

PREDICTION OF DEMENTIA SUBTYPES BY MACHINE LEARNING: FEASIBILITY OF ERROR CONTROL

Ye Tian[†], Henry Rusinek[‡], Alok Vedvyas[‡] and Yang Feng[§]

[†]Department of Statistics, Columbia University [‡]Grossman School of Medicine, New York University [§]School of Global Public Health, New York University

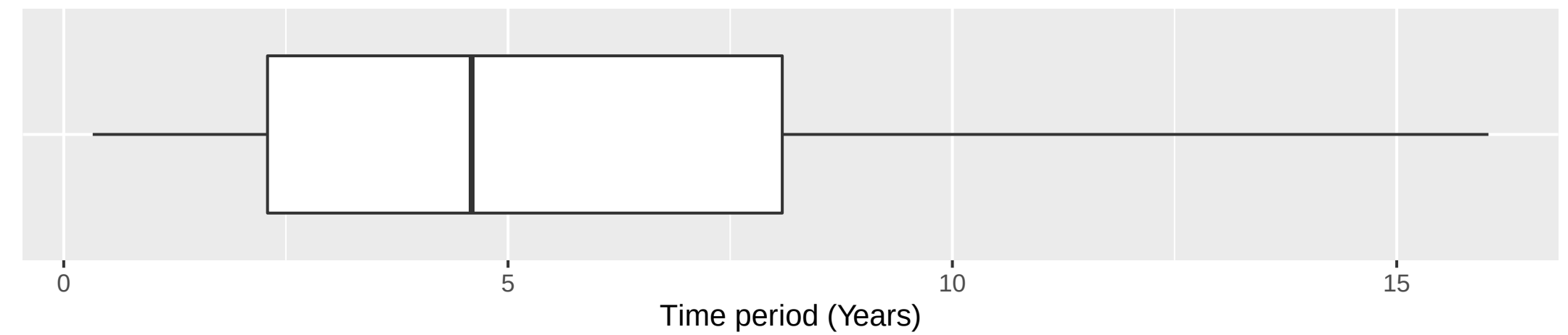
Introduction

Background

Accurate prediction of dementia subtypes is important in geriatric decision-making and in designing clinical trials. We used machine learning and a novel error control algorithm to analyze longitudinal data collected by the National Alzheimer Coordinating Center (NACC) from June 2005 to November 2021. Our goals were

- (1) Predicting the conversion to specific dementia subtypes in patients who were dementia-free at baseline;
- (2) Identifying factors that best contribute to the prediction;
- (3) Controlling the prediction error rate on specific dementia subtypes to help design clinical trials.

Our longitudinal multi-class analysis is an important extension of prior efforts to model conversion to dementia [2]. The study includes all available NACC data, covering time periods from baseline to the most recent diagnosis (See the box plot). Previous models predict risk of dementia in older adults only for a fixed period [1, 3, 4]



Method

We studied the 13,243 NACC participants who had normal clinical etiologic diagnosis results at their first visit. Among these patients, 10,679 remained dementia-free (NL). Remaining participants converted to Alzheimer’s disease (AD) 1,793, Lewy body disease (LBD) 153, and Vascular dementia (VD) 290 as the primary diagnosis at their last exam. Predictors with over 10% missing values were removed. For remaining predictors (<10% missing) unavailable data were imputed. The final set of 249 predictors included clinical, family history, and behavioral features. Using 500 cross-validation rounds, cases were randomly split into training (90%) and testing (10%) subsets. To accommodate unbalanced distribution of dementia subtypes, the same number (153) of cases per subtype were selected for model fitting. Penalized multinomial regression was fitted to predict the diagnosis result of their last visit.

Key clinical metrics and risk factors

The most important features to distinguish AD and NL

Features	Coefficients	Odds ratios	Descriptions
BIRTHYR	-0.5951	0.5515	Subject’s year of birth
NACCDAYS	0.2110	1.2349	Days from initial visit to most recent visit
ALCOCCAS1	-0.1049	0.9004	The subject consumed any alcohol in the past three months.
COGSTAT1	-0.0850	0.9185	The subject’s cognitive status is deemed better than normal for age
DECIN1	0.0739	1.0767	The co-participant reports a decline in subject’s memory.
VEG	-0.0590	0.9427	Vegetables - Total number of vegetables named in 60 seconds
ALCFREQ4	-0.0544	0.9471	The subject have at least one drink of any alcoholic beverage daily or almost daily during the past three months.
ALCFREQ3	-0.0530	0.9484	The subject have at least one drink of any alcoholic beverage a few times a week during the past three months.

