Universiteit
Leiden
The Netherlands

Multimodal MRI-based classification of Alzheimer's disease

Vos, F. de

Citation
Vos, F. de. (2021, December 9). *Multimodal MRI-based classification of Alzheimer's disease*. Retrieved from https://hdl.handle.net/1887/3245855

Version:	Publisher's Version
License:	Licence agreement concerning inclusion of doctoral thesis in the Institutional Repository of the University of Leiden
Downloaded from:	https://hdl.handle.net/1887/3245855

Note: To cite this publication please use the final published version (if applicable).

General discussion

6.1 Summary of results

In this thesis we studied magnetic resonance imaging based (MRI-based) Alzheimer's disease (AD) classification. We derived a wide range of features from anatomical MRI, diffusion MRI and resting state functional MRI (fMRI) scans. This is important, because MRI scans contain a wealth of information that might help clinicians to improve AD diagnosis. However, anatomical MRI scans are usually only used to inspect specific AD characteristics like the size and shape of the hippocampus, and diffusion MRI and resting state fMRI scans are not at all used in clinical practice. We also studied the use of MRI scans for early identification of AD, and for future prediction of cognitive decline. This is important, because there is need for reliable AD diagnosis at the early symptomatic phase, or ideally even before the onset of the disease (Frisoni et al., 2017). This would provide patients with a prognosis so they can prepare for their future trajectory. In addition, early phase AD diagnosis is important for drug research, because early phase AD patients might still be susceptible for drugs (Scheltens et al., 2016).

We have tried to answer several hypotheses. In **chapter two** we used anatomical MRI scans to calculate multiple anatomical features, and used those to classify AD patients and healthy elderly controls. The grey matter density of the cortical structures and the volumes of subcortical structures yielded the highest classification scores. Moreover, combining multiple types of features increased accuracy. To put these results in perspective, we compared them with hippocampal volume and whole brain atrophy, because these are well established AD biomarkers. The combination of different anatomical MRI features yielded higher accuracies than these two simpler measures. This suggests that anatomical MRI scans contain more information than can be captured by simple measures only, and extracting this extra information benefits AD classification. These improvements are substantial, and might be considered for clinical use.

In **chapter three** we took a similar approach to resting state fMRI scans. We used those scans to calculate multiple feature types, and to classify AD patients and healthy elderly controls. We calculated functional connectivity (FC) using multiple approaches, like large scale connectivity matrices and whole brain voxel-wise seed correlation analyses. In addition, we derived indirect FC measures, including FC dynamics and graph measures. We also quantified the amplitude of low frequency fluctuations that reflect neuronal signal. The FC matrices, FC dynamics and the amplitude of low frequency

fluctuations yielded the highest classification scores. Also here, The combination of different types of features increased accuracy.

In **chapter four** we evaluated our MRI-based models for individual AD classification in diverse clinical populations from various memory clinics. This is an important step towards clinical applicability, because it reflects the day-to-day challenge of clinicians. This effort is difficult for at least two reasons. First, our AD classification models are developed using only AD patients and healthy elderly controls, and therefore they do not capture the full heterogeneity found in diverse clinical populations. Second, MRI scans collected at different sites can be largely different due to different scanner vendors and scanner settings (Ewers et al., 2006; Takao et al., 2014; Zhu et al., 2011). Consequently, MRI-based AD classification models do not always translate well to MRI scans acquired from another scanner.

To evaluate our MRI-based AD classification models for diverse clinical populations, we trained them on a single-centre data set consisting of AD patients and controls, and used them to assign AD probability scores to AD patients, mild cognitive impairment (MCI) patients and SMC patients from a multi-centre memory clinic data set. We used models based on anatomical MRI, diffusion MRI and resting state fMRI scans. For all three MRI modalities, the AD patients were on average assigned higher AD probability scores than MCI patients, and the MCI patients were on average assigned higher AD probability scores than SMC patients. However, for some models the three clinical groups show large overlap in AD probability scores. The anatomical MRI models generalised best to the memory clinic data. Especially the grey matter density model could differentiate well between all three groups. However, the diffusion MRI model did not generalise well to the multi-centre memory clinic data set, probably due to large scanner site differences in the diffusion MRI data. The resting state fMRI model generalised reasonably well to the memory clinic data, but it was inferior to the anatomical MRI model.

