Feature Selection to Forecast Cognitive Decline Using Multimodal Alzheimer's Disease Models

Benoît Sauty, Etienne Maheux, Stanley Durrleman

Feature Selection to Forecast Cognitive Decline Using Multimodal Alzheimer's Disease Models

Benoît Sauty, Etienne Maheux and Stanley Durrleman

Inria, Sorbonne Université, Institut du Cerveau - Paris Brain Institute - ICM

Abstract. Multimodal medical data (e.g. MR and PET imaging, CSF measurements, clinical assessments) reflect different aspects of Alzheimer's Disease, including early changes in brain structure and function that can occur before the onset of the associated cognitive impairment. We propose to use a feature selection method within a disease progression model to identify the combinations of imaging and non-imaging biomarkers across modalities that allow the best predictions of the cognitive decline. We first demonstrate that the chosen non-linear mixed-effect model outperforms all benchmarked methods in the TADPOLE challenge, with increasing performance as various modalities are added. We then introduce a controlled protocol to compare the added value of each feature for the forecast of cognition, at different stages of the disease, and for varying time-to-predictions. Notable findings include that the volumes of the ventricles are predictive features at the later AD stages but not at early stages, hippocampal volume is mostly important for intermediate stages and cognitively unimpaired subjects, cortical thickness of temporal cortex is most important for short-term predictions in AD patients at any stages, and cortical summaries of glucose and amyloid PET uptakes are only useful for intermediate AD stages. These conclusions may inform the design of efficient prognosis scores that have been shown to decrease sample size in clinical trials and can be adapted to the targeted disease stages and the trial duration.

Keywords: Multimodal modeling · Cognitive Forecast · Disease progression models · Alzheimer's Disease.

1 Introduction

Alzheimer's disease (AD) is a neurodegenerative disorder that affects millions of individuals worldwide. It is characterized by an accumulation of pathological proteins and the loss of neurons in the brain that lead to a gradual decline in cognitive function, including memory loss and difficulty with language, reasoning, and perception. Despite its high prevalence and significant personal and societal impact, current diagnostic and treatment strategies are limited. Data-driven forecasting methods and enrichment strategies have the potential to substantially decrease the sample size and inform the design of targeted clinical trials [4, 16]. In that spirit, disease progression models offer prognostic tools, identify the disease stage best suited for clinical intervention and predict individual progression patterns.

Forecasting cognitive decline in AD patients is challenging due to the complex nature of the disease and the variability in its progression. Traditional diagnostic tools, such as cognitive testing and clinical evaluations, may not be sensitive enough to detect early changes in brain function. Multimodal imaging data, including structural and functional imaging, may be valuable in detecting early changes in brain function that occur before the onset of cognitive symptoms. Longitudinal data, which capture changes in the brain over time, are particularly informative in understanding the progression of the disease. However, the combinations of modalities that best forecast the decline at different stages of the disease have not yet been elucidated.

Disease progression models for biomarkers and imaging features can be formulated using differential equations [9], time-to-event models for discretized abnormality thresholds [30], neural networks for scalar [19], imaging [6, 24] and multi-modal data [5], Gaussian processes [15] and various regression frameworks. Well suited to irregularly spaced or missing data, a very flexible approach is proposed with mixed-effects models, which account for both the average trajectory of the population, called the fixed-effects, and individual variations to that trajectory that account for inter-subjects variability called the random-effects. Early models used linear modeling [28] while non-linearities were later added with polynomial [29], logistic [11], and exponential [20] regressions. Such models can make prediction about future outcomes. However, little focus has been put on the selection of the most predictive features for cognitive forecast. In this study, we investigate the potential of multimodal longitudinal data in forecasting cognitive scores in AD patients using mixed-effects models, highlighting the most relevant combinations of modalities for each stage of the disease progression.

