Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment

Ming Lin[1], **Pinghua Gong**[1], **Tao Yang**[1], **Jieping Ye**[1], **Roger L. Albin**[2,3], and **Hiroko H. Dodge**[2,4]

[1]Department of Computational Medicine and Bioinformatics, University of Michigan, Ann Arbor, MI

[2]Department of Neurology and Michigan Alzheimer's Disease Center, University of Michigan, Ann Arbor, MI

[3]Neurology Service & GRECC, VAAAHS, Ann Arbor, MI

[4]Department of Neurology and Layton Aging and Alzheimer's Disease Center, Oregon Health & Science University, Portland, OR

Abstract

Background—Clinical trials increasingly aim to retard disease progression during pre-symptomatic phases of Mild Cognitive Impairment (MCI) and thus recruiting study participants at high risk for developing MCI is critical for cost-effective prevention trials. However, accurately identifying those who are destined to develop MCI is difficult. Collecting biomarkers is often expensive.

Methods—We used only non-invasive clinical variables collected in the National Alzheimer's Coordinating Center (NACC) Uniform Data Sets version 2.0 and applied machine learning techniques to build a low-cost and accurate Mild Cognitive Impairment (MCI) conversion prediction calculator. Cross-validation and bootstrap were used to select as few variables as possible accurately predicting MCI conversion within 4 years.

Results—31,872 unique subjects, 748 clinical variables and additional 128 derived variables in NACC data sets were used. 15 non-invasive clinical variables are identified for predicting MCI/aMCI/naMCI converters, respectively. Over 75% Receiver Operating Characteristic Area Under the Curves (ROC AUC) was achieved. By bootstrap we created a simple spreadsheet calculator which estimates the probability of developing MCI within 4 years with a 95% confidence interval.

Conclusions—We achieved reasonably high prediction accuracy using only clinical variables. The approach used here could be useful for study enrichment in pre-clinical trials where enrolling participants at risk of cognitive decline is critical for proving study efficacy, and also for developing a shorter assessment battery.

Corresponding Author: Hiroko H. Dodge. dodgeh@ohsu.edu or hdodge@umich.edu.

Keywords

Study enrichment; National Alzheimer's Coordinating Center Uniform Data Set (NACC UDS); mild cognitive impairment; incidence; prediction; dementia; bootstrap; machine learning; ROC AUC

Introduction

Clinical trials increasingly aim to retard disease progression during pre-symptomatic phases of Mild Cognitive Impairment (MCI) and Alzheimer's disease (AD), when it is more likely that pathologic changes can be arrested or reversed[1,2]. Recruiting study participants at high risk for developing MCI, i.e., study enrichment, is critical for cost-effective prevention trials[3,4]. Pre-symptomatic populations include both those who will convert to MCI (true at-risk subjects) and those who will retain normal cognitive status over time (false at-risk subjects). As little or no prevention effects can be detected among false at-risk subject in conventional trial durations, the higher the fraction of this group in trials, the more challenging to demonstrate intervention efficacy. Sample size power calculations are especially problematic as these estimates are often based on longitudinal trajectories of cognitive or functional outcomes among those who developed MCI vs. those who retained normal cognitive status during follow-up in prior studies. In reality, when recruiting at-risk subjects, some proportion of subjects will not experience cognitive decline as expected, reducing power to detect intervention efficacy.

Biofluid and imaging biomarkers are extensively evaluated as early indicators of pathological processes in clinical AD[5], but assessing these biomarkers is expensive and often challenging to apply widely among pre-symptomatic older adults. Recent findings suggest neuropsychological test results (i.e., involving less invasive methods and interviews only) might have as high discriminatory ability as biofluid or imaging biomarkers in stratifying at risk subjects[6]. It would be advantageous to use non-invasively collected clinical variables to identify accurately those at high risk for developing MCI within conventional trial durations. The National Alzheimer's Coordinating Center (NACC) was established by the National Institute on Aging (NIA, U01 AG016976) in 1999 to facilitate collaborative research among Alzheimer's Disease Centers (ADCs) across the United States. NACC developed and maintains a large relational database of standardized clinical and neuropathological research data called the Uniform Data Set (UDS) and standardized neuropsychological test results. Data are uploaded to the central repository from each ADC, with data cleaned and ready to be distributed to research communities (https://www.alz.washington.edu/). This dataset contains over 30,000 subjects (December, 2016). We applied big data analytics approaches to this rich dataset to derive the best model for distinguishing those developing MCI within a 4 year follow-up period from those retaining normal cognition. Previous study concluded that it is more cost effective to improve specificity than sensitivity.[3, 4] Therefore, we aimed to select the model with the highest specificity in the current study. By estimating weights of selected variables, we also developed a risk score calculator based subjects baseline characteristics to obtain the probability of conversion to MCI within 4 years.

