Medical Image Analysis 66 (2020) 101714

Contents lists available at ScienceDirect

Medical Image Analysis

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

The reliability of a deep learning model in clinical out-of-distribution MRI data: A multicohort study☆

Gustav Mårtensson[a,*], Daniel Ferreira[a], Tobias Granberg[b,c], Lena Cavallin[b,c], Ketil Oppedal[d,e,f], Alessandro Padovani[g], Irena Rektorova[h], Laura Bonanni[i], Matteo Pardini[j], Milica G Kramberger[k], John-Paul Taylor[l], Jakub Hort[m], Jón Snædal[n], Jaime Kulisevsky[o,p,q,r], Frederic Blanc[s,t], Angelo Antonini[u], Patrizia Mecocci[v], Bruno Vellas[w], Magda Tsolaki[x], Iwona Kłoszewska[y], Hilkka Soininen[z,A], Simon Lovestone[B], Andrew Simmons[C,D,E], Dag Aarsland[d,F], Eric Westman[a,E]

[a] Division of Clinical Geriatrics, Department of Neurobiology, Care Sciences and Society, Karolinska Institutet, Stockholm, Sweden
[b] Department of Clinical Neuroscience, Karolinska Institutet, Stockholm, Sweden
[c] Department of Radiology, Karolinska University Hospital, Stockholm, Sweden
[d] Centre for Age-Related Medicine, Stavanger University Hospital, Stavanger, Norway
[e] Stavanger Medical Imaging Laboratory (SMIL), Department of Radiology, Stavanger University Hospital, Stavanger, Norway
[f] Department of Electrical Engineering and Computer Science, University of Stavanger, Stavanger, Norway
[g] Neurology Unit, Department of Clinical and Experimental Sciences, University of Brescia, Brescia, Italy
[h] 1st Department of Neurology, Medical Faculty, St. Anne's Hospital and CEITEC, Masaryk University, Brno, Czech Republic
[i] Department of Neuroscience Imaging and Clinical Sciences and CESI, University G d'Annunzio of Chieti-Pescara, Chieti, Italy
[j] Department of Neuroscience (DINOGMI), University of Genoa and Neurology Clinics, Polyclinic San Martino Hospital, Genoa, Italy
[k] Department of Neurology, University Medical Centre Ljubljana, Medical faculty, University of Ljubljana, Slovenia
[l] Institute of Neuroscience, Newcastle University, Newcastle upon Tyne, UK
[m] Memory Clinic, Department of Neurology, Charles University, 2nd Faculty of Medicine and Motol University Hospital, Prague, Czech Republic
[n] Landspitali University Hospital, Reykjavik, Iceland
[o] Movement Disorders Unit, Neurology Department, Sant Pau Hospital, Barcelona, Spain
[p] Institut d'Investigacions Biomédiques Sant Pau (IIB-Sant Pau), Barcelona, Spain
[q] Centro de Investigación en Red-Enfermedades Neurodegenerativas (CIBERNED), Barcelona, Spain
[r] Universitat Autónoma de Barcelona (U.A.B.), Barcelona, Spain
[s] Day Hospital of Geriatrics, Memory Resource and Research Centre (CM2R) of Strasbourg, Department of Geriatrics, Hôpitaux Universitaires de Strasbourg, Strasbourg, France
[t] University of Strasbourg and French National Centre for Scientific Research (CNRS), ICube Laboratory and Fédération de Médecine Translationnelle de Strasbourg (FMTS), Team Imagerie Multimodale Intégrative en Santé (IMIS)/ICONE, Strasbourg, France
[u] Department of Neuroscience, University of Padua, Padua & Fondazione Ospedale San Camillo, Venezia, Venice, Italy
[v] Institute of Gerontology and Geriatrics, University of Perugia, Perugia, Italy
[w] UMR INSERM 1027, gerontopole, CHU, University of Toulouse, France
[x] 3rd Department of Neurology, Memory and Dementia Unit, Aristotle University of Thessaloniki, Thessaloniki, Greece
[y] Medical University of Lodz, Lodz, Poland
[z] Institute of Clinical Medicine, Neurology, University of Eastern Finland, Finland
[A] Neurocenter, Neurology, Kuopio University Hospital, Kuopio, Finland
[B] Department of Psychiatry, Warneford Hospital, University of Oxford, Oxford, UK
[C] NIHR Biomedical Research Centre for Mental Health, London, UK
[D] NIHR Biomedical Research Unit for Dementia, London, UK
[E] Department of Neuroimaging, Centre for Neuroimaging Sciences, Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK
[F] Institute of Psychiatry, Psychology and Neuroscience, King's College London, London, UK

☆ Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://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.

* Corresponding author.
E-mail address: gustav.martensson@ki.se (G. Mårtensson).

https://doi.org/10.1016/j.media.2020.101714

ARTICLE INFO

Article history:
Received 7 November 2019
Revised 17 April 2020
Accepted 24 April 2020
Available online 1 May 2020

Keywords:
Neuroimaging
Deep learning
Domain shift
Clinical application

ABSTRACT

Deep learning (DL) methods have in recent years yielded impressive results in medical imaging, with the potential to function as clinical aid to radiologists. However, DL models in medical imaging are often trained on public research cohorts with images acquired with a single scanner or with strict protocol harmonization, which is not representative of a clinical setting. The aim of this study was to investigate how well a DL model performs in unseen clinical datasets–collected with different scanners, protocols and disease populations–and whether more heterogeneous training data improves generalization. In total, 3117 MRI scans of brains from multiple dementia research cohorts and memory clinics, that had been visually rated by a neuroradiologist according to Scheltens' scale of medial temporal atrophy (MTA), were included in this study. By training multiple versions of a convolutional neural network on different subsets of this data to predict MTA ratings, we assessed the impact of including images from a wider distribution during training had on performance in external memory clinic data. Our results showed that our model generalized well to datasets acquired with similar protocols as the training data, but substantially worse in clinical cohorts with visibly different tissue contrasts in the images. This implies that future DL studies investigating performance in out-of-distribution (OOD) MRI data need to assess multiple external cohorts for reliable results. Further, by including data from a wider range of scanners and protocols the performance improved in OOD data, which suggests that more heterogeneous training data makes the model generalize better. To conclude, this is the most comprehensive study to date investigating the domain shift in deep learning on MRI data, and we advocate rigorous evaluation of DL models on clinical data prior to being certified for deployment.

1. Introduction

The use of deep learning (DL) models in neuroimaging has increased rapidly in the last few years, often showing superior diagnostic abilities compared to traditional imaging softwares (see Litjens et al., 2017; Lundervold and Lundervold, 2019 for reviews). This makes DL models promising to use as diagnostic aid to clinicians. However, for a software to function in a clinical setting it should work on images acquired from different scanners, protocol parameters, and of varying image quality–a scenario reflective of most clinical settings today. Fig. 1 shows illustrative examples of the variability in images from some different centers included in this study.

Training a DL model on magnetic resonance imaging (MRI) scans requires a large dataset to obtain good performance. However, (labeled) clinical data is generally difficult (and expensive) to acquire due to strict privacy regulations on medical data. Most researchers are therefore constrained to rely on publicly available neuroimaging datasets, which are typically research cohorts that differ from a clinical setting in several ways: 1) Images are acquired from the same scanner and protocol, or protocols have been harmonized across machines. This is done to reduce image variability and confounding effects, which are problematic also for traditional neuroimaging softwares such as FSL, FreeSurfer and SPM (Guo et al., 2019). 2) Research cohorts often have strict inclusion and exclusion criteria for the individuals enrolled in order to study a particular effect of interest. For instance, to study the progression of patients suffering from Alzheimer's Disease (AD) it may be necessary to exclude comorbidities, such as cerebrovascular pathology or history traumatic brain injury, in order to reduce heterogeneity not relevant to the research question. This is the case of the Alzheimer's Disease Neuroimaging Initiative (ADNI) cohort–the most extensive public neuroimaging dataset in AD and used for training and evaluation in multiple DL studies on AD (Litjens et al., 2017). However, since comorbidities are frequent alongside AD the ADNI cohort is hardly reflective of the heterogeneous AD profiles of patients in the clinics (Boyle et al., 2018; Ferreira et al., 2017). Thus, training a DL model on data from a research cohort may perform worse in a clinical setting due to difficulties generalizing to new scanners/protocols (point 1) and/or a more heterogeneous population (point 2). Investigating the performance in *out-of-distribution data* (OOD data, i.e. images acquired with different scanners/protocols than the ones included in the training set) is an important step in order to investigate clinical applicability of DL models and understanding the challenges that can arise when deploying.