1.1 Related work

The automatic prediction of the patient's current diagnosis from multimodal data at the time of the acquisition has been extensively studied, with an emphasis on T1-MRI [7, 14] and more broadly neuroimaging data [3, 21]. The study of future conversion from MCI to AD, which is a more challenging and clinically relevant task, has also been tackled with a variety of approaches. [2] proposes a quantitative review of such studies to exhibit the best-performing methods and set of modalities for this particular task. It concludes that for short-term predictions (<2y), predicting the patient does not change clinical status works as well as prediction models and that using FDG-PET data and cognition improves prediction while T1-MRI does not. This task, however, depends on the current clinical diagnostic practices that may not be reproducible among clinicians and countries and provides little dynamic information about the cognitive changes of patients. Predicting the cognitive outcomes at varying time-points in the future is of greater interest for clinical decision support systems. The most notable step in this direction was the introduction of The Alzheimer's Disease Prediction Of Longitudinal Evolution (TADPOLE) challenge [18] that uses data from the Alzheimer's Disease Neuroimaging Initiative (ADNI) to provide a benchmark to compare the forecasting performance of more than 90 multimodal

progression models. However, amongst the official submissions, no method significantly outperformed a simple linear mixed model for the prediction of cognitive decline [17]. The challenge has remained open and new methods have since provided better estimates, although the added value of each selected feature has not been assessed. In this work, we:

- demonstrate that a mixed-effect logistic model outperforms all benchmarked methods in the TADPOLE challenge, with increasing performance as imaging data from various modalities are added,
- describe a protocol to fairly compare the added value of each imaging feature for the forecast of cognition in different scenarios depending on the targeted disease stage and the time to prediction,
- exhibit, for each prediction task (a given cohort and a given prediction horizon), the features that are most useful or detrimental for the forecast.

2 Methodology

2.1 Longitudinal geometric model

Mixed-effects models describe each patient's progression over time as a variation – the random-effects – around the average population trajectory – the fixed-effects. As suggested in [27], the inter-subjects variability is parametrized as the combined effects of an individual onset age τ and pace of decline α – for the temporal variability – and a spaceshift w – for the intrinsic differences that are independent of time. An affine time-warp $t \mapsto \alpha(t - \tau)$ maps the age at visits to a pathological age on the common timeline of the average trajectory. These individual parameters form the random-effects, which we model with Gaussian priors. The reference positions, velocities, and time that define the average trajectory, together with the variances of the random-effects and residual noise, form the fixed-effects of the model. These models have also been demonstrated to generalize well to different databases [16] and reach test-retest noise levels for the modeling of cognitive scores and regional imaging features [12, 23, 25], hinting that the models could not be improved without over-fitting.

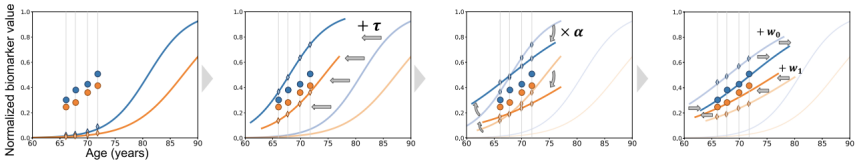


Fig. 1: Illustration of the random-effects "transforming" the average trajectory to account for the variability between patients, for a model with two features (blue and orange). Dots are the actual measurements and plain lines are the modeled trajectory. More details on this model are provided in [27]. Figure courtesy of [12].

2.2 Calibration and goodness of fit

These longitudinal models are part of a family of geometric models [26] that can be calibrated using a Maximum a Posteriori estimation of both random and

fixed-effects, using MCMC-SAEM procedure in which the estimation step of an Expectation-Maximization algorithm is replaced by a stochastic approximation. See [1, 13] for details on the procedure, description of the complete likelihood and proof of convergence and stability. Once a model is calibrated and the fixed-effects are estimated for a given cohort – a set of patients following inclusion criteria, used as training data – random-effects can be learned for unseen patients using a gradient-descent on the likelihood. All analysis is done using the open-source Leaspy software with a 2.3GHz CPU. Fitting one model and making predictions takes between 20" in a univariate setting and 10' for a multivariate model with 10 features.

2.3 Data processing

All analysis in this study is done on publicly available data from the ADNI database (adni.loni.usc.edu), which provides repeated MRI, FDG-PET and AV45-PET scans, as well as various clinical assessments for AD patients. Firstly, we demonstrate the usefulness of using multimodal data on the well-studied TADPOLE benchmark [18]. This challenge uses a subset of ADNI database to make monthly forecasts of the ADAS-Cog score that can be compared with the actual measurements. We use the processed features from the challenge data for the volumes (hippocampi and ventricles), thicknesses (enthorinal and fusiform cortex), and brain-average PET uptake (FDG and AV45). This challenge allows demonstrating that a model performs well on a real life dataset, but does not allow describing the most predictive features. In practice, the training cohort is very heterogeneous (1667 patients with 8068 visits total, including 615 AD, 594 MCI and 458 healthy patients), and the prediction tasks are not designed with a fixed amount of training visits per patients (7.6 $\pm$ 3.9 visits), nor for a fixed time-to prediction (2.7 $\pm$ .8 years, with 219 predictions to make).