Methods

Data

We used the clinical variables collected in the National Alzheimer's Coordinating Center (NACC) Uniform Data sets version 2.0 (UDS 2.0) downloaded April 2015. Thirty four National Institutes of Health (NIH) funded Alzheimer's Disease centers contributed the data. UDS visits conducted between September 2005 and March 2015 was included. The dataset contained 31,872 unique subjects. Out of 31,872 subjects, 15,516 subjects (48.7%) had three or more visits/assessments and 4 years of follow-up. Among them, 7026 subjects had normal cognition at baseline and were used in the current analyses. Out of 7026 subjects, 5883 subjects retained intact cognition (stable normal) during follow-up and 1143 subjects developed MCI or dementia within 4 years from their baseline evaluations (converters); 748 clinical variables were collected at initial visits and additional 128 derived variables were computed from UDS Version 2.0. After initial cleaning of variables (treating missing values correctly and re-categorizing responses if they are ordinal or categorical responses), 348 variables were included as candidate variables in analyses to select informative variables.

Diagnosis

MCI incidence was determined based on consensus meetings at each ADC. The amnestic MCI category includes single and multi-domain amnestic MCI. Non-amnestic MCI was MCI without memory impairment.

Statistical model

Our aim is to differentiate between those who converted to MCI and/or dementia within 4 years versus those retaining normal cognition. We compared discriminatory abilities indicated by Receiver Operating Characteristics Area Under Curve (ROC AUC) by using the following classifiers: Support Vector Machine (SVM), Logistic Regression (LR), and Random Forest (RF). For each classifier, we examined the following univariate feature selection methods: Information Gain (InfoGain)[7], χ^2-test (Chi2)[8], minimum Redundancy Maximum Relevance Feature Selection (mRMR)[9], Gini-index (Gini)[10], BLogReg[11], Fisher score[12], and Kruskal-Wallis test (KW)[13] . Briefly, InfoGain analysis first calculates the information gain for each clinical variable independently and then the features with specified numbers of highest information gains are selected as informative variables. Chi2 uses chi-square test to estimate the independence between clinical variables and diagnostic categories. A high value of χ^2 statistic indicates the failure of the hypothesis of independence of the two variables, indicating the high associativity of clinical variables and diagnostic categories. The mRMR selects the variables one-by-one where the selected variables are maximally relevant to the diagnostic categories and their dependence between each other is minimized. The Gini feature selection uses the Gini-index as the measurement of dependence between clinical variables and diagnostic categories. The BLogReg models the feature/label dependence based on an improved sparse logistic regression algorithm. The Fisher score feature selection is based on a so-called Fisher criterion to select variables. This criterion prefers the feature presentation where the distance of the same type of subjects is minimized while the distance of different type of subjects is maximized. The Kruskal-Wallis

test is a non-parametric statistical test method to detect the independence between clinical variables and diagnostic categories. It is a non-parametric method applicable to non-Gaussian distributions but may require more samples. The above methods are implemented by the software package FeatureMiner[14]. In addition to the above univariate feature selection approaches, jointly feature selection method based on L1-norm regularizer such as LASSO is another popular approach to simultaneously select features and learn feature weights. We use the liblinear[15] implementation of sparse Logistic Regression in our experiment (listed as LR-LASSO in Table 2). The parameters of LR-LASSO are tuned by five-fold cross-validation and subsampling in the same way as in the univariate feature selection approach.