In **chapter five** we evaluated MRI's predictive accuracy for future cognitive decline. To this end, baseline multimodal MRI scans were used to predict two-year follow-up cognitive decline. Change in general cognitive functioning as measured by the mini-mental state examination (MMSE) was predicted above chance level, but the effect size was small. Change on six other tests, that measure specific aspects of memory and executive functioning, was not predicted above chance level. Our effort was complicated by a lack of observed cognitive decline. At average, there was significant

decline on only three out of seven cognitive tests, and the amount of decline for these tests was little. The lack of observed cognitive decline may have been caused by the large drop-out rate. Patients that decline more are probably more likely to drop out.

6.2 Strengths and limitations

Comprehensive comparison of MRI measures

A major strength of this thesis is the use of both anatomical MRI, diffusion MRI and resting state fMRI scans for AD classification. Moreover, for each scan type we used an extensive list of features. Especially for resting state fMRI scans we derived a large number of different features, including FC networks and its graph measures, the dynamics of these networks over time, FC within resting state networks, and the amplitude of low frequency fluctuations. For the purpose of AD classification, this has not been done as extensively before. This enabled a comparison between different MRI modalities, and between the different features per modality. Furthermore, we could combine all these different features, and we showed that they can complement each other for AD classification. Our results can aid researchers in choosing relevant MRI measures for AD research.

Extensive evaluation of generalisability

We took a careful approach to evaluate our AD classification models. Importantly, we used cross-validation to ensure that our models are not the result of overfitting. More specifically, we employed nested cross-validation. In nested cross-validation, there is an outer cross-validation loop that prevents ordinary overfitting. Ordinary overfitting is the practice of fitting too complex models that do not generalise to external data. Additionally, there is an inner loop that prevents over-hyping. Over-hyping is overfitting on the hyperparameter by trying out many different hyperparameter values and reporting the best result. Whereas overfitting is a problem that most researchers are aware of nowadays, over-hyping is still happening a lot, making results overoptimistic (Hosseini et al., 2020). We put effort in preventing both these types of overfitting to ensure that our results represent honest estimates of AD classification accuracy.

Although cross-validation uses 'out of sample' data for model evaluation, these data are still similar to the training data, because they originate from the same data set. Cross-validation does not ensure external validity beyond data similar to the analysed sample. In the case of MRI-based AD classification, there are two important issues

that can limit external validity. First, the patient population used for training a model might be too homogeneous. In chapter two and three we used data sets containing only diagnosed AD patients and healthy elderly controls. However, in a clinical setting, the patient group is more diverse, consisting of patients ranging from mild cognitive complaints to progressed dementia. Classification models trained on a too homogeneous sample might not generalise well to a clinically diverse sample. Second, MRI scans acquired at different MRI scanners can have different characteristics, due to different scanner vendors or due to different acquisition settings (Ewers et al., 2006; Takao et al., 2014; Zhu et al., 2011). The MRI scans used in chapter two and three were acquired at a single scan site, and the AD classification models might not generalise well to MRI scans from different scanners.

For these reasons, AD classification models should also be evaluated on external data. Therefore, in **chapter four** we evaluated external validity of our AD classification models on a multi-centre memory clinic data set. This data set was collected at four different scan sites, and consisted of clinically diverse patients. This evaluation is important for judging the clinical applicability of MRI-based AD classification models.

Classification methods

To create classification models, we used penalised logistic regression methods. These methods are well suited for MRI-based AD classification (Schouten et al., 2016; Teipel et al., 2015). They use penalties to hinder predictors entering the regression model. Predictors will only be included in the model if their added value for prediction outweighs the penalty. Consequently, only the most relevant predictors will enter the regression model (Friedman et al., 2010; Zou and Hastie, 2005). This is crucial for MRI-based classification studies, because in these studies the number of predictors largely outnumbers the number of subjects. This comes with the risk of including too many predictors and overfitting the model.

Penalised logistic regression methods only include main effects of the predictors, and they only estimate (log)linear effects between the predictors and the outcome variable. Some other methods also include interaction effects or non-linear effects. These methods can potentially include more information, which could increase accuracy. For example, random forests (Breiman, 2001a) include both non-linear effects and interaction effects, and they generally show high classification accuracy (Fernández-Delgado et al., 2014). An even more flexible method is deep learning. Deep learning is gaining increasing popularity, because it outperforms other classification methods

in a variety of tasks (LeCun et al., 2015), and more recently in medical image analysis (Esteva et al., 2017; Vieira et al., 2017). It is unclear whether deep learning is also beneficial for MRI-based AD classification. Due to its flexibility, it requires many subjects (e.g., thousands) to train the classification models, especially when there are many predictors. For MRI-based AD classification studies this is problematic, because they typically use many predictors while not many subjects are available. Nevertheless, deep learning was successfully applied to the ADNI data set for AD classification based on anatomical MRI scans. An accuracy of 99% was reported for classifying AD patients vs healthy elderly controls (Basaia et al., 2019). On the other hand, deep learning did not outperform penalised logistic regression in various MRI-based classification tasks, even though 9,300 MRI scans were used (Schulz et al., 2019). However, it was argued that this was due to using pre-engineered features, depriving deep learning of its main advantage - representation learning (Abrol et al. 2021).