We thus select four homogeneous cohorts from the ADNI database using the baseline cognitive score and amyloid status based on Aβ42 measurements in CSF or average AV45-PET uptake [8]. Cognitively unimpaired patients are defined as having less than median score (<14/85 for ADAS-Cog13) and mildly (resp. strongly) impaired patients as being below (resp. above) the third quantile (< and >23/85 for ADAS-Cog13) of the overall distribution. We refer to those cohorts as A-/C-, A+/C-, A+/C+ and A+/C++. Subjects with A- and C+ or C++ are few and likely to be suffering from a variety of conditions, so they are not considered in our study. We process raw T1-MRI, FDG-PET and AV45-PET images using the Clinica software [22] that wraps common image processing pipelines to extract volumes (hippocampi, ventricles and striatum), average thickness of the temporal cortex, average FDG uptakes (temporal and occipital lobes) and brain-average AV45 uptake. Volumes are normalized by intracranial volume, and we also use p-tau protein concentration from CSF.

In practice, biomarkers have long been hypothesized to follow sigmoid shapes [10] so we elect to use logistic trajectories and features should be normalized to the range $[0, 1]$. We thus discard outliers using the so-called three sigma rule for features that are not naturally bounded and add a min-max normalization

within each cohort, providing realistic asymptotes. Since cortical thicknesses, brain volumes and FDG uptakes decrease over time, we flip data around .5, using the rotation $x \mapsto 1-x$, in order to ensure increasingness, which is required for logistic modeling. For each cohort, we split the dataset in 5-folds and make predictions based on the first two visits for all patients of one fold, using the calibration made on all visits from the other folds.

	A+/C++	A+/C+	A+/C-	A-/C-
$N_{patients}/N_{visits}$	349 / 1621	253 / 1431	321 / 1788	428 / 2596
Age at baseline (y)	74.0 ± 7.6	74.4 ± 6.8	72.5 ± 6.7	71.1 ± 6.8
Average follow-up (y)	2.3 ± 1.8	3.5 ± 2.5	4.2 ± 3.0	4.8 ± 3.1
Sex (M/F)	193/156	173/80	135/186	203/225
ADAS-cog baseline	28.5 ± 7.5	17.1 ± 4.4	9.1 ± 3.8	8.2 ± 3.7
ADAS-cog last visit	39.0 ± 12.6	23.5 ± 11.8	12.4 ± 10.5	9.1 ± 5.1
N_{scans} (MRI/FDG/AV45)	1097/523/362	933/479/318	1018/474/583	1510/629/915

Table 1: Demographics for the four selected cohorts.

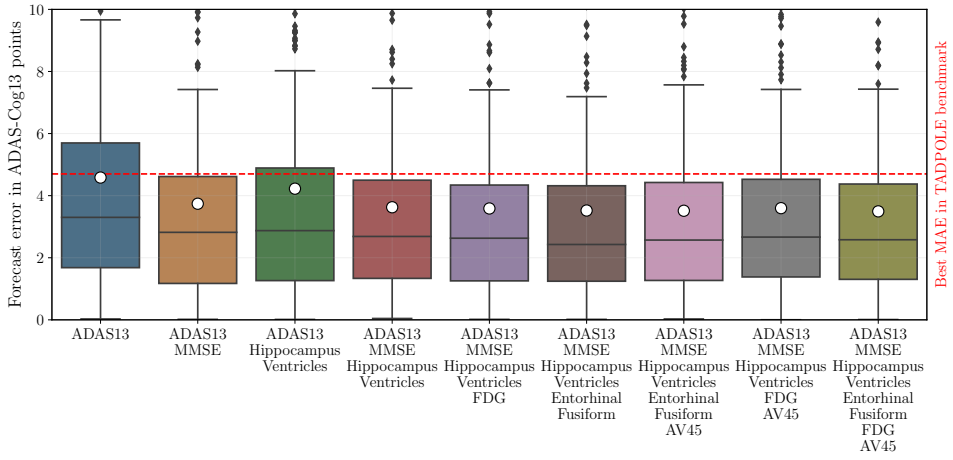
2.4 Statistical testing of the usefulness of each feature