Our first experiment is to find the best trade-off between the number of selected clinical variables and the discriminative power of the classifiers trained on these variables. We first examined the gain in ROC AUC by selecting top 10, 15, 20 and 25 variables. Preliminary study showed that after selecting 15 variables, there were little further gains in AUC by increasing the number of variables to predict MCI converters i.e., ROC AUC stabilizes after 15 variables (see Figure 1). Therefore, we used 15 variables as the number of candidate variables for ROC AUC for each model.

We randomly split 7026 subjects into 5 folds and used 4 folds for training and 1 fold for testing (i.e, 20% of the data is for testing). We repeated the five-fold cross-validation 100 times and then averaged over trials. We compared the average accuracy, sensitivity, specificity and AUC of each model for identifying converters. We first assessed incidence of overall MCI conversion, followed by limiting the outcomes to incidence of amnestic MCI or to incidence of non-amnestic MCI. Our focus was to find predictors of MCI incidence, but some subjects received diagnoses of dementia without prior incident MCI incidence (MCI skippers). These subjects were included as incidence of overall MCI if the transition to dementia occurred within 4 years from baseline assessment, presupposing that these subjects went through an undetected MCI stage prior to diagnoses of dementia. If responses to questionnaire items are yes/no, we created one dummy variable. If responses are categorical with multiple response categories, we created multiple dummy variables with the lowest category as a reference group. Other variables were treated as continuous variables in all models.

Since we used cross-validation and sub-sampling, the feature weights are random variables derived from multiple trials. We applied bootstrap method to estimate the mean value and the standard deviation of the feature weights. Once we estimated the mean value and the standard deviation, we generated the 95% confidential intervals using these values and a standard approach: Let the number of bootstrap loops be n, the empirical mean value be μ and the standard deviation be σ. The lower bound x_- and the upper bound x_+ of the 95% interval is given by

$$x_-=\mu - 1.96\frac{\sigma}{\sqrt{n}},\ x_+=\mu+1.96\frac{\sigma}{\sqrt{n}}\ .$$

Using the weights, we were able to estimate the probability of each subject converting to MCI by using his or her information for the selected variables. Given a new subject, we first

use the 100 SVM models to predict his/her affiliation of normal or MCI. Then we could estimate the probability of this patient converting to MCI by counting how many SVM models give positive predictions. Similarly, the 95% confidence interval was estimated as described above. Probability calculators in excel formats were generated.

Results

Table 1 shows baseline characteristics of the overall study cohort, incident MCI groups, and subjects retaining normal cognition. Mean age (std) of the total sample used here was 76.0 (10.0) with 66.1% female. Table 2 shows top-10 models with the highest AUC for predicting each outcome (MCI, aMCI and naMCI). AUC, accuracy, sensitivity and specificity are shown. There were not large differences in AUC across models within each outcome as shown in 2a, 2b, and 2c. For overall MCI, AUC was around 0.76 in all models and for aMCI around 0.77. For naMCI, AUC was slightly lower, ranging from 0.72 to 0.73. In addition to AUC, the specificity is of significant importance for cost-effective trial enrichment design[4]. Therefore, we chose the model with best specificity among the top-10 AUC models. The chosen models are exhibited in Table 2 with bold fonts.

