADNI (ADNI-1, GO and 2), using MRI preprocessed with the procedure described in section 3.1.1. We used different types of inputs by extracting sub-parts of the MRI:

- **image** corresponds to the whole 3D MRI (this is the input that was used in the previous section of this study),
- **patch** corresponds to 36 patches of size 50^3 voxels with no overlapping covering the whole MRI,
- **roi** corresponds to two rectangular prisms encompassing the left and right hippocampi.

We then used the following method: the balanced accuracy was evaluated separately for 1.5 T and 3 T images on the test set. If these balanced accuracies are both lower than the original one (on the whole test set) then results are biased towards the field strength. To evaluate the difference between the balanced accuracy on the whole test or only one field strength, we performed a paired t-test on the two series of five folds for each experiment. If the differences in balanced accuracy between the 1.5 T set and the whole set, and the 3 T set and the whole test are significant, and if the balanced accuracies of the 1.5 T and 3 T sets are similar, then we can conclude that the network partly learned the field strength. We considered that the difference between two series was significant if the t-test comparing them resulted in a p-value lower than 0.05.

4. RESULTS

Input	1.5 T	3 T	all	p-values (1)	(2)	(3)
image	0.80 [0.79, 0.87, 0.80, 0.75, 0.81]	0.84 [0.79, 0.82, 0.84, 0.86, 0.90]	0.82 [0.79, 0.84, 0.82, 0.80, 0.85]	0.21	0.21	0.21
roi	0.86 [0.85, 0.86, 0.88, 0.88, 0.86]	0.91 [0.89, 0.91, 0.91, 0.94, 0.92]	0.89 [0.86, 0.88, 0.90, 0.91, 0.89]	< 0.01	< 0.01	< 0.01
patch	0.78 [0.76, 0.76, 0.82, 0.78, 0.76]	0.86 [0.89, 0.86, 0.86, 0.83, 0.83]	0.81 [0.82, 0.81, 0.84, 0.80, 0.79]	< 0.01	< 0.01	< 0.01

Table 2: Comparison of balanced accuracies for the AD vs CN task with deep learning methods obtained on 1.5 T, 3 T and the combination of the two (all). P-values correspond to the following paired t-tests: (1) 1.5 T vs all, (2) 3 T vs all, (3) 1.5 T vs 3 T.

Input	1.5 T	3 T	all	p-values (1)	(2)	(3)
image	0.63 [0.61, 0.73, 0.49, 0.69, 0.66]	0.60 [0.63, 0.62, 0.51, 0.66, 0.60]	0.68 [0.68, 0.71, 0.64, 0.73, 0.67]	0.16	< 0.01	0.27
roi	0.70 [0.67, 0.70, 0.68, 0.72, 0.72]	0.70 [0.74, 0.64, 0.74, 0.70, 0.70]	0.74 [0.75, 0.72, 0.76, 0.74, 0.75]	0.02	0.03	0.82
patch	0.56 [0.50, 0.51, 0.60, 0.59, 0.60]	0.58 [0.68, 0.50, 0.54, 0.61, 0.58]	0.68 [0.71, 0.64, 0.64, 0.71, 0.69]	0.01	< 0.01	0.64

Table 3: Comparison of balanced accuracies of deep learning methods for the sMCI vs pMCI task obtained on 1.5 T, 3 T and the combination of the two (all). P-values correspond to the following paired t-tests: (1) 1.5 T vs all, (2) 3 T vs all, (3) 1.5 T vs 3 T.

4.1. Field strength classification task

To assess whether there is a risk that a network learns the field strength instead of the diagnosis status, we trained CNNs to detect the field strength, i.e. 1.5 T vs 3 T, using the T1-weighted MR images of sMCI and pMCI patients from ADNI.

The CNN perfectly learns to differentiate field strengths in our population by obtaining a balanced accuracy of 0.98 [0.98, 0.96, 0.98, 0.98, 0.98]. Moreover, the direct application of the networks to the sMCI vs pMCI led to a balanced accuracy higher than chance 0.65 [0.65, 0.65, 0.65, 0.64, 0.66] and of similar value as the ones that could be obtained by networks trained on sMCI vs pMCI (0.68 when using the whole image as input, see Table 3). Then we checked whether our previously published results were contaminated by this bias.

4.2. Quantifying bias in published results

We evaluated the presence of bias in our previous work [6]. Results are displayed in Tables 2 and 3. The original value always lies between the values obtained for 1.5 T and 3 T for AD vs CN, and we cannot conclude to the learning of the field strength by the network with the p-values (though we note that for **roi** and **patch** the results are much better on 3 T images than 1.5 T images).

This is not the case for sMCI vs pMCI. Indeed, each time the 1.5 T and 3 T results are not significantly different, but they are both significantly different from the original values (except for **image** where only the 3 T series is significantly different from the original values). Then we observe a significant drop in balanced accuracies of 1.5 T and 3 T compared to the original one for **patch** CNN (12 and 10 percent points). The **image** and **roi** CNNs are also affected by this bias, but not to the same extent, with drops of 5 and 8 percent points for **image** and drops of 4 percent points for **roi**. We guess that the **patch** experiments are more affected than **image** or **roi** ones as in some patches at the edge of the brain no information relevant to the diagnosis can be found, then the only useful information is the field strength.