For a given cohort, once a model is calibrated for each possible combination of features and predictions are made using the first two visits, we rank the models by quality of prediction using R2 scores for a given time horizon. We then compare the number of occurrences of each feature in the best and the worst-performing models. We use 8 features, which yield a total of 256 models, and compare the top and bottom 32 models. Under the null hypothesis that a feature does not improve nor worsen prediction, the number of models that use it should follow a binomial distribution under the normal approximation. We thus proceed to Welch t-tests that compare the proportion of models that use the feature in the best and worst performing models, providing a proxy for how much a feature improves or hurts predictive abilities.

3 Experiments and results

3.1 Benchmark on the TADPOLE challenge

As displayed in Fig. 2, we evidence that multivariate models perform better, both when adding a second cognitive score (MMSE) or imaging data, with the best-performing model being the one that features all modalities. However, it is likely that all modalities do not convey relevant information about neurodegeneration at the same disease stage, and the added value of some effects are mitigated by the heterogeneity of the cohort. Besides, some patients – especially the ones that are most healthy and decline the least – have a lot more visits to train on, driving the average error down. In order to rigorously assess the added value of each modality for the forecast of cognitive decline at the different stages of AD, a more controlled protocol is necessary.



MAE	4.59	3.74	4.22	3.63	3.59	3.52	3.51	3.60	**3.49**
R2	0.53	0.66	0.50	0.67	0.67	0.68	**0.70**	0.68	**0.70**

Fig. 2: Distribution of prediction errors for the TADPOLE challenge. Circles are the MAE for each model. Below are the first and second order statistics that measure MAE and explained variance. Models without any cognitive features display MAE of 7.0 and more on a scale of 85, so we only display models that include cognition.

3.2 Experiments on homogeneous cohorts

Fig. 3 represents the average progression of the multivariate model with all the selected features for each cohort. Fig. 4 compares the best performing multivariate model to a univariate model that only uses the cognitive score and to a constant prediction model. Predictions are stratified based on the time to the last visit used for training (rounded to the unit). Fig. 5 displays, for each cohort and time horizon, the features that increase or decrease prediction accuracy.

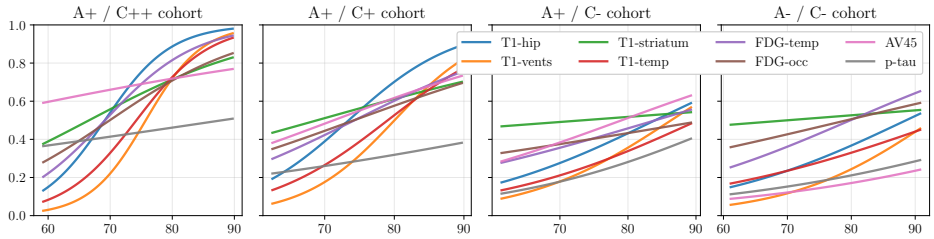
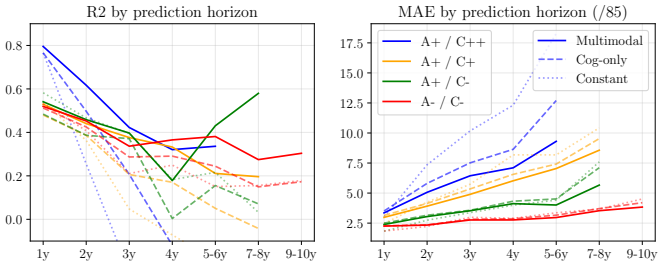


Fig. 3: Average progression for the 4 learned models and corresponding cohorts. T1-MRI features are the volumes of hippocampi, ventricles, striatum, and thickness of the temporal cortex. FDG-PET features are the average FDG uptake in temporal and occipital cortex, and AV45 is the average AV45 brain uptake. p-tau is the concentration of phosphorylated tau protein in the CSF.



