PREDICTION OF ALZHEIMER'S DISEASE FROM ASYMPTOMATIC STAGES USING MACHINE LEARNING MODELS

Shengnan Huang,[1] Luyuan Qi,[2] Billy Amzal [2]
[1]Analytica Laser – a Certara company, New York, USA
[2]Analytica Laser – a Certara company, Paris, France

- Current research in Alzheimer's disease (AD) focuses on detecting the disease from asymptomatic stages[1].
- Machine learning models for predicting time-to-event (i.e. time-to-diagnosis of AD), a bivariate outcome, are not extensively studied due to censoring information, despite a number of classification models were built considering the diagnosis of AD as an univariate outcome.
- We comparatively explored the models predicting long-term risk of AD, i.e. time-to-diagnosis of AD, from asymptomatic stages using real-world data from Alzheimer's Disease Neuroimaging Initiative (ADNI).

OBJECTIVES

- To explore the capabilities of machine learning survival models in predicting time-to-diagnosis of AD.
- To explore potential clinical use of these models.

METHODS

Population

- We downloaded the dataset ended on Sep 9th 2018 from ADNI.
- 527 subjects who were cognitively normal at baseline (i.e., the first visit), and had at least one follow-up visit after baseline were selected.

Outcome

- Time-to-diagnosis of AD, defining the diagnosis of AD as the first mild cognitive impairment (MCI) or dementia due to AD, whichever happened first.
- Subjects free of MCI or dementia due to AD till the end of follow-up were set as censored.

Covariates

- Two sets of covariates were used in modeling (Table 1).
  - Covariate set 1 only included static covariates.
  - Covariate set 2 further incorporated the changing rates of time-dependent covariates.

Table 1. Covariates used in the machine learning models*

Covariate set 1	Covariate set 2
• Demographic variables • Gene marker • Neuropsychological tests** at baseline	• Demographic variables • Gene marker • Neuropsychological tests at baseline • Neuropsychological tests changing rates***

** All covariates used in modeling were non-missing.*
*** 20 neuropsychological tests: Clinical Dementia Rating (CDR): Community affairs, Home and hobbies, Judgment and problem solving, Orientation; CDR sum of boxes (CDRSB); Mini Mental State Exam (MMSE); Alzheimer's Disease Assessment Scale-Cognitive total 11 and 13 (ADAS11, ADAS13); Logical memory delayed recall (LDEL); Trail-making test-part B (TRABSCOR); Boston naming test (BNT); Functional Assessment Questionnaire (FAQ); Category fluency-animal (CATANIMSC); Geriatric Depression (GDTOTAL); Modified Hachinski (HMSCORE); Rey Auditory Verbal Learning Test (RAVLT): Immediate, Learning, Forgetting, Percent Forgetting*
**** Changing rates = (test value from the last available non-missing record – test value at baseline) / follow-up time between these two records.*

Modelling strategy

- 80% randomly selected subjects were used for training while the remaining 20% were used for testing.
- Survival tree and random forest models were trained with two covariate sets, respectively.
- Integrated Brier Score (IBS) was used to compare model performance.
- The models were fitted using the {partykit} {randomForestSRC} R packages, R version 3.6.0.

RESULTS

Survival tree model

- Sub-groups and corresponding Kaplan - Meier (KM) curves were generated based on covariates values (Figure 1). Median diagnosis-free days were provided per subgroup (Table 2). From Figure 1, we can see that no patient developed the diagnosis of AD during follow-up in Node 4.
- IBS was lower when using Covariate set 2, showing that neuropsychological tests changing rates played an important role in prediction and increased the model performance (Table 3).

Random forest model

- The KM curve of each subject type was produced (Figure 2).
- Covariate influence on the probability of developing the diagnosis of AD was obtained (Figure 3). For example, as time went, the homozygote carriers (APOE4 = 2) had higher probabilities to develop diagnosis of AD, compared to heterozygote carriers (APOE4 = 1) and non-carriers (APOE4 = 0).
- Adding information on neuropsychological tests changing rates increased model performance measured by IBS (Table 3).
- Random forest model outperformed survival tree model under Covariate set 2 on both training and test sets (Table 3).