EVENTS1	0.0390	1.0398	The subject have any difficulty with keeping track of current events but does it by self.
NACCFADM1	0.0371	1.0378	Evidence of a dominantly inherited AD mutation exists in this family.

The most important features to distinguish LBD and NL

Features	Coefficients	Odds ratios	Descriptions
PD1	2.5347	12.6126	Recent/Active Parkinson’s disease (PD)
BEREM1	0.9630	2.6195	REM sleep behavior disorder
PDOTHR1	0.6969	2.0075	Recent/Active other parkinsonian disorder
HALL1	0.6736	1.9613	Hallucinations in the last month
NACCBEHF8	0.6091	1.8388	The predominant symptom that was first recognized as a decline in the subject’s behavior - REM sleep behavior disorder
BIRTHYR	-0.5718	0.5645	Subject’s year of birth
PARK1	0.3411	1.4065	Present Parkinson’s disease
NACCDAYS	0.2109	1.2348	Days from initial visit to most recent visit
NACCLIVS2	0.1963	1.2169	The subject lives with spouse or partner
SEX2	-0.1728	0.8413	Subject’s sex - Female

The most important features to distinguish VD and NL

Features	Coefficients	Odds ratios	Descriptions
BIRTHYR	-0.5735	0.5635	Subject’s year of birth
CBSTROKE2	0.5213	1.6842	Remote/Inactive stroke
CBSTROKE1	0.2807	1.3241	Recent/Active stroke
NACCDAYS	0.2109	1.2348	Days from initial visit to most recent visit
MINTTOTS_abn1	0.1569	1.1699	Abnormal total score in Multilingual Naming Test (MINT)
CBTIA1	0.1453	1.1564	Recent/Active Transient ischemic attack (TIA)
CBTIA2	0.1443	1.1552	Remote/Inactive Transient ischemic attack (TIA)
NACCNHR2	0.1347	1.1442	Black or African American
HYPERTEN1	0.1318	1.1409	Recent/Active hypertension
ALCOCCAS1	-0.1052	0.9001	The subject consumed any alcohol in the past three months.

Key findings:

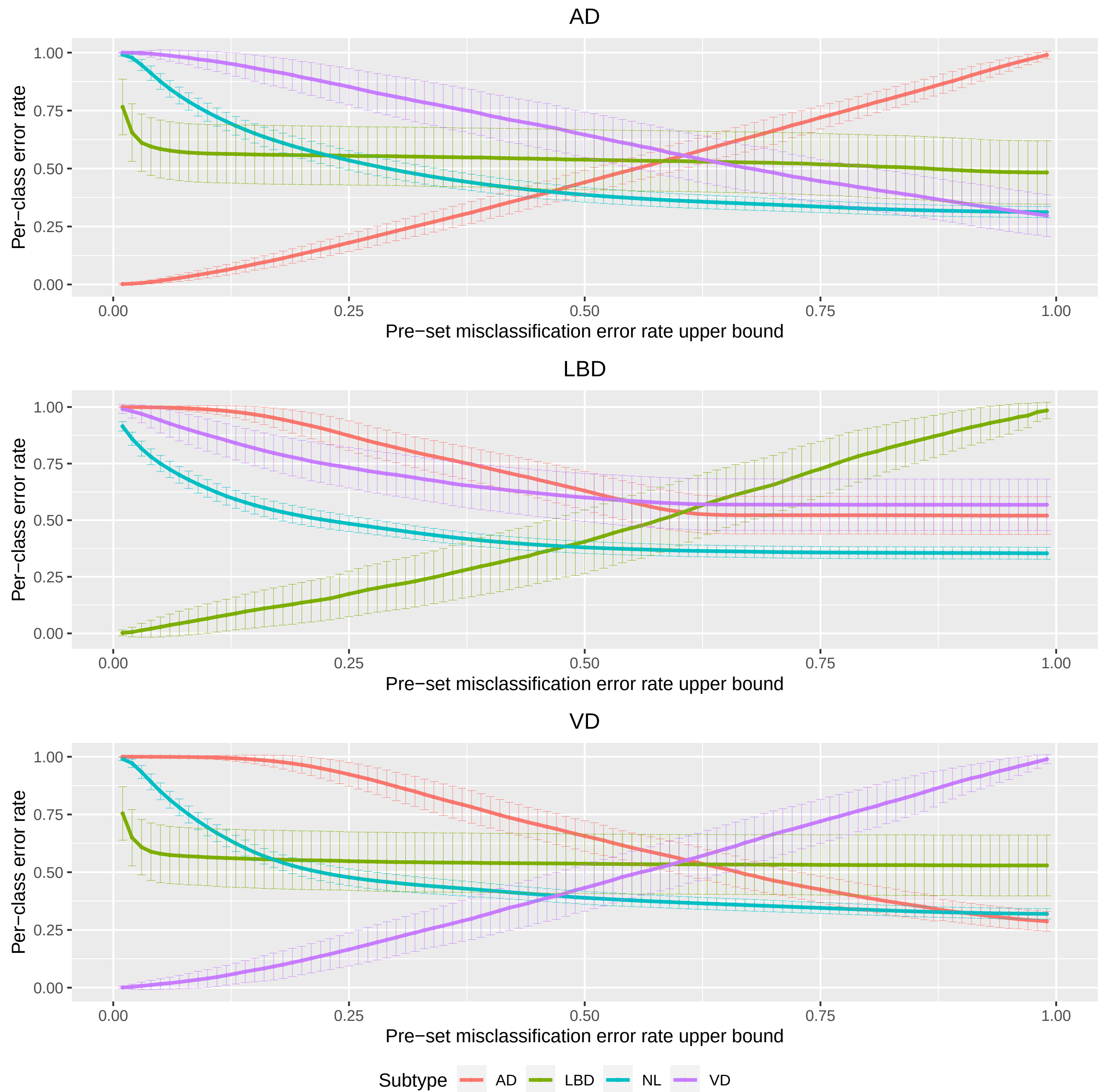
- (1) Year of birth is an important feature to distinguish AD or VD and NL. Controlling other factors, the later the patient was born, the less likely AD or VD is developed. Interestingly, birth year has more predictive power than age.
- (2) As expected, low baseline memory measures imply decline to AD.
- (3) Moderate alcohol consumption was found to protect against AD and VD. This might be caused by other unknown confounders.
- (4) The baseline diagnosis of Parkinson’s disease predisposes development of LBD.
- (5) Disorder REM sleep and hallucinations indicate decline to LBD.
- (6) Controlling other factors, decline to LBD is more prevalent in men than in women.
- (7) Stroke, transient ischemic attack and hypertension are import risk factors for future VD.

Prediction error control

We further manipulated the penalized multinomial regression model by controlling the prediction error rate on a given dementia subtype [5]. The method provides theoretical guarantee to solve constrained modeling. We tested above model by solving the following problems:

$$\begin{aligned} & \text{minimize } \Pr(\text{not predicted as NL}|\text{NL}) + \Pr(\text{not predicted as LBD}|\text{LBD}) + \Pr(\text{not predicted as VD}|\text{VD}) \\ & \text{subject to } \Pr(\text{not predicted as AD}|\text{AD}) \leq \alpha, \end{aligned}$$

where α is a pre-set error upper bound. We controlled error rates of three subtypes with different α ’s to examine how sensitive each subtype prediction error is to the control of the other subtype errors.



References

- [1] DE Barnes et al. “Predicting risk of dementia in older adults: The late-life dementia risk index”. In: *Neurology* 73.3 (2009), pp. 173–179.
- [2] Charlotte James et al. “Performance of machine learning algorithms for predicting progression to dementia in memory clinic patients”. In: *JAMA network open* 4.12 (2021), e2136553–e2136553.
- [3] Jinlei Li et al. “Practical risk score for 5-, 10-, and 20-year prediction of dementia in elderly persons: Framingham Heart Study”. In: *Alzheimer’s & Dementia* 14.1 (2018), pp. 35–42.
- [4] Dina Silva et al. “Prediction of long-term (5 years) conversion to dementia using neuropsychological tests in a memory clinic setting”. In: *Journal of Alzheimer’s Disease* 34.3 (2013), pp. 681–689.