	A+/C++	A+/C+	A+/C-	A-/C-
1y	382	298	288	385
2y	215	206	263	386
3y	93	122	124	181
4y	54	88	127	206
5-6y	53	95	177	268
7-8y	22	48	109	166
9-10y	3	18	24	56

Fig. 4: Best predictions across all models for each prediction horizons (y. is years). Amount of predicted visits are shown in the right table. Predictions task with sample size <30 visits are discarded.

For predictions at less than 2 years, the constant prediction outperforms progression models for the C- cohorts because of the very little change happening over such a period of time, which concurs with [2, 17]. For longer time-to-prediction, multimodal models always outperform univariate models, even more as the patients are at an advanced disease stage. The R2 score is more informative of the quality of the model as it penalizes large errors more and represents the ability to explain the variability in the progression of patients, effectively separating the slow and the fast progressors, while the MAE can be very small for patients that do not progress a lot, while providing no specific information about each patient. For instance, the lower MAE for A-/C- and A+/C- cohorts correspond to similar R2 scores as the A+/C+ (see Table 1 for the average progression of each cohort).

In the most advanced stages of the disease (A+/C++), ventricles volumes and thickness of temporal cortex are the most useful features. In the A+/C+ cohort, PET and CSF data are significantly more retained in the best models, highlighting the usefulness of multimodality during the "core progression" of AD. In the A+/C- cohort, multimodality yields less information about the progression, and only the thickness of temporal cortex and ptau concentration significantly improve the forecast. The healthy A-/C- cohort, on the other hand, only uses the volume of the hippocampus to predict the (small) progression of patients. Interestingly, the ventricles are very useful for advanced AD patients but get progressively detrimental as cohorts are selected at earlier stages of the disease. It is also striking to see that gathering multimodal imaging from patients, at the cost of great expenses and inconveniences for the patient, does not yield much improvement for the early stages of the disease, while it is most useful for A+/C+ patients.

To validate those findings, we run the same pipeline with a standard linear mixed-effects model for prediction and recover the same feature selection patterns, except for the T1-striatum being more retained in the best models for all horizons in the A-/C- cohort. Besides, we decided to rank models by R2 scores for statistical testing, but it could be argued that progression models should

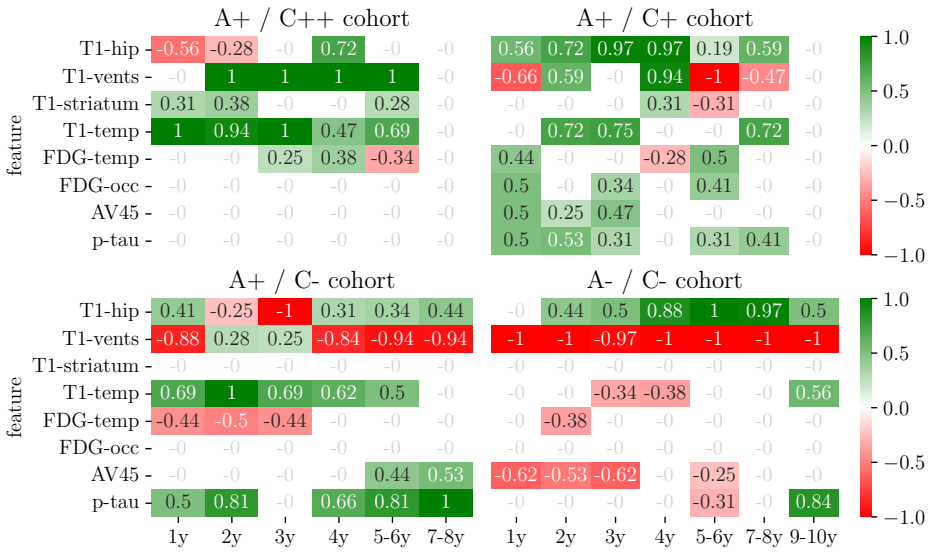


Fig. 5: Difference in proportion of models that feature each outcome in the best and worst models for a given cohort and time to prediction. 1 means the feature is selected in all the best models and none of the worst, while -1 is the opposite. Non-significant differences are not displayed. Green means that features are significantly more present in the models with best R2 scores, and red means significantly less.

minimize the mean error instead of the tail of the distribution. The same figure obtained by maximization of the MAE score yields the exact same trend, with only slight variations in the t-values. Besides, selecting 16, 32, 64 or 128 of the best/worst models do not change the trends of usefulness of features.