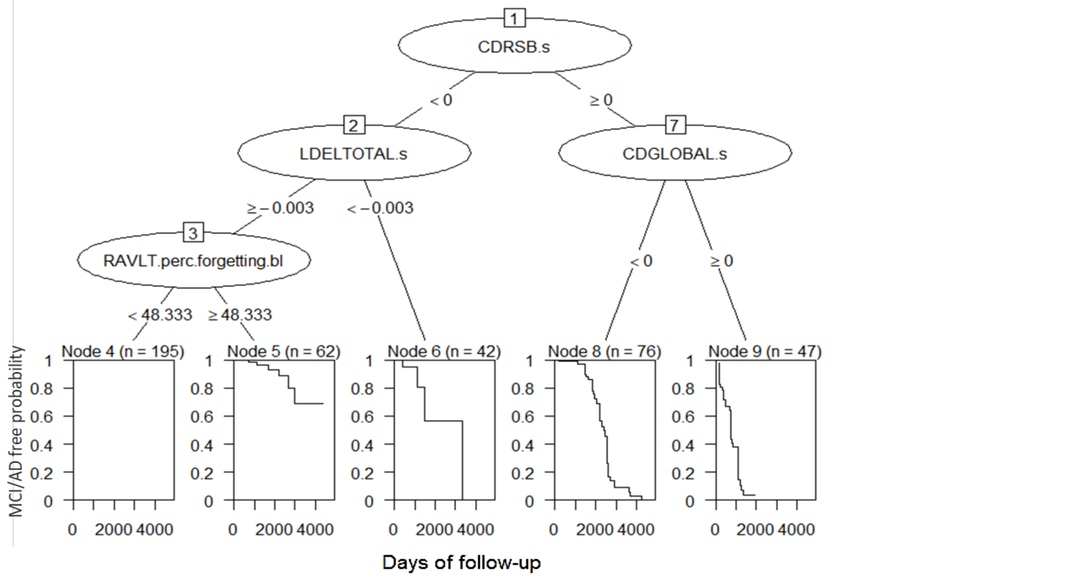

Figure 1. The trained survival tree model together with the predicted KM curve for each node.
.bl represents baseline value while .s represents changing rate of a covariate.
CDGLOBAL: CDR Global; LDELTOTAL: Logical memory delayed recall total score; Other covariates are the same as in Table 1.

Table 2. Median diagnosis-free time (days after baseline) provided by the survival tree model for each subgroup

	Node 4	Node 5	Node 6	Node 8	Node 9
Median diagnosis free days	> 4929	> 4929	3355	2408	745

Table 3. IBS of survival tree and random forest models on both training and test sets by different covariate sets

Covariate set	Dataset	Trained model	
		Survival tree*	Random forest
Covariate set 1	Training set	0.141	0.118
	Test set	0.143	0.164
Covariate set 2	Training set	0.0957	0.026
	Test set	0.1152	0.059