Tables 3a – 3c shows the 15 variables selected as predictors of outcomes and their descriptions. Tables 4a- 4c summarize predictor variables specific to aMCI (4a), naMCI (4b), common to both subtypes (4c) and those selected only when overall MCI (including MCI skippers) was an outcome (4d). We also list weights for each variable in Supplemental table (Table S1:a-c) where weights are averaged weights over the five-fold cross-validation which were repeated 100 times. Predictor variables specific for aMCI incidence included older age, lower logical memory immediate and delayed recall scores (latter with a higher weight than the former), clinician's impression of subject's memory decline, difficulty in travelling independently, Unified Parkinson's Disease Rating Scale (UPDRS) non-dopaminergic deficiency indicators, and an indicator that subjects had motor neuron disease. However, the motor neuron disease variable has a very low weight (0.002) compared with other selected variables (see supplemental Table 1b). For example, one rank before this variable has weight of −0.015 (verbal fluency category animals) and therefore there is a large gap in importance in predicting an aMCI outcome between the 14th and 15th variables. For aMCI outcome, 14 variables are sufficient. naMCI specific variables included more frequent falls, disease status at enrollment (not being proband (those with 1st degree relatives being diagnosed as AD)), motor slowing, lower Boston naming test scores, lower length of digit span backwards, impairment in CDR's judgment/problem solving component, and having tremor. Variables selected across both MCI sub-types included (from high to low weights in order) UDS cognitive status based on UDS neuropsychological test results, informant's report of a decline in subject's memory, lower category fluency vegetables score, lower digit symbol scores, higher Trail B scores (i.e., taking longer time to complete), higher CDR sum of box, impairment in memory component of CDR and category fluency animals. Finally, using these selected variables and weights generated from models, we created Excel calculators which estimate probability of converting to MCI, aMCI and naMCI diagnoses within 4 years, given specific characteristics of subjects (Supplemental excel sheets 1 – 3), with 95% confidence interval of the probabilities.

Discussion

As clinical trials increasingly target pre-symptomatic subjects, enriching study populations with those likely to convert to MCI within conventional trial durations may reduce required sample sizes and costs associated with evaluating study participants[3,4]. NACC UDS is used in all NIH-funded Alzheimer's disease centers in the United States and the approach shown in this study may be useful for selecting candidate subjects for preclinical trials from the NACC sampling pool. In addition to enriching clinical trial samples, it proved possible to create a short battery of selected variables and associated weights for screening at-risk subjects when administering long batteries such as UDS is not feasible. Our results showed that we could achieve over 75% AUC and over 70% specificity in distinguishing incidence of MCI within 4 years from those who remained normal cognition without invasive and costly indicators such as biofluid or imaging biomarkers.

The selected variables predicting MCI incidence, which include age, CDR sum of boxes and CDR memory score, cognitive domains that tap memory (logical memory immediate and delayed test scores) and executive functions (Trail B, category fluency), attention (digit symbol), informant's report of subject's memory decline, are all well-established predictors of MCI. Two financial management abilities – managing taxes and paying bills – are also known FAQ items which decline early in the course of dementing disorders[16-19]. For predicting conversion to MCI, aMCI, and naMCI, useful predictor variables included items from the UPDRS, history of falls, presence of tremor (captured in UPDRS), bradykinesia, hypomimia, and speech changes. The selection of motor slowing is consistent with previous findings where decline in walking speed was an early indicator of cognitive impairment[20-22]. The inclusion of motor items among our useful variables emphasizes the utility of a careful, standardized motor evaluation, much of which is captured by the widely used and easily implemented UPDRS scale. This result may reflect the impact of forebrain amyloid burden on motor function[23].

One recent study applied a multimodal SVM method to identify those who converted from normal cognition to MCI or AD (vs. stable normal) within 24 months using Alzheimer's Disease Neuroimaging Initiative (ADNI) study participants[24]. Applying the modalities which include selected features of MRI, AV45-PET, and FDG-PET, they showed 70% accuracy, 75% sensitivity and 67% specificity in predicting MCI converters. Also using ADNI cohort, another study[25] examined the predictive ability of FDG-PET a priori specified ROI in identifying MCI converters. This study obtained 82% AUC ROC, identifying 11 subjects who received diagnosis of MCI or AD dementia within 4 years out of 54 initially healthy control subjects. After including results of Trail Making Test B, the AUC improved to 93.4%. In another study of ADNI participants[26], the authors used a model combining MRI & FDG-PET measures to achieve 81.2% accuracy (80.0% sensitivity, 82.4% specificity) in predicting MCI converters within 4 years. Unlike past studies, we used only clinical variables and obtained the possible combination of items with each variable weighted to maximize prediction of MCI converters. During ADC consensus conferences, clinicians and neuropsychologists gather all clinical information available, including those obtained in previous assessments and establish clinical diagnoses. Experienced clinicians weigh all the available information and provide the best possible judgment. Our approach

can be regarded as an algorithmically operationalized summary of clinical judgments predicting clinical MCI incidence within 4 years. Based on estimates obtained from these models, we generated user friendly Excel spreadsheets which return the probability of developing MCI, aMCI, and naMCI with 95% confidence intervals.