4 Discussion

Several methodological limitations need to be acknowledged. Firstly, as is typical of neuroimaging databases, the visits we use do not all feature each selected modality, which can in turn mitigate the added-value of choosing a feature in our model (see Table 1 for information on the amount of scans for each modality). Secondly, it must be noted that this study used a selection of 8 features that are plausibly associated with AD, but further work that scales this statistical pipeline up for a wide variety of available features without pre-selection is underway. It must also be noted that patients with very long follow-up periods are usually less impaired, which can bias the prediction scores and selected features for the long-term predictions.

As [16] demonstrates, improving the forecast of cognitive decline allows to drastically decrease sample size for clinical trials and could help provide personalized care and prognosis. Understanding the features that should be used and are worth acquiring for each population is a crucial first step in that direction.

References

1. Allassonnière, S., Kuhn, E., Trouvé, A.: Construction of bayesian deformable models via a stochastic approximation algorithm: a convergence study. Bernoulli **16**(3), 641–678 (2010)
2. Ansart, M., Epelbaum, S., Bassignana, G., Bône, A., Bottani, S., Cattai, T., Couronné, R., Faouzi, J., Koval, I., Louis, M., et al.: Predicting the progression of mild cognitive impairment using machine learning: A systematic, quantitative and critical review. Medical Image Analysis **67**, 101848 (2021)
3. Arbabshirani, M.R., Plis, S., Sui, J., Calhoun, V.D.: Single subject prediction of brain disorders in neuroimaging: Promises and pitfalls. Neuroimage **145**, 137–165 (2017)
4. Ballard, C., Atri, A., Boneva, N., Cummings, J.L., Frölich, L., Molinuevo, J.L., Tariot, P.N., Raket, L.L.: Enrichment factors for clinical trials in mild-to-moderate alzheimer's disease. Alzheimer's & Dementia: Translational Research & Clinical Interventions **5**, 164–174 (2019)
5. Couronné, R., Louis, M., Durrleman, S.: Longitudinal autoencoder for multi-modal disease progression modelling (Apr 2019), preprint
6. Cui, R., Liu, M.: RNN-based longitudinal analysis for diagnosis of Alzheimer's disease. Computerized Medical Imaging and Graphics **73**, 1–10 (2019)
7. Falahati, F., Westman, E., Simmons, A.: Multivariate data analysis and machine learning in alzheimer's disease with a focus on structural magnetic resonance imaging. Journal of Alzheimer's disease **41**(3), 685–708 (2014)
8. Hansson, O., Seibyl, J., Stomrud, E., Zetterberg, H., Trojanowski, J.Q., Bittner, T., Lifke, V., Corradini, V., Eichenlaub, U., Batrla, R., et al.: Csf biomarkers of alzheimer's disease concord with amyloid-β pet and predict clinical progression: a study of fully automated immunoassays in biofinder and adni cohorts. Alzheimer's & dementia **14**(11), 1470–1481 (2018)
9. Ito, K., Corrigan, B., Zhao, Q., French, J., Miller, R., Soares, H., Katz, E., Nicholas, T., Billing, B., Anziano, R., et al.: Disease progression model for cognitive deterioration from Alzheimer's Disease Neuroimaging Initiative database. Alzheimer's & Dementia **7**(2), 151–160 (2011)
10. Jack Jr, C.R., Knopman, D.S., Jagust, W.J., Petersen, R.C., Weiner, M.W., Aisen, P.S., Shaw, L.M., Vemuri, P., Wiste, H.J., Weigand, S.D., et al.: Update on hypothetical model of alzheimer's disease biomarkers. Lancet neurology **12** (2013)
11. Jedynak, B., Lang, A., Liu, B., Katz, E., Zhang, Y., Wyman, B., et al.: A computational neurodegenerative disease progression score: method and results with the Alzheimer's disease neuroimaging initiative cohort. Neuroimage **63** (2012)
12. Koval, I., Bône, A., Louis, M., Lartigue, T., Bottani, S., Marcoux, A., Samper-Gonzalez, J., Burgos, N., Charlier, B., Bertrand, A., et al.: Ad course map charts alzheimer's disease progression. Scientific Reports **11**(1), 1–16 (2021)
13. Kuhn, E., Lavielle, M.: Coupling a stochastic approximation version of em with an mcmc procedure. ESAIM: Probability and Statistics **8**, 115–131 (2004)
14. Leandrou, S., Petroudi, S., Kyriacou, P.A., Reyes-Aldasoro, C.C., Pattichis, C.S.: Quantitative mri brain studies in mild cognitive impairment and alzheimer's disease: a methodological review. IEEE reviews in biomedical engineering **11** (2018)
15. Lorenzi, M., Filippone, M., Frisoni, G.B., Alexander, D.C., Ourselin, S., Initiative, A.D.N., et al.: Probabilistic disease progression modeling to characterize diagnostic uncertainty: application to staging and prediction in alzheimer's disease. NeuroImage **190**, 56–68 (2019)