Some previous studies have investigated the clinical applicability of machine learning models, or *domain shift* (training a model on data from one domain and applying it in data from another). A recent paper by De Fauw et al. (2018) trained and applied a deep learning model on a clinical dataset of 3D optical coherence tomography scans, which managed to predict referral decisions with similar performance as experts. However, when applied to images from a new scanning device the performance was poor. Since they used a two-stage model architecture, where the first part segmented the image into different tissue types (making subsequent analysis scanner independent), it was sufficient to retrain only the segmentation network with a (much smaller) dataset from the new device. Klöppel et al. (2015) investigated the performance of a trained SVM-classifier to diagnose dementia in a clinical dataset of a more heterogeneous population. Their models were also fed tissue-segmentation maps, preprocessed using SPM, and found a drop in performance compared to the "clean" training set, as well as lower performance than previous studies had reported (typically cross-validation performance). Zech et al. (2018) explicitly investigated how a convolutional neural network (CNN) trained for pneumonia screening on chest X-rays generalized to new cohorts. They found significantly lower performance in OOD cohorts. Further, they demonstrated that a CNN could accurately classify which hospital an image was acquired at and thus potentially leverage this information to adjust the prediction method due to different disease prevalences in the cohorts. Some recent studies have investigated MRI segmentation performance across centers and again found drops in performance (Albadawy et al., 2018; Kamnitsas et al., 2017; Perone et al., 2019). These analyses were made on a small number of images, as segmented data is typically expensive and time-consuming to label. In contrast to segmented data, visual ratings of atrophy, which still serve as the main tools to quantify neurodegeneration in memory clinics, offer a faster method to annotate brain images that make it feasible to label large datasets ($>$ 1000 images) from multiple clinics. Our group recently proposed *AVRA* (Automatic Visual Ratings

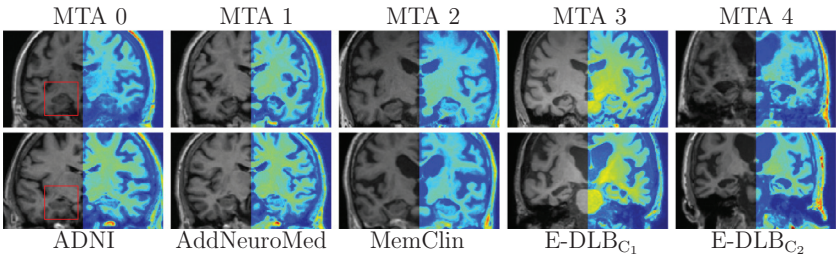

Fig. 1. Two randomly selected images from five different cohorts in the study to illustrate image intensity variability between cohorts, and examples of Scheltens' scale of medial temporal atrophy (MTA) rated by a radiologist. The red boxes show the region of interest for the MTA scale with a progressive worsening in the hippocampus and surrounding regions. The images are normalized to have zero mean and unit variance, with the same intensity color scale for all images. The jet color map on the right-hand part of the images is used to visibly highlight intensity differences between centers. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

of Atrophy), a DL model based on convolutional neural networks (CNN) (Mårtensson et al., 2019). AVRA inputs an unprocessed T_1-weighted MRI image and predicts the ratings of Scheltens' Medial Temporal Atrophy (MTA) scale, commonly used clinically to diagnose dementia (Scheltens et al., 1992) (see Fig. 1 for examples of the MTA scale).

The aim of this study is to systematically investigate the performance of a CNN based model (AVRA) in OOD data from clinical neuroimaging cohorts. We study the impact more heterogeneous training data has on generalization to OOD data by training and evaluating AVRA on images from different combinations of cohorts. Two of these cohorts are research oriented: similar to each other in terms of disease population (AD) and protocol harmonization. The other two datasets consist of clinical data from multiple European sites including individuals of different and mixed types of dementia, not just AD. Additionally, we assess the inter- and intra-scanner variability of AVRA in a systematic test-retest set. To our knowledge this is the largest and most comprehensive study on the generalization of DL models in neuroimaging and MRI data.

2. Material and methods

2.1. MRI data and protocols

The 3117 images analyzed in this study came from five different cohorts described in Table 1, where we also list the reasons for including these datasets in the current study. Full lists of scanners and scanning protocols are provided as Supplementary Data. TheHiveDB was used for data management in this study (Muehlboeck et al., 2014).

Data used in the preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (http://adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of mild cognitive impairment (MCI) and early Alzheimer's disease (AD). In brief, the ADNI dataset is a large, public dataset that has helped advance the field of AD and neuroimaging. However, the strictly harmonized protocols and strict exclusion criteria make ADNI unrepresentative of a clinical setting. Some subjects were scanned multiple times (within a month) in both a 1.5T and a 3T scanner in which case one of the images was selected at random during training for the current study. *AddNeuroMed* is an imaging cohort collected in six sites across Europe with the aim of developing and validating novel biomarkers for AD (Simmons et al., 2011). The MRI images were acquired with protocols designed to be compatible with data from ADNI, and the two cohorts have been successfully combined in previous studies (Falahati et al., 2016; Mårtensson et al., 2018; Westman et al., 2011). AddNeuroMed was an interesting cohort to assess AVRA's reliability in due to having consistent scanning parameters and acquisition methods similar to ADNI. Thus, this dataset represented a research cohort where we expected our DL model to show good performance in when trained on ADNI data. A subset of the images (N =122) of patients diagnosed with AD had been visually rated for MTA. Exclusion criteria for both these studies included no histories of head trauma, neurological or psychiatric disorders (apart from AD), organ failure, or drug/alcohol abuse.

The *MemClin* dataset was used for training also in our previous study detailing AVRA (Mårtensson et al., 2019). MemClin consists of images of AD or frontotemporal lobe dementia collected from the memory clinic at Karolinska Hospital in Huddinge, Sweden. This dataset better resembled a clinical setting with varying scanning parameters and field strengths, while the disease population was not completely representative of patients in a memory clinic. The only exclusion criteria was history of traumatic brain injury. Images and ratings have previously been analyzed in (Ferreira et al., 2018; Lindberg et al., 2009).

The fourth cohort in this study consists of clinical MRI images from the European consortium for Dementia with Lewy Bodies (referred to as *E-DLB* from here on) previously described in (Kramberger et al., 2017; Oppedal et al., 2019). Patients with referrals to memory, movement disorders, psychiatric, geriatric or neurology clinics that had undergone an MRI were selected from 12 sites in Europe. These individuals were diagnosed with Dementia with Lewy Bodies (DLB), AD, Parkinson's Disease with Dementia (PDD), mild cognitive impairment (MCI, due to AD or DLB), or were healthy elderly controls (HC). The images were acquired as part of the clinical routine, and consequently without protocol harmonization, and can thus be considered to reflect a clinical setting well. Exclusion criteria for the E-DLB cohort were having received a recent diagnosis of major somatic illness, history of psychotic or bipolar disorders, acute delirium, or terminal illness.