Clearly selecting a population to maximize treatment benefit (predictive enrichment, e.g., selecting those with high amyloid burden for anti-amyloid trials) is critical. If not, the trial is likely to fail due to lack of benefits among the experimental group. It is ideal if our proposed risk calculation approach based solely on clinical variables could be combined with a biomarker-based enrichment strategy. This will likely ensure the efficacy shown in clinical outcomes (not just biomarker modifications) as well as reduction in cost of following false positive subjects longitudinally.

Limitations of our analyses include potential sampling bias. Those enrolling in ADC cohorts are not representative community samples, even when subjects exhibit normal cognition. Therefore, application of probability calculators to non-NACC data might not be valid. The education level of this group, for example, is likely above the population average. This may limit generalizability of our study results. Even though standard criteria and procedures are applied across all ADCs, there may be some variability in selection and diagnoses factors among centers[27,28].

The neuropsychological (NP) test battery in UDS used in the current study (Version 2.0) was replaced by new tests in March of 2015 (Version 3)[29] as part of an effort for NACC to use non-proprietary cognitive tests (see detail: https://www.alz.washington.edu/WEB/researcher_home.html). Once data with Version 3 NP test battery are accumulated, we plan to construct an equivalent calculator based on the new battery. We will post our current conversion probability calculator on the NACC researcher website, after permission from NACC, for wider use among AD center researchers. Considering uncertainty in psychological distress posed by disclosing estimated probability of getting MCI diagnosis,[30] our plan is to release the calculator only for research purposes.

Conclusions

Using only non-invasively collected clinical variables, we achieved over 75% ROC AUC for identifying subjects converting to MCI within 4 years of initial evaluations. The proposed variable selection and MCI converter identification approaches may be useful in clinical trial enrichment and also assist in creating a shorter battery for screening at-risk subjects.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

References

1. Aisen PS, Andrieu S, Sampaio C, et al. Report of the task force on designing clinical trials in early (predementia) AD. Neurology. 2011; 76(3):280–286. [PubMed: 21178097]

2. Vellas B, Aisen PS, Sampaio C, et al. Prevention trials in Alzheimer's disease: an EU-US task force report. Prog Neurobiol. 2011; 95(4):594–600. [PubMed: 21925234]

3. Breitner JCS. How can we really improve screening methods for AD prevention trials? Alzheimer Disease and Associated Disorders. 2016; 2(1):45–47.

4. Leoutsakos JM, Bartlett AL, Forrester SN, Lyketsos CG. Simulating effects of biomarker enrichment on Alzheimer's disease prevention trials: conceptual framework and example. Alzheimers Dement. 2014; 10(2):152–161. [PubMed: 23954029]

5. Dodge HH, Zhu J, Harvey D, et al. Biomarker progressions explain higher variability in stage-specific cognitive decline than baseline values in Alzheimer disease. Alzheimers Dement. 2014; 10(6):690–703. [PubMed: 25022534]

6. Ewers M, Walsh C, Trojanowski JQ, et al. Prediction of conversion from mild cognitive impairment to Alzheimer's disease dementia based upon biomarkers and neuropsychological test performance. Neurobiol Aging. 33(7):1203–1214 e1202.

7. Mitchell, TM. Machine Learning. McGraw-Hill, Inc; New York: 1997.

8. Liu H, Setiono R. Chi2: Feature selection and discretization of numeric attributes. Paper presented at: ICTAI. 1995

9. Peng H, Long F, Ding C. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy. IEEE Transactions on pattern analysis and machine intelligence. 2005; 27(8):1226–1238. [PubMed: 16119262]

10. Shang W, Huang H, Zhu H, Lin Y, Qu Y, Wang Z. A novel feature selection algorithm for text categorization. Expert Systems with Applications. 2007; 33(1):1–5.

11. Cawley GC, Talbot NL. Gene selection in cancer classification using sparse logistic regression with Bayesian regularization. Bioinformatics. 2006; 22(19):