5. CONCLUSION

This study started from the observation that the sMCI/pMCI status was associated with MRI field strength because of a recruitment bias in ADNI. We showed that CNNs could successfully learn to differentiate 1.5 T from 3 T MRI, and that a field strength predictor would achieve a 65% balanced accuracy in ADNI. We further observed that sMCI/pMCI predictors would learn the data structure, leading to inflated prediction accuracy. Our case example demonstrates how field strength acts as a confounder on sMCI vs pMCI results. We showed that previous results (including a previous publication from our group) reported inflated prediction accuracy of the sMCI vs pMCI task. This could partly explain the low generalizability of the prediction onto other test sets, such as the Australian Imaging, Biomarkers and Lifestyle (AIBL).

Beyond this specific example, bias may be present in other studies or data sets, and may cause an overestimation of the performance of machine learning algorithms. In addition to the MRI field strength, several other confounders have also been flagged in the neuroimaging literature. They include age of the participants, sex, site, MRI machine, body size (e.g. height, weight) or head motion. Recently, a large scale examination has suggested many possible confounders of structural MRI studies [15]. Importantly, the presence and effect of the putative confounding factor are dependent on each data set and trait/disorder of interest, and in some cases several confounding factors can contribute to prediction bias.

To avoid this pitfall we can only recommend future studies to more systematically take into account putative con-

founders. Several approaches may be used, such as the post-hoc ones we implemented here, which consists in evaluating prediction accuracy in subsets of the sample, or evaluating generalizability of the prediction into specific subsets of participants (e.g. into 1.5 T images). Another approach consists in controlling for known confounders when evaluating the prediction accuracy. For example, one may use a generalized linear regression framework, with confounders fitted as covariates. On the other hand, confounders may also be dealt with during the training of algorithms. For example one could over-sample or put more weight on rare samples (here sMCI patients with 1.5 T images and pMCI patients with 3 T images). Finally, even though it is tempting to use as much data as possible, the most reliable solution could be to train networks with images of the same field strength.

6. COMPLIANCE WITH ETHICAL STANDARDS

Ethical approval was not required as confirmed by the license attached with the open access data.

7. ACKNOWLEDGMENTS

The authors have no relevant financial or non-financial interests to disclose. The research leading to these results has received funding from the French government under management of Agence Nationale de la Recherche as part of the "Investissements d'avenir" program, reference ANR-19-P3IA-0001 (PRAIRIE 3IA Institute) and reference ANR-10-IAIHU-06 (Agence Nationale de la Recherche-10-IA Institut Hospitalo-Universitaire-6). BCD is supported by the NHMRC (app1161356).

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

8. REFERENCES

[1] E. Nichols et al., "Global, regional, and national burden of Alzheimer's disease and other dementias, 1990–2016: A systematic analysis for the Global Burden of Disease Study 2016," *The Lancet Neurology*, vol. 18, no. 1, pp. 88–106, 2019.

[2] M. Ansart et al., "Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review," *Medical Image Analysis*, vol. 67, pp. 101848, 2021.

[3] L. Zhang, M. Wang, M. Liu, and D. Zhang, "A Survey on Deep Learning for Neuroimaging-Based Brain Disorder Analysis," *Frontiers in Neuroscience*, vol. 14, pp. 779, 2020.

[4] S. Basaia et al., "Automated classification of Alzheimer's disease and mild cognitive impairment using a single MRI and deep neural networks," *NeuroImage: Clinical*, p. 101645, 2018.

[5] S. El-Sappagh, T. Abuhmed, S. M. Riazul Islam, and K. S. Kwak, "Multimodal multitask deep learning model for Alzheimer's disease progression detection based on time series data," *Neurocomputing*, vol. 412, pp. 197–215, 2020.

[6] J. Wen et al., "Convolutional neural networks for classification of Alzheimer's disease: Overview and reproducible evaluation," *Medical Image Analysis*, vol. 63, pp. 101694, 2020.

[7] G. Lee et al., "Predicting Alzheimer's disease progression using multi-modal deep learning approach," *Scientific Reports*, vol. 9, no. 1, pp. 1952, 2019.

[8] F. Gao et al., "AD-NET: Age-adjust neural network for improved MCI to AD conversion prediction," *NeuroImage: Clinical*, vol. 27, pp. 102290, 2020.

[9] Y. Shmulev and M. Belyaev, "Predicting Conversion of Mild Cognitive Impairments to Alzheimer's Disease and Exploring Impact of Neuroimaging," in *Graphs in Biomedical Image Analysis and Integrating Medical Imaging and Non-Imaging Modalities*, Cham, 2018, vol. 11044, pp. 83–91, Springer International Publishing.

[10] T. Zhang and M. Shi, "Multi-modal neuroimaging feature fusion for diagnosis of Alzheimer's disease," *Journal of Neuroscience Methods*, vol. 341, pp. 108795, 2020.

[11] N. J. Tustison et al., "N4ITK: Improved N3 Bias Correction," *IEEE Transactions on Medical Imaging*, vol. 29, no. 6, pp. 1310–1320, 2010.

[12] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee, "Symmetric Diffeomorphic Image Registration with Cross-Correlation: Evaluating Automated Labeling of Elderly and Neurodegenerative Brain," *Medical image analysis*, vol. 12, no. 1, pp. 26–41, 2008.

[13] A. Routier et al., "Clinica: An Open-Source Software Platform for Reproducible Clinical Neuroscience Studies," *Frontiers in Neuroinformatics*, vol. 15, pp. 689675, 2021.

[14] E. Thibeau-Sutre et al., "ClinicaDL: An open-source deep learning software for reproducible neuroimaging processing," preprint, 2021.

[15] F. Alfaro-Almagro et al., "Confound modelling in UK Biobank brain imaging," *NeuroImage*, vol. 224, pp. 117002, 2021.