16. Maheux, E., Koval, I., Ortholand, J., Birkenbihl, C., Archetti, D., Bouteloup, V., Epelbaum, S., Dufouil, C., Hofmann-Apitius, M., Durrleman, S.: Forecasting individual progression trajectories in alzheimer's disease. Nature Communications **14**(1), 761 (2023)
17. Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Eshaghi, A., Toni, T., et al.: The alzheimer's disease prediction of longitudinal evolution (tadpole) challenge: Results after 1 year follow-up. arXiv preprint arXiv:2002.03419 (2020)
18. Marinescu, R.V., Oxtoby, N.P., Young, A.L., Bron, E.E., Toga, A.W., Weiner, M.W., Barkhof, F., Fox, N.C., Klein, S., Alexander, D.C., et al.: Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease. arXiv preprint arXiv:1805.03909 (2018)
19. Nguyen, M., He, T., An, L., Alexander, D., Feng, J., Yeo, B., et al.: Predicting Alzheimer's disease progression using deep recurrent neural networks. NeuroImage **222**, 117203 (2020)
20. Raket, L.L.: Statistical disease progression modeling in Alzheimer disease. Frontiers in big Data **3** (2020)
21. Rathore, S., Habes, M., Iftikhar, M.A., Shacklett, A., Davatzikos, C.: A review on neuroimaging-based classification studies and associated feature extraction methods for alzheimer's disease and its prodromal stages. NeuroImage **155** (2017)
22. Routier, A., Burgos, N., Díaz, M., Bacci, M., Bottani, S., El-Rifai, O., Fontanella, S., Gori, P., Guillon, J., Guyot, A., et al.: Clinica: An open-source software platform for reproducible clinical neuroscience studies. Frontiers in Neuroinformatics **15**, 689675 (2021)
23. Sauty, B., Durrleman, S.: Impact of sex and apoe-ε4 genotype on patterns of regional brain atrophy in alzheimer's disease and healthy ageing (2022)
24. Sauty, B., Durrleman, S.: Progression models for imaging data with longitudinal variational auto encoders. In: Medical Image Computing and Computer Assisted Intervention–MICCAI 2022: 25th International Conference, Singapore, September 18–22, 2022, Proceedings, Part I. pp. 3–13. Springer (2022)
25. Sauty, B., Durrleman, S.: Impact of sex and apoe-ε4 genotype on regional brain metabolism in alzheimer's disease. In: 2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI) (2023)
26. Schiratti, J.B., Allassonniere, S., Colliot, O., Durrleman, S.: Learning spatiotemporal trajectories from manifold-valued longitudinal data. In: Neural Information Processing Systems. No. 28 in Advances in Neural Information Processing Systems, Montréal, Canada (Dec 2015)
27. Schiratti, J.B., Allassonnière, S., Colliot, O., Durrleman, S.: A bayesian mixed-effects model to learn trajectories of changes from repeated manifold-valued observations. The Journal of Machine Learning Research **18**(1), 4840–4872 (2017)
28. Verbeke, G.: Linear mixed models for longitudinal data. In: Linear mixed models in practice. Springer (1997)
29. Wu, H., Zhang, J.: Local polynomial mixed-effects models for longitudinal data. Journal of the American Statistical Association **97**(459), 883–897 (2002)
30. Young, A., Oxtoby, N., Daga, P., Cash, D., Fox, N., Ourselin, S., Schott, J., Alexander, D.: A data-driven model of biomarker changes in sporadic Alzheimer's disease. Brain **137**(9), 2564–2577 (2014)