We also investigated AVRA's rating consistency on MRI images (without lesion filling) of three healthy and nine individuals with Multiple Sclerosis (MS, mean disease duration 7.3 $\pm$ 5.2 years) that were scanned twice with repositioning in three different Siemens scanners (i.e. six scans in total) in a single day. Six of the patients had relapsing-remitting MS, two secondary progressive MS, and one primary progressive MS. This dataset was collected for a previous study (Guo et al., 2019), and we will refer to this small set as the *test-retest* dataset. These individuals were not rated for MTA by a radiologist.

Table 1
An overview of how the cohorts used for training and/or evaluation differ from each other, and the purpose of including them in the present study. The E-DLB cohort (denoted as E-DLB$_{all}$, referring to all images in the cohort) was stratified into different subsets in order to isolate specific features of interest. N_{train}/N_{test} refers to the number of labeled images used during training/evaluation, where some cohorts were split into training and test set. Abbreviations: Deep Learning (DL); Out-of-distribution (OOD) data; Alzheimer's disease (AD); Healthy controls (HC); Frontotemporal lobe dementia (FTLD); Dementia with Lewy Bodies (DLB); Parkinson's disease with dementia (PDD).

Cohort	Scanners/Protocols	Disease population	Purpose of inclusion
ADNI N_{train}=1568 N_{test}=398	Multiple scanners and sites, but strictly harmonized with phantom. Both 1.5T and 3T.	AD spectrum and HC.	Common cohort to train and evaluate DL models in, which we hypothesize should not generalize well.
AddNeuroMed N=122	Harmonized, designed to be compatible with ADNI.	AD patients only.	Assess AVRA in an external research cohort similar to ADNI.
MemClin N_{train}=318 N_{test}=66	Unharmonized, part of clinical routine from a single memory clinic.	Mainly AD spectrum and HC, with 37 FTLD patients.	Large clinical cohort with similar disease population as ADNI and AddNeuroMed.
E-DLB$_{all}$ N=645	Retrospective unharmonized data of varying quality from 12 European sites as part of their clinical routine.	Mainly DLB spectrum, but also HC, AD and PDD.	To assess performance of AVRA in a large, realistic clinical cohort.
E-DLB$_{AD}$ N=193	Same as E-DLB$_{all}$	Only individuals with AD pathology from E-DLB$_{all}$.	To isolate effects of scanners/protocols not seen during training from disease population.
E-DLB$_{\{DLB,PDD\}}$ N={266,97}	Same as E-DLB$_{all}$	Only individuals with DLB or PDD pathology from E-DLB$_{all}$, respectively.	To assess the impact scanners/protocols *and* disease populations not seen during training have on AVRA performance.
E-DLB$_{\{25\%,50\%\}}$ N_{train}={173,312} N_{test}= 333	Same as E-DLB$_{all}$	Randomly selected images with a probability of 25% (or 50%) from all centers in E-DLB$_{all}$.	To assess effect of including training data from test set distribution has on AVRA performance.
E-DLB$_{\{C_1,C_2\}}$ N={101,165}	Both centers have used a single scanner (3T) and protocol.	Only images from center C_1 and C_2 from E-DLB$_{all}$, respectively.	"External validation sets": how would AVRA perform if deployed in two external memory clinics?
E-DLB$_{C_{3-12}}$ N=379	Same as E-DLB$_{all}$	All images in E-DLB$_{all}$ *except* from center C_1 and C_2.	Large clinical cohort with a more heterogeneous disease population than MemClin.
Test-Retest N=72	Three Siemens scanners (two 1.5T, one 3T) with similar protocols but unharmonized.	Young (38 ± 13 years old) MS patients and healthy controls.	Systematic evaluation of the impact scanner variability has on AVRA predictions.

Table 2
Distribution of MTA ratings from a neuroradiologist in the different cohorts, together with sex (female percentage) and age demographics. The lines in bold refers to the statistics of the *whole* cohort, whereas the rows not in boldface text are the subsets used for during training. *N* is the total number of rated images, and since both left and right hemispheres were rated there were 2*N* ratings. *MTA distribution* shows the percentage of each radiologist rating per (sub-)cohort. A small linespace are added between some E-DLB subsets to illustrate the grouping of the subsets where no overlap between training and test sets occur.

Cohort	N	MTA distribution, (%)					Females	Age
Subset		0	1	2	3	4	(%)	(mean ± std)
ADNI$_{all}$	1966	11	40	29	14	6	41	76.9 ± 6.6
ADNI$_{train}$	1568	11	40	29	14	6	41	77.0 ± 6.6
ADNI$_{test}$	398	12	39	28	16	5	43	76.6 ± 6.9
AddNeuroMed	122	2	21	41	23	13	66	75.7 ± 6.1
MemClin$_{all}$	384	3	35	39	18	6	57	68.0 ± 8.2
MemClin$_{train}$	318	3	34	40	17	6	56	68.0 ± 8.2
MemClin$_{test}$	66	4	39	33	21	4	61	68.3 ± 8.2
E-DLB$_{all}$	645	14	41	29	12	4	44	73.7 ± 8.0
E-DLB$^{train}_{25\%}$	149	15	40	28	12	4	43	74.2 ± 8.1
E-DLB$^{train}_{50\%}$	324	15	41	29	11	3	45	74.0 ± 8.1
E-DLB$^{test}_{50\%}$	321	12	42	29	12	5	43	73.4 ± 8.0
E-DLB$_{C_1}$	101	16	40	29	11	4	23	75.9 ± 6.5
E-DLB$_{C_2}$	165	19	41	28	6	5	41	72.3 ± 9.5
E-DLB$_{C_{3-12}}$	379	11	42	30	15	3	51	73.7 ± 7.5
E-DLB$_{AD}$	193	4	30	38	20	7	55	75.7 ± 7.7
E-DLB$_{DLB}$	266	14	43	28	11	4	44	73.6 ± 8.2
E-DLB$_{PDD}$	97	19	46	27	7	1	15	71.8 ± 7.0

2.2. Radiologist ratings

An experienced neuroradiologist (Lena Cavallin, L.C.) visually rated 3117 T_1-weighted brain images (blind to age and sex) according to the established MTA rating scale. These ratings have been used in previous studies on AD (Ferreira et al., 2015) and DLB (Oppedal et al., 2019) by our group, and the distribution of ratings are shown in Table 2. These rating scales provide a quantitative measure of atrophy in specific regions, and while they are often used for dementia diagnosis the rating scales themselves are independent of diagnosis, age and sex. L.C. has previously demonstrated excellent inter- and intra-rater agreements in research studies (Mårtensson et al., 2019).

2.3. Model description

Our group recently proposed a method we call *AVRA* (Automatic Visual Ratings of Atrophy) that provides computed scores of three visual rating scales commonly used clinically: Scheltens' MTA scale (see Fig. 1), Pasquier's frontal subscale of global cortical atrophy (GCA-F), and Koedam's scale of posterior atrophy (PA) (Mårtensson et al., 2019). AVRA showed substantial rating-agreement to an expert neuroradiologist in all three scales on a hold-out test set (N=464) that was drawn from the same distribution as the training data (N=1886) from two AD cohorts. Since the measures are independent of diagnosis, sex and age, a DL tool such as AVRA (trained end-to-end and does its own feature-extraction from the entire brain volume) should work equally well on different disease populations.

For this experiment we focused only on the MTA scale and used the same network architecture as previously described in Mårtensson et al., 2019, but with different combinations of cohorts in the training set. Briefly, AVRA is a Recurrent Convolutional Neural Network (R-CNN) that inputs an MRI volume, which is processed slice-by-slice by the model. A residual attention network (Wang et al., 2017) is used to extract features from each slice, and these are forwarded to a Long-Short Term Memory (LSTM) network (Hochreiter and Schmidhuber, 1997). The LSTM modules remember relevant information provided from each slice and use it to predict the atrophy score the radiologist would give. This prediction is continuous, but when studying the inter-rater agreement with the radiologist, expressed in kappa statistics or accuracy, we round AVRA's output to the nearest integer.

A trained version of AVRA targeted towards neuroimaging researchers is publicly available at https://github.com/gsmartensson/avra_public.

2.4. Training procedure

To systematically investigate the performance in new data distributions we trained multiple versions of AVRA on images from different combinations of cohorts. For each training set we kept the number of images fixed to the maximum size of the ADNI training dataset ($N = 1568$), since more training data generally leads to better performance and could bias the results. ADNI was the largest dataset with ratings available to us, and needed to be part of all training sets in order for the number of images to be large enough for training. When combining data from an additional cohort, we replaced a subject in ADNI with one from the new cohort that had the same ratings from the radiologist. This way, both the size and the distribution of the training data were kept constant.

We used the same hyper-parameters (and network architecture) as in our previous paper (Mårtensson et al., 2019), which were decided through cross-validation performance across all three rating scales (MTA, PA, and GCA-F). To avoid the process of hyper-parameter tuning in this study, we replicated our previous training procedure. Briefly, for all training set combinations the data was split into five partitions. We trained five models per dataset, each leaving one partition out of its training set resulting in an average training set size of $N = 1254$ images. All models were trained for 200 epochs, optimized through stochastic gradient descent, with a minibatch size of 20 and a cyclic learning rate varying between 0.01 and 0.0005, following a cosine annealing schedule (Loshchilov and Hutter, 2016). In Mårtensson et al., 2019 we used the five networks as an ensemble model on the hold-out test set for more accurate ratings. In this study we used the five models to indicate performance variability due to subject composition and weight initialization, and to reduce the risk of spurious findings that can occur when only reporting a single metric. All images were a priori registered to the MNI standard brain with FSL FLIRT 6.0 (FMRIB's Linear Image Registration Tool) (Greve and Fischl, 2009; Jenkinson and Smith, 2001) through rigid transformation (translation and rotation only), similar to aligning images to the anterior and posterior commissures. Registration was performed to facilitate the training procedure by yielding consistent voxel resolutions ($1x1x1mm^3$) and centering of images. Apart from normalizing each image to zero mean and unit variance, no additional image preprocessing (such as intensity normalization or scull-stripping) was performed.

Each of the cohorts have different characteristics, as outlined in Table 1. Since the E-DLB cohort was highly diverse in terms of scanners and disease population, we stratified it into different partitions (some with overlap, but no training/test set pairs shared any images) in order to isolate specific features. To investigate the performance drop due to OOD test data, we randomly assigned each subject into $\text{E-DLB}^{\text{train}}_{25\%}$, $\text{E-DLB}^{\text{train}}_{50\%}$ and $\text{E-DLB}^{\text{test}}_{50\%}$, where the numbers refer to the percentage of subject from the whole cohort and with no overlap between $^{\text{train}}$ and $^{\text{test}}$. This setup aims to simulate realistic ways of introducing a DL model into a new clinic: 1) *as is* (i.e. no additional labeled data from the new clinic), 2) retraining, or finetuning, the existing model with *some* additional labeled data from the same clinics ($\text{E-DLB}^{\text{train}}_{25\%}$), 3) same as 2) but with twice as much additional data ($\text{E-DLB}^{\text{train}}_{50\%}$).

To assess the impact of disease population we sampled individuals on the AD spectrum (E-DLB_{AD}), DLB spectrum ($\text{E-DLB}_{\text{DLB}}$), or with PDD ($\text{E-DLB}_{\text{PDD}}$) into three subsets. Since the main bulk of training images comes from ADNI–an AD cohort–it is of interest to see if the models overfit to AD atrophy patterns and are influenced by neighboring regions in the medial temporal lobe not part of the MTA scale.

To study if AVRA's generalizability improved when widening the training data distribution we also computed the performance on data from two clinics that we refer to as E-DLB_{C_1} and E-DLB_{C_2}. A single 3T scanner and protocol was used at each site for scanning, yet with visibly different image intensities (see image examples in Fig. 1). We view these centers as "external validation sets" to estimate the performance we may expect if implementing AVRA in a new memory clinic (although single-scanner usage and study populations may not perfectly represent a memory clinic sample). We included data from all other centers (C_3, C_4, ..., C_{12}) in our training set ($\text{E-DLB}_{C_{3-12}}$) to study if more heterogeneous training data improved generalization to new protocols.

2.5. Evaluation metrics

We assess the performance of AVRA using Cohen's linearly weighted kappa κ_w, which is the most common metric to assess inter- and intra-rater agreement for visual ratings in the literature. It ranges from [-1,1] where $\kappa_w \in$ [0.2,0.4) is generally considered *fair*, $\kappa_w \in$ [0.4,0.6) *moderate*, $\kappa_w \in$ [0.6,0.8) *substantial* and $\kappa_w \in$ [0.8,1] *almost perfect* (Landis and Koch, 1977). As opposed to accuracy, κ_w takes the rating distributions of the two sets into account, which is particularly useful when the number of ratings in each class are imbalanced. As comparison, AVRA achieved inter-rater agreements of $\kappa_w = 0.72$ - 0.74 (left and right MTA, respectively) to an expert radiologist on a test set from the same data distribution as the training data in Mårtensson et al., 2019, similar to reported inter-rater agreements between two radiologists. We computed mean and standard deviations of the predictions from the five models trained on each training set combination. Since using κ_w required rounding AVRA's continuous predictions to the nearest integer, mean squared error (MSE) was also reported. Ensemble model performance and accuracies are included as Supplementary Data.

Table 3
Rating agreement between AVRA and a neuroradiologist expressed in Cohen's κ_w and mean squared error (MSE) for various test sets when trained on different combinations of training cohorts. The values represent the mean and standard deviations of five networks independently trained on 80% of $N = 1568$ images, with a fixed label distribution in each training set. A ✓ symbol in a column denotes that the cohort of that row was part of the training set. E.g. the first column shows the rating agreement and MSE for different test sets when trained only on ADNI, the second when trained on ADNI+AddNeuroMed, etc. If there was any overlap between images in a training and test set combination no agreement was computed (listed as '–' in the table). The greatest agreement values for each test set are in bold.