**The IBS for survival tree was calculated as the mean of IBS for each node.*

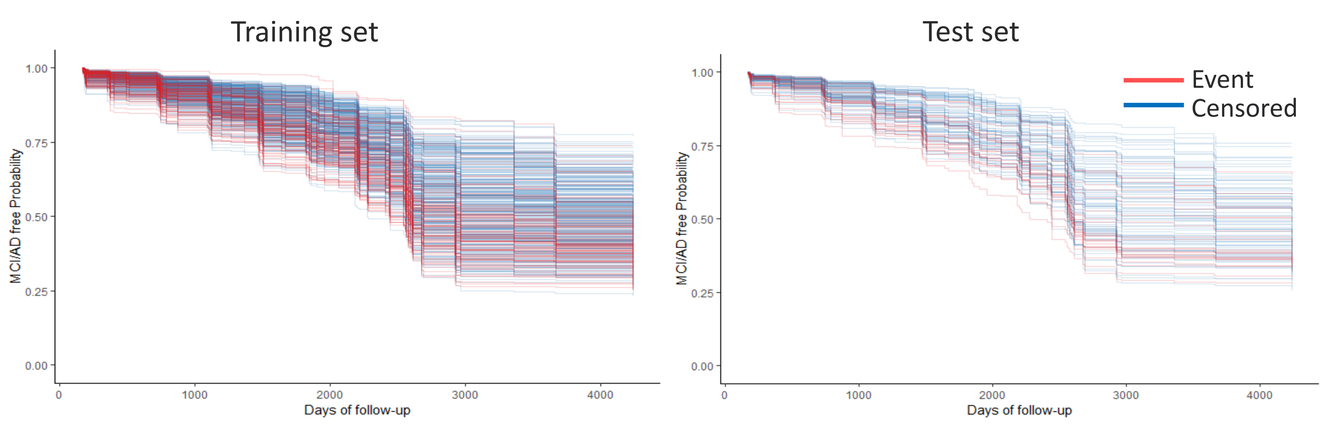

Figure 2. Predicted KM curve for each participant type given by random forest model.

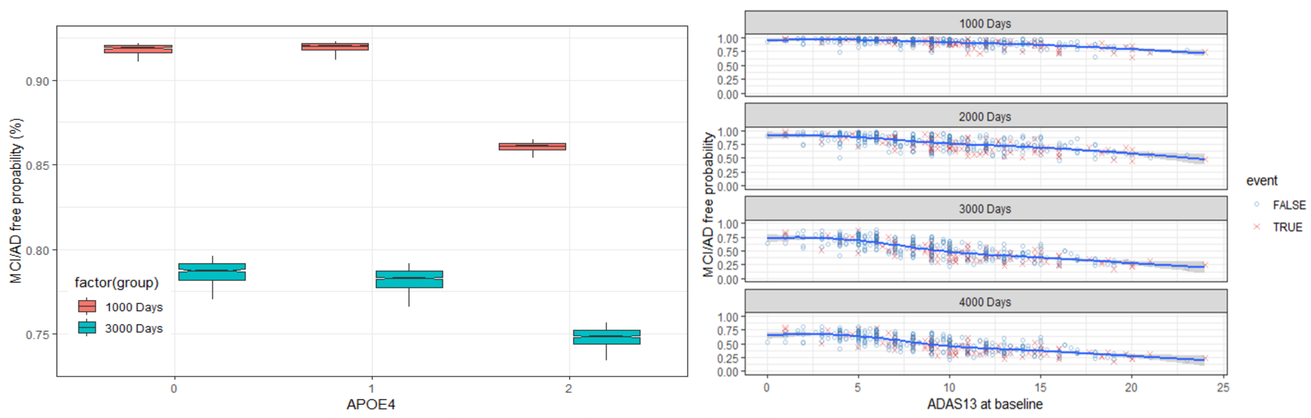

Figure 3. Covariate influence information from random forest.

LIMITATIONS AND FURTHER WORK

- The models explored in this study offered different perspectives and the performances varied. Thus, clinical consideration should be involved when fitting specific models.
- Random forest model had a better performance, but it's hard to get a reasonable single tree out from the random forest model.
- The model performance on test sets suggested possible overfitting, thus models need to be further refined (Table 3).

CONCLUSION

- The individual-level real-word datasets from asymptomatic stages onwards would be crucial for predicting AD progression.
- The time-to-event data analysis requires dedicated machine learning models.

REFERENCES

[1]Chong MS and Sahadevan S. Lancet Neurol. 2005;4:576–9.

CERTARA | EVIDENCE & ACCESS