In conclusion, it is yet unclear whether MRI-based AD classification could benefit from deep learning. However, it is certainly worth investigating more. Due to data sharing initiatives (e.g., ADNI; Jack et al., 2008) and publicly available large data sets (e.g., UK biobank; Alfaro-Almagro et al., 2018), it has become easier to apply deep learning to MRI data.

Model interpretation

In this thesis we have mainly focused on the accuracy of our classification models, but less so on the content of these models. To some extent, we interpreted the models, by summarising which features were most important. However, the interpretation of these models has some limitations. For example, we have forced sparsity in our models, which hinders potentially relevant features from entering the model. Also, when highly correlated features are all informative, only one of those features is necessary for the model. Consequently, training the model on slightly different training sets can result in largely different models. Furthermore, the effects of the features are conditional on the effect of all the other features in the model, and should therefore be interpreted in that relative context. When many features are involved, this further complicates interpretation.

A focus on classification accuracy usually comes at the cost of interpretability. This is referred to as the distinction between 'predicting' and 'explaining' (Breiman, 2001b; Shmueli, 2010). In statistical modelling one normally chooses to focus on either one of these two goals. We choose to focus on 'predicting', because classification accuracy is

most important for individual patient classification. However, preferably the predictive mode provides some explanation as well (Ribeiro et al., 2016). This is important because clinicians are more willing to accept a model-based diagnosis if they know what the model bases its diagnosis on. Therefore, one should make accurate models, and subsequently try to make these models interpretable (Breiman, 2001b). This could increase the likeliness of AD classification models to be used by clinicians.

Sample sizes

In this thesis we used medium sized samples (N = 42 for chapter 2, N = 250 for chapter 3, and N = 189 for chapter 4 and 5), which are considered small sample sizes for machine learning studies. Therefore, the uncertainty of the accuracy estimates is relatively high. For example, for sample sizes of 300 the discrepancy between accuracy measured by cross-validation and expected accuracy on new data is estimated between 4% and 6% (Varoquaux, 2018). Consequently, it is difficult to evaluate small effect sizes. In this thesis, this is not a problem for comparing the accuracies against chance, because many reported AUC values were over .80, which is much higher than chance level. However, comparisons of accuracies against each other mostly concern small differences. For example, in chapter 2 the AUC values of the anatomical MRI measures range from 0.67 to 0.94, and the AUC of the combined model is 0.98. In chapter 3 the AUC values of the resting state fMRI measures range from 0.51 To 0.85, and the AUC of the combined model is 0.85. The increases of the combined models over the single measures are small, and we can therefore not rule out that these increases are due to chance.

Still, we think it is useful that we have chosen these data sets for our analyses, because they have never been analysed in this context before. Most other AD classification studies have used the ADNI data set (Jack et al., 2008), and it is essential to use other data sets as well to not blind-stare on the results on the ADNI data set alone. Nevertheless, we acknowledge that our data sets are not large enough for firm conclusions, and we advise to replicate our results on a larger data set. This can be accomplished using either the ADNI data set, or large imaging population studies like the UK biobank (Alfaro-Almagro et al., 2018).

6.3 Future research

Early phase AD identification

In this thesis we included clinically diagnosed AD patients. This diagnosis requires at least cognitive and behavioural symptoms, which are known to start in a late phase of the disease (Jack et al., 2013). Consequently, the included AD patients are relatively progressed. Brain changes, as observed on MRI scans, are known to start already before cognitive and behavioural symptoms occur (Jack et al., 2013). Therefore, MRI scans might be used to diagnose AD in an earlier phase. In fact, anatomical MRI scans have been used to predict future conversion to AD in MCI patients (Davatzikos et al., 2011).