Cohort	Cohorts incl. in training									
$\text{ADNI}^{\text{train}}$	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
AddNeuroMed		✓				✓	✓			✓
$\text{MemClin}^{\text{train}}$			✓		✓		✓			✓
$\text{E-DLB}_{C_{3-12}}$				✓	✓	✓	✓			
$\text{E-DLB}^{\text{train}}_{25\%}$								✓		
$\text{E-DLB}^{\text{train}}_{50\%}$									✓	✓
	Cohen's κ_w									
$\text{ADNI}^{\text{test}}$	0.67 ± .02	**0.69 ± .01**	0.67 ± .02	0.69 ± .02	0.66 ± .01	0.67 ± .02	0.67 ± .02	0.66 ± .01	0.67 ± .01	0.67 ± .02
AddNeuroMed	**0.66 ± .01**	–	0.64 ± .02	0.65 ± .01	0.61 ± .05	–	–	0.63 ± .03	0.63 ± .03	–
MemClin	0.62 ± .02	0.62 ± .02	–	0.63 ± .02	–	**0.64 ± .03**	–	0.62 ± .05	0.61 ± .03	–
$\text{MemClin}^{\text{test}}$	0.64 ± .04	0.65 ± .03	0.72 ± .03	0.67 ± .03	**0.74 ± .04**	0.66 ± .05	0.69 ± .02	0.65 ± .07	0.59 ± .05	0.71 ± .02
$\text{E-DLB}_{\text{all}}$	0.58 ± .02	0.58 ± .02	**0.61 ± .01**	–	–	–	–	–	–	–
$\text{E-DLB}^{\text{test}}_{50\%}$	0.59 ± .02	0.58 ± .01	0.60 ± .02	–	–	–	–	0.62 ± .02	0.63 ± .02	**0.65 ± .02**
E-DLB_{AD}	0.52 ± .03	0.52 ± .01	**0.57 ± .03**	–	–	–	–	–	–	–
$\text{E-DLB}_{\text{DLB}}$	0.59 ± .03	0.58 ± .03	**0.61 ± .01**	–	–	–	–	–	–	–
$\text{E-DLB}_{\text{PDD}}$	0.58 ± .04	0.58 ± .06	**0.60 ± .05**	–	–	–	–	–	–	–
E-DLB_{C_1}	0.30 ± .04	0.31 ± .04	0.49 ± .07	0.42 ± .07	0.51 ± .05	0.52 ± .05	**0.52 ± .03**	–	–	–
E-DLB_{C_2}	0.64 ± .04	0.61 ± .02	0.64 ± .01	0.64 ± .04	**0.65 ± .02**	0.63 ± .03	0.64 ± .02	–	–	–
	Mean squared error									
$\text{ADNI}^{\text{test}}$	0.31 ± .02	**0.29 ± .01**	**0.29 ± .01**	**0.29 ± .01**	0.32 ± .01	0.30 ± .02	0.30 ± .02	0.31 ± .01	0.31 ± .01	0.31 ± .02
AddNeuroMed	**0.27 ± .01**	–	0.28 ± .01	0.30 ± .01	0.32 ± .05	–	–	**0.27 ± .01**	0.29 ± .03	–
MemClin	0.34 ± .02	0.31 ± .02	–	0.31 ± .02	–	**0.28 ± .02**	–	0.31 ± .04	0.32 ± .02	–
$\text{MemClin}^{\text{test}}$	0.33 ± .02	0.29 ± .04	0.23 ± .02	0.27 ± .03	**0.22 ± .03**	0.26 ± .03	0.24 ± .01	0.29 ± .03	0.31 ± .04	0.25 ± .02
$\text{E-DLB}_{\text{all}}$	0.41 ± .02	0.41 ± .03	**0.36 ± .02**	–	–	–	–	–	–	–
$\text{E-DLB}^{\text{test}}_{50\%}$	0.41 ± .02	0.40 ± .03	0.36 ± .03	–	–	–	–	0.35 ± .02	0.34 ± .02	**0.33 ± .01**
E-DLB_{AD}	0.50 ± .05	0.48 ± .02	**0.39 ± .05**	–	–	–	–	–	–	–
$\text{E-DLB}_{\text{DLB}}$	0.41 ± .04	0.42 ± .03	**0.38 ± .01**	–	–	–	–	–	–	–
$\text{E-DLB}_{\text{PDD}}$	0.30 ± .03	0.30 ± .05	**0.27 ± .02**	–	–	–	–	–	–	–
E-DLB_{C_1}	0.83 ± .11	0.79 ± .13	0.49 ± .12	0.53 ± .09	0.46 ± .08	0.45 ± .04	**0.44 ± .05**	–	–	–
E-DLB_{C_2}	**0.28 ± .03**	0.32 ± .02	0.30 ± .01	0.30 ± .04	0.29 ± .03	0.30 ± .02	0.30 ± .03	–	–	–

3. Results

The rating agreements between AVRA and the neuroradiologist are summarized in Table 3. When only training on the research cohort ADNI we saw a general drop in performance in clinical cohorts compared to the test set of ADNI–particularly in the E-DLB_{C_1} set. Adding data from the similar cohort AddNeuroMed helped little in improving generalization, whereas the inclusion of clinical MemClin had a positive impact on performance. The overall impression was that including data from clinical cohorts in the training set improved the rating agreements and accuracies in the clinical test sets, although not consistently. Surprisingly, the rating agreement was greater in the sub-cohorts $\text{E-DLB}_{\text{DLB}}$ and $\text{E-DLB}_{\text{PDD}}$ than in E-DLB_{AD} when only training on images from AD cohorts.

In Fig. 2 we focus on the centers E-DLB_{C_1} and E-DLB_{C_2}, where AVRA's performance metrics were particularly low (C_1) or close to within-distribution test set performances (C_2) when trained on research data. We compared the predictions made by the ensemble models trained only on ADNI (x-axis) to when trained on data from ADNI and clinical images from the MemClin and $\text{E-DLB}_{C_{3-12}}$ cohorts. Thus, no images from these centers had been part in either of the training sets, but the latter included clinical images acquired from a wider range of protocols. We observed systematic differences in the predictions between the two models, most notably in the C_1 cohort. Note the intensity differences in tissue types between images from ADNI, C_1 and C_2 in Fig. 1.

AVRA's ensemble predictions on the test-retest cohort are plotted in Fig. 3 for the models trained on the least and most heterogeneous data. We observed small intra-subject rating variability for most subjects, within the same model. It was mainly the predictions of the two images acquired with the Siemens Trio 3T that stood out. While the direction of the rating prediction differences were not consistent across subjects, it suggests that AVRA may systematically rate images acquired from some protocols/scanners differently. Comparing the two versions we see that the model trained only on ADNI systematically rates images lower than when trained also on clinical data–same as in Fig. 2. Further, it should be noted that these participants were younger than in any of the training cohorts and–for the patients suffering from MS–from a different disease population.

4. Discussion

In this study we systematically showed that the performance of a CNN trained on MRI images from homogeneous research cohorts generally drops when applied to clinical data. In one center–where image intensity was visibly different to images from the training data–the performance of AVRA was lower due to a systematic underestimation. However, by including images acquired from a wider range of scanners and protocols in the training set we observed an increase in robustness/reliability of the DL model in unseen OOD data–without a substantial damage to the within-distribution test set performance. This is the first study on a large MRI neuroimaging dataset labeled by the same expert neuroradiologist (thus no inter-observability bias) and with fixed training set sizes and label distribution. These results add to the evidence that rigorous testing of DL applications in medical imaging needs to be performed on external data before being used in clinics.

From our results in Table 3 we note several interesting findings. First, the level of agreement is lower in the clinical cohorts MemClin and $\text{E-DLB}_{\text{all}}$ when only trained on research cohorts (ADNI with or without AddNeuroMed). This suggests that we can expect

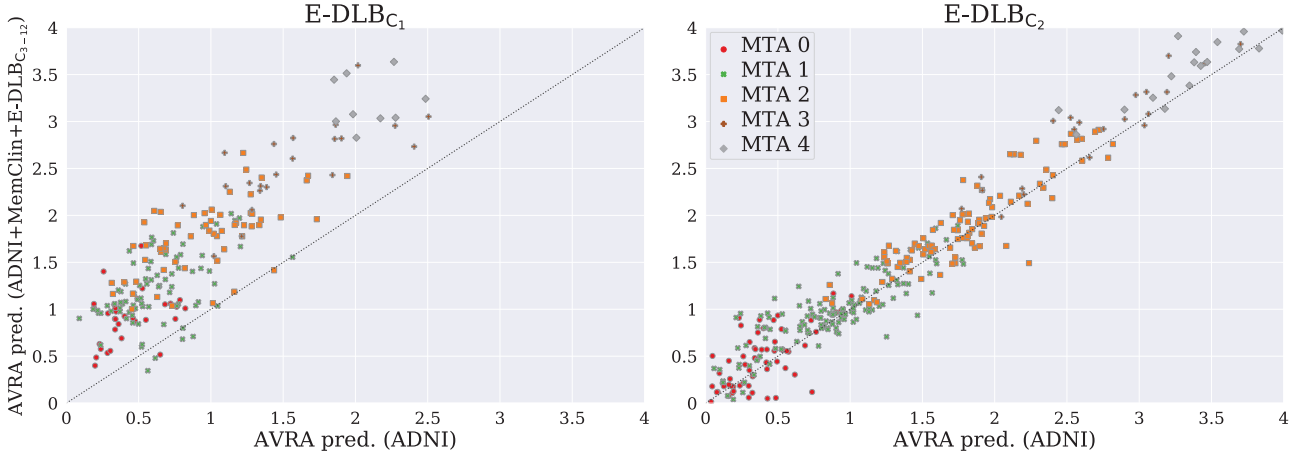

Fig. 2. Scatter plot of AVRA's ensemble predictions of the images from E-DLB$_{C_1}$ (left) and E-DLB$_{C_2}$ (right), which respectively showed poor and good agreement using the baseline model trained only on ADNI. Each dot represents a subject, where the x-coordinate is the prediction when trained only on the ADNI cohort, and the y-coordinate where images from clinical cohorts were also represented in the training data. The marker symbols and colors indicate radiologist's ("ground truth") ratings. The dotted line show $x = y$, making it clear that AVRA's predictions were systematically lower than if including data from a wider distribution in the training set. This was very prominent in the E-DLB$_{C_1}$ cohort, but also notable in E-DLB$_{C_2}$.

Fig. 3. Boxplot of AVRA's ensemble ratings of left MTA (left column) and right MTA (right column) for all participants in the test-retest dataset. Top row: model trained only on ADNI; Bottom row: model trained on ADNI+AddNeuroMed+MemClin+E-DLB$_{C_{3-12}}$. Each subject was scanned twice with repositioning in three different scanners, and each image's AVRA rating is plotted in different colors depending on scanner. Individuals denoted with the prefix "HC" were healthy controls and "MS" were patients with Multiple Sclerosis.

a degradation of a CNN model when applied to MRI images acquired with protocols not seen during training, which is problematic for scalable deployment in clinics. Similar findings have previously been reported on segmentation tasks on cross-institutional MRI data (Albadawy et al., 2018; Perone et al., 2019) and chest x-ray data (Pooch et al., 2019; Yao et al., 2019). While inter-rater agreement levels of $\kappa_w > 0.6$ might be considered acceptable in many clinical situations for visual ratings (reported κ_w levels between radiologists are typically between 0.6 and 0.8 in previous studies (Mårtensson et al., 2019)) we see that the agreement in E-DLB$_{C_1}$ is substantially lower when only trained on data from harmonized research cohorts. The degree of variation in performance

across the multiple test sets is concerning and makes it difficult to assess how well a deep learning model generalizes to clinical data.

Second, including images of larger variability from clinical cohorts improved performance even when keeping the training set size and label distribution fixed. Including data from MemClin in the training set had a positive impact on the E-DLB sets and vice versa. This implies that by training a supervised DL model on data from a wide range of scanners, protocols, field strengths and diagnoses/labels it is possible to achieve acceptable performance on new unseen data. The systematic prediction differences for E-DLB$_{C_1}$ in Fig. 2 illustrates this point well, where training data from other memory clinics had a large impact on predictions.

Third, we investigated the performance of AVRA in DLB and PDD populations when trained on images of subjects on the AD spectrum (from healthy controls, to patients with mild cognitive impairment and AD). Unexpectedly, the agreement was higher in both the DLB and PDD populations than in the AD population from the E-DLB cohort. These results could potentially be explained by the differences in rating distributions between the disease populations. PDD and DLB individuals generally had lower MTA ratings than the AD patients, and from Fig. 2 we see that the model trained only on ADNI tends to rate too low–particularly for higher MTA values. Thus, this systematic error could affect AVRA's performance in the AD population more. However, the relatively high agreements of E-DLB$_{DLB}$ and E-DLB$_{PDD}$ show potential that AVRA has the ability to generalize across disease populations. This finding is likely attributed to the strength of the clinical visual rating scales–which are disease-unspecific by design–and demonstrate the power of incorporating domain knowledge when building DL models. A previous study on applying machine learning models (SVM) on unseen clinical data reported and discussed difficulties in determining if subjects suffered from mixed pathologies (e.g. both AD and FTD) or a misdiagnosis (Klöppel et al., 2015). A model trained to discriminate between e.g. AD patients from healthy controls–both generally defined by strict inclusion and exclusion criteria in research cohorts–does just that. Applying an "AD model" like this in a more heterogeneous cohort with controls, AD and DLB subjects, would thus most probably misdiagnose DLB as AD due to resembling patterns of atrophy (Oppedal et al., 2019).

The test-retest results (Fig. 3) show impressive consistency for each DL model in most predictions. The ratings from the version trained on multiple datasets seems to yield higher variability for many subjects compared to when only trained on ADNI. Given that this model showed better generalization in the analyses summarized in Table 3, this is a bit counterintuitive. It should be noted however that these differences are small considering being trained on integer ratings with some degree of intra-rater variability. The explanation for this inter-scanner variability could partially be due to a minor overfit to scanner and protocol. This is however to prefer to the ADNI-model where the ratings seems to be systematically too low. Within-scanner and within-field strength variability was practically non-existent, and it is only the images of the 3T scanner that notably deviates for some patients. This means that we expect AVRA to be useful for longitudinal studies, where the data is typically collected in a harmonized way. Guo et al. (2019) analyzed the same dataset using different (non-machine learning) neuroimaging softwares and reported smaller within- than between-scanner variability. A previous study investigating the impact choice of scanner and field strength have on the performance of an SVM-classifier found the largest performance drop when training on 1.5T data and testing on 3T data and vice versa, while generalizing well to new scanners within the same field strength (Abdulkadir et al., 2011). Their analyses were done in the ADNI cohort, with protocols harmonized using a phantom to reduce scanner and site variability. For computer scientists it would solve many practical issues if protocols were harmonized across clinics, and that these protocols were used as default. However, this seems unlikely given the enormous effort of implementing it, the development of new (improved) sequences, and disrupting habits and workflows of clinicians. Further, the real gain of machine learning applications would be on CT images–as it is cheaper and more commonly available–where image quality variation is even greater. Thus, scanner/protocol generalization remains an important issue that needs solving prior to deploying DL models as clinical aid. Since labeled data in medicine is often difficult or expensive to acquire semi-supervised approaches may play a big role in medical machine learning applications as it allows the inclusion of unlabeled images in the training data. This has been shown to improve generalization on medical OOD data (Kamnitsas et al., 2017; Orbes-Arteaga et al., 2019; Perone et al., 2019).

The current study has some limitations that we leave as future studies. Foremost, we trained and evaluated a single network architecture and we cannot say to what degree the results are representative of DL models in general. By using the same hyperparameters as in Mårtensson et al., 2019 (tuned to optimize performance on a within-distribution cross-validation set) nothing prevented AVRA from overfitting to the training protocol. Further, while the kappa metric is the most common way to quantify reliability of visual ratings, it can be noisy since we need to round the prediction to nearest integer. The MSE metric does not require rounding but is on the other hand sensitive to outliers. Furthermore, since AVRA takes an unprocessed MRI image (apart from a rotation and translation) as input–just as a radiologist would do–we did not explore the impact that additional preprocessing could have on generalization. E.g. intensity normalization has been shown to improve image synthesis using DL (Reinhold et al., 2019), and may reduce the inter-scanner variability. However, it should be noted that traditional neuroimaging softwares–such as SPM, FreeSurfer, and FSL–apply extensive preprocessing but are still vulnerable to the effect of inter-scanner variability (Guo et al., 2019).

5. Conclusion

In this study we assessed how well a supervised deep learning model (AVRA), trained on MRI brain images to predict Scheltens' MTA score, generalizes to external clinical data. More specifically, we trained multiple versions of AVRA on data from different combinations of research and clinical cohorts, while keeping training set size and label distribution fixed. We found that AVRA trained on homogeneous data from a research cohort generalized well to cohorts with similar protocols, but worse when applied to clinical data. On images from one specific memory clinic the performance dropped to an unacceptably low level. Including more heterogeneous data from a wider range of scanner and protocols during training improved the performance also in out-of-distribution data. Furthermore, when applying AVRA on images of patients suffering from other neurological disorders than AD we did not observe a noticeable decrease in performance. From these findings we advocate that DL models need to be rigorously tested in OOD data before being deployed in clinics. This is, to our knowledge, the largest and most comprehensive study to date on the effect of domain shift in MRI images and deep learning models.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Gustav Mårtensson: Conceptualization, Methodology, Software, Data curation, Investigation, Visualization, Writing - original draft. **Daniel Ferreira:** Conceptualization, Methodology, Data curation, Writing - review & editing. **Tobias Granberg:** Resources, Writing - review & editing. **Lena Cavallin:** Resources, Writing - review & editing. **Ketil Oppedal:** Resources, Writing - review & editing. **Alessandro Padovani:** Resources, Writing - review & editing. **Irena Rektorova:** Resources, Writing - review & editing. **Laura Bonanni:** Resources, Writing - review & editing. **Matteo Pardini:** Resources, Writing - review & editing. **Milica G Kramberger:** Resources, Writing - review & editing. **John-Paul Taylor:** Resources, Writing - review & editing. **Jakub Hort:** Resources, Writing - review & editing. **Jón Snædal:** Resources, Writing - review & editing. **Jaime Kulisevsky:** Resources, Writing - review & editing. **Frederic Blanc:** Resources, Writing - review & editing. **Angelo Antonini:** Resources, Writing - review & editing. **Patrizia Mecocci:** Resources, Writing - review & editing. **Bruno Vellas:** Resources, Writing - review & editing. **Magda Tsolaki:** Resources, Writing - review & editing. **Iwona Kłoszewska:** Resources, Writing - review & editing. **Hilkka Soininen:** Resources, Writing - review & editing. **Simon Lovestone:** Resources, Writing - review & editing. **Andrew Simmons:** Resources, Writing - review & editing. **Dag Aarsland:** Resources, Writing - review & editing. **Eric Westman:** Conceptualization, Methodology, Supervision, Writing - review & editing.

Acknowledgements

We would like to thank the Swedish Foundation for Strategic Research (SSF), The Swedish Research Council (VR), the Strategic Research Programme in Neuroscience at Karolinska Institutet (StratNeuro), Swedish Brain Power, Centrum för innovativ medicin (CIMED), Stiftelsen Olle Engkvist Byggmästare, the regional agreement on medical training and clinical research (ALF) between Stockholm County Council and Karolinska Institutet, Hjärnfonden, Alzheimerfonden, the Åke Wiberg Foundation and Birgitta och Sten Westerberg for additional financial support. The Titan X Pascal used for this research was donated by the NVIDIA Corporation.

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

Supplementary material

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.media.2020.101714.

References

Abdulkadir, A., Mortamet, B., Vemuri, P., Jack, C.R., Krueger, G., Klöppel, S., 2011. Effects of hardware heterogeneity on the performance of SVM Alzheimer's disease classifier. NeuroImage 58 (3), 785–792. doi:10.1016/j.neuroimage.2011.06.029.

Albadawy, E.A., Saha, A., Mazurowski, M.A., 2018. Deep learning for segmentation of brain tumors: impact of cross-institutional training and testing: impact. Med. Phys. 45 (3), 1150–1158. doi:10.1002/mp.12752.

Boyle, P.A., Yu, L., Wilson, R.S., Leurgans, S.E., Schneider, J.A., Bennett, D.A., 2018. Person-specific contribution of neuropathologies to cognitive loss in old age. Ann. Neurol. 83 (1), 74–83. doi:10.1002/ana.25123.

De Fauw, J., Ledsam, J.R., Romera-Paredes, B., Nikolov, S., Tomasev, N., Blackwell, S., Askham, H., Glorot, X., O'Donoghue, B., Visentin, D., van den Driessche, G., Lakshminarayanan, B., Meyer, C., Mackinder, F., Bouton, S., Ayoub, K., Chopra, R., King, D., Karthikesalingam, A., Hughes, C.O., Raine, R., Hughes, J., Sim, D.A., Egan, C., Tufail, A., Montgomery, H., Hassabis, D., Rees, G., Back, T., Khaw, P.T., Suleyman, M., Cornebise, J., Keane, P.A., Ronneberger, O., 2018. Clinically applicable deep learning for diagnosis and referral in retinal disease.. Nat. Med. 24 (9), 1342–1350. doi:10.1038/s41591-018-0107-6.

Falahati, F., Ferreira, D., Soininen, H., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Lovestone, S., Eriksdotter, M., Wahlund, L.O., Simmons, A., Westman, E., 2016. The effect of age correction on multivariate classification in Alzheimer's disease, with a focus on the characteristics of incorrectly and correctly classified subjects. Brain Topogr. 29 (2), 296–307. doi:10.1007/s10548-015-0455-1.

Ferreira, D., Cavallin, L., Larsson, E.-M., Muehlboeck, J.-S., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Simmons, A., Wahlund, L.-O., Westman, E., 2015. Practical cut-offs for visual rating scales of medial temporal, frontal and posterior atrophy in Alzheimer's disease and mild cognitive impairment. J. Intern. Med. 278 (3), 277–290. doi:10.1111/joim.12358.

Ferreira, D., Hansson, O., Barroso, J., Molina, Y., Machado, A., Hernández-Cabrera, J.A., Muehlboeck, J.-S., Stomrud, E., Nägga, K., Lindberg, O., Ames, D., Kalpouzos, G., Fratiglioni, L., Bäckman, L., Graff, C., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Ahlström, H., Lind, L., Larsson, E.-M., Wahlund, L.-O., Simmons, A., Westman, E., 2017. The interactive effect of demographic and clinical factors on hippocampal volume: a multicohort study on 1958 cognitively normal individuals. Hippocampus 27 (6), 653–667. doi:10.1002/hipo.22721.

Ferreira, D., Shams, S., Cavallin, L., Viitanen, M., Martola, J., Granberg, T., Shams, M., Aspelin, P., Kristoffersen-Wiberg, M., Nordberg, A., Wahlund, L.O., Westman, E., 2018. The contribution of small vessel disease to subtypes of Alzheimer's disease: a study on cerebrospinal fluid and imaging biomarkers. Neurobiol. Aging 70, 18–29. doi:10.1016/j.neurobiolaging.2018.05.028.

Greve, D.N., Fischl, B., 2009. Accurate and robust brain image alignment using boundary-based registration. NeuroImage 48 (1), 63–72. doi:10.1016/j.neuroimage.2009.06.060.

Guo, C., Ferreira, D., Fink, K., Westman, E., Granberg, T., 2019. Repeatability and reproducibility of FreeSurfer, FSL-SIENAX and SPM brain volumetric measurements and the effect of lesion filling in multiple sclerosis. Eur. Radiol. 29 (3), 1355–1364. doi:10.1007/s00330-018-5710-x.

Hochreiter, S., Schmidhuber, J., 1997. Long short-term memory. Neural Comput. 9, 1735–1780.

Jenkinson, M., Smith, S., 2001. A global optimisation method for robust affine registration of brain images. Med. Image Anal. 5 (2), 143–156. doi:10.1016/S1361-8415(01)00036-6.

Kamnitsas, K., Baumgartner, C., Ledig, C., Newcombe, V., Simpson, J., Kane, A., Menon, D., Nori, A., Criminisi, A., Rueckert, D., Glocker, B., 2017. Unsupervised domain adaptation in brain lesion segmentation with adversarial networks. In: Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics), 10265 LNCS, pp. 597–609. doi:10.1007/978-3-319-59050-9_47. arXiv:1612.08894v1.

Klöppel, S., Peter, J., Ludl, A., Pilatus, A., Maier, S., Mader, I., Heimbach, B., Frings, L., Egger, K., Dukart, J., Schroeter, M.L., Perneczky, R., Häussermann, P., Vach, W., Urbach, H., Teipel, S., Hüll, M., Abdulkadir, A., 2015. Applying automated MR-based diagnostic methods to the memory clinic: a prospective study. J. Alzheimers Dis. 47 (4), 939–954. doi:10.3233/JAD-150334.

Kramberger, M.G., Auestad, B., Garcia-Ptacek, S., Abdelnour, C., Olmo, J.G., Walker, Z., Lemstra, A.W., Londos, E., Blanc, F., Bonanni, L., McKeith, I., Winblad, B., De Jong, F.J., Nobili, F., Stefanova, E., Petrova, M., Falup-Pecurariu, C., Rektorova, I., Bostantjopoulou, S., Biundo, R., Weintraub, D., Aarsland, D., 2017. Long-Term cognitive decline in dementia with lewy bodies in a large multicenter, international cohort. J. Alzheimers Dis. 57 (3), 787–795. doi:10.3233/JAD-161109.

Landis, J.R., Koch, G.G., 1977. The Measurement of Observer Agreement for Categorical Data. Biometrics 33 (1), 159. doi:10.2307/2529310. NIHMS150003.

Lindberg, O., Östberg, P., Zandbelt, B.B., Öberg, J., Zhang, Y., Andersen, C., Looi, J.C., Bogdanović, N., Wahlund, L.O., 2009. Cortical morphometric subclassification of frontotemporal lobar degeneration. Am. J. Neuroradiol. 30 (6), 1233–1239. doi:10.3174/ajnr.A1545.

Litjens, G., Kooi, T., Bejnordi, B.E., Setio, A.A.A., Ciompi, F., Ghafoorian, M., van der Laak, J.A., van Ginneken, B., Sánchez, C.I., 2017. A survey on deep learning in

medical image analysis. Medical Image Analysis 42 (December 2012), 60–88. doi:10.1016/j.media.2017.07.005. 1702.05747.

Loshchilov, I., Hutter, F., 2016. SGDR: Stochastic gradient descent with warm restarts. arXiv:1608.03983, 1–16. 10.1002/fut

Lundervold, A.S., Lundervold, A., 2019. An overview of deep learning in medical imaging focusing on MRI. Z. Med. Phys. 29 (2), 102–127. doi:10.1016/j.zemedi.2018.11.002.

Mårtensson, G., Ferreira, D., Cavallin, L., Muehlboeck, J.-S., Wahlund, L.-O., Wang, C., Westman, E., 2019. AVRA: Automatic visual ratings of atrophy from MRI images using recurrent convolutional neural networks. NeuroImage: Clinical 23 (March), 101872. doi:10.1016/j.nicl.2019.101872. arXiv:1901.00418.

Mårtensson, G., Pereira, J.B., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Soininen, H., Lovestone, S., Simmons, A., Volpe, G., Westman, E., 2018. Stability of graph theoretical measures in structural brain networks in Alzheimer's disease. Sci. Rep. 8 (1), 11592. doi:10.1038/s41598-018-29927-0.

Muehlboeck, J.-S., Westman, E., Simmons, A., 2014. TheHiveDB image data management and analysis framework. Front. Neuroinform. 7 (January), 49. doi:10.3389/fninf.2013.00049.

Oppedal, K., Ferreira, D., Cavallin, L., Lemstra, A.W., ten Kate, M., Padovani, A., Rektorova, I., Bonanni, L., Wahlund, L.O., Engedal, K., Nobili, F., Kramberger, M., Taylor, J.P., Hort, J., Snædal, J., Blanc, F., Walker, Z., Antonini, A., Westman, E., Aarsland, D., 2019. A signature pattern of cortical atrophy in dementia with Lewy bodies: a study on 333 patients from the European DLB consortium. Alzheimer's Dement. 15 (3), 400–409. doi:10.1016/j.jalz.2018.09.011.

Orbes-Arteaga, M., Varsavsky, T., Sudre, C.H., Eaton-Rosen, Z., Haddow, L.J., Sørensen, L., Nielsen, M., Pai, A., Ourselin, S., Modat, M., Nachev, P., Cardoso, M.J., 2019. Multi-Domain Adaptation in Brain MRI through Paired Consistency and Adversarial Learning. In: Multi-Domain Adaptation in Brain MRI through Paired Consistency and Adversarial Learning. DART, 11795, pp. 54–62. 1908.05959.

Pooch, E. H. P., Ballester, P. L., Barros, R. C., 2019. Can we trust deep learning models diagnosis? The impact of domain shift in chest radiograph classification. arXiv:1909.01940.

Perone, C.S., Ballester, P., Barros, R.C., Cohen-Adad, J., 2019. Unsupervised domain adaptation for medical imaging segmentation with self-ensembling NeuroImage 194, 1–11 doi:10.1016/j.neuroimage.2019.03.026.

Reinhold, J.C., Dewey, B.E., Carass, A., Prince, J.L., 2019. Evaluating the impact of intensity normalization on MR image synthesis. In: Angelini, E.D., Landman, B.A. (Eds.), Medical Imaging 2019: Image Processing. SPIE, p. 126. doi:10.1117/12.2513089. 1812.04652.

Scheltens, P., Leys, D., Barkhof, F., Huglo, D., Weinstein, H.C., Vermersch, P., Kuiper, M., Steinling, M., Wolters, E.C., Valk, J., 1992. Atrophy of medial temporal lobes on MRI in "probable" Alzheimer's disease and normal ageing: diagnostic value and neuropsychological correlates. J. Neurol. Neurosurg. Psychiatry 55, 967–972. doi:10.1136/jnnp.55.10.967.

Simmons, A., Westman, E., Muehlboeck, S., Mecocci, P., Vellas, B., Tsolaki, M., Kłoszewska, I., Wahlund, L.O., Soininen, H., Lovestone, S., Evans, A., Spenger, C., 2011. The AddNeuroMed framework for multi-centre MRI assessment of Alzheimer's disease: experience from the first 24 months. Int. J. Geriatr. Psychiatry 26 (1), 75–82. doi:10.1002/gps.2491.

Wang, F., Jiang, M., Qian, C., Yang, S., Li, C., Zhang, H., Wang, X., Tang, X., 2017. Residual attention network for image classification. In: Proceedings - 30th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, vol. 2017-Janua, pp. 6450–6458. doi:10.1109/CVPR.2017.683.

Yao, L., Prosky, J., Covington, B., Lyman, K., 2019. A strong baseline for domain adaptation and generalization in medical imaging. arXiv:1904.01638, 1–5.

Westman, E., Simmons, A., Muehlboeck, J.S., Mecocci, P., Vellas, B., Tsolaki, M., Kloszewska, I., Soininen, H., Weiner, M.W., Lovestone, S., Spenger, C., Wahlund, L.O., 2011. AddNeuroMed and ADNI: similar patterns of Alzheimer's atrophy and automated MRI classification accuracy in Europe and North America. NeuroImage 58 (3), 818–828. doi:10.1016/j.neuroimage.2011.06.065.

Zech, J.R., Badgeley, M.A., Liu, M., Costa, A.B., Titano, J.J., Oermann, E.K., 2018. Variable generalization performance of a deep learning model to detect pneumonia in chest radiographs: a cross-sectional study. PLoS Med. 15 (11), 1–17. doi:10.1371/journal.pmed.1002683.