In line with this thesis, MCI to AD conversion prediction could be improved by calculating multiple anatomical MRI measures, or by using multimodal MRI models. Also, this would enable a comparison between the three MRI modalities on predictive accuracy of AD conversion. In this thesis, anatomical MRI was unquestionably superior to diffusion MRI and resting state fMRI for AD classification. However, for the prediction of AD conversion, this may be different since resting state fMRI has been put forward as a potential early AD marker (Buckner et al., 2005; Sheline and Raichle, 2013). However, for MCI to AD conversion prediction this has not yet been investigated thoroughly using a wide range of resting state fMRI measures in combination with machine learning. In chapter 3 we presented the methodology for such a study.

Taking into account the heterogeneity of clinical populations

In this thesis we set the focus on classifying AD vs. healthy elderly controls. In chapter 2 and 3 we developed AD classification models using only AD patients and healthy elderly controls. In chapter 4 we extended our focus, and applied these models to a memory clinic data set including SMC patients, MCI patients and AD patients. But still, this data set does not reflect the heterogeneity of clinical populations as seen in memory clinics. These populations consist of patients suffering from other dementia types as well, like frontotemporal dementia, Lewy body dementia or vascular dementia. In addition, some of these patients may experience their symptoms due to non-dementia causes, like depression, medication or alcohol abuse.

In order to be clinically useful, the full heterogeneity of clinical populations should be taken into account. Therefore, MRI-based AD classification should proceed from AD vs. control classification to classification of AD in a more diverse sample. This task is more complex, and requires larger data sets, including a wider variety of patient types.

Combining MRI measures with other biomarkers

The goal of our studies was to extensively evaluate MRI scans for AD classification. We did not study the inclusion of other types of biomarkers in our models. In future efforts, our models can be extended with other biomarkers that are known to be predictive for AD. In a first step (cerebrospinal fluid) CSF markers, PET scans and cognitive measures could be added, because these are proven to be discriminative for AD (Jack et al., 2016). In a second step one could add markers that are less discriminative, but could still be of additive value, like genetic markers (Genin et al., 2011) or metabolic markers (de Leeuw et al., 2017).

Scanner site differences

A challenge for MRI-based classification is the presence of scanner variability. It is known that technical variabilities across scan sites can have large effects on MRI scans. This is especially so for diffusion MRI scans (Zhu et al., 2011), but it is also the case for anatomical MRI (Takao et al., 2014) and fMRI (Feis et al., 2015). Consequently, MRI models based on scans from one scanner are not easily applied to MRI scans from another scanner. Partly, this problem can be solved by applying data harmonisation methods (Fortin et al., 2017; Fortin et al., 2018; Yu et al., 2018), as done in chapter 4 and 5 in this thesis. However, this only works for scanner sites for which MRI scans are available when fitting the MRI-based classification model. The created model will not take into account scan site effects of other scanner sites. Consequently, the use of the model is limited to MRI scans of scanners that were used when creating the model.

To facilitate clinical usefulness of MRI-based classification, care should be taken to standardise MRI scan protocols as much as possible, such that later data harmonisation is not or less needed. This is difficult though, because standardisation is only possible to a certain extent, and limited by the type of scanner hardware. Moreover, as shown by the UK Biobank data, even when using identical hardware for MRI acquisition at each site, site is still an important confounder (Alfaro-Almagro et al., 2021). This shows the persistent character of scan site effects, and indicates the necessity to incorporate this effect in MRI-based classification models.

6.4 Conclusion

This thesis includes two important results. First, AD classification accuracy increases when combining multiple types of measures from a single MRI scan. This is the case for both anatomical MRI scans and resting state fMRI scans. For anatomical MRI scans, this implies that a combination of measures may increase accuracy of AD diagnoses in clinical practice. Usually, anatomical MRI scans are only used to inspect specific AD characteristics like the size and shape of the hippocampus, or to rule out other causes for symptoms like a brain tumour or other dementia types. This thesis shows that more information can be extracted from a single anatomical MRI scan, and this may add to clinical AD diagnosis. The same applies to resting state fMRI scans. Combining multiple types of measures from these scans increases AD classification accuracy. This result is not of direct clinical importance, but it is important for further research. Resting state fMRI scans have potential for early AD diagnosis (Viviano & Damoiseaux, 2020), and to further investigate this potential we advise to combine multiple types of resting state fMRI measures.

Second, MRI-based AD classification models are not only useful for the data set that was used for training the model. They also generalise to some extent to heterogeneous patient populations, and to MRI scans acquired at different scan sites. This result is important, because MRI-based classification models are usually trained on a single-centre data set, including only AD patients and healthy elderly controls. The external validity of these models was yet unclear. This result adds to the translation of MRI-based AD classification to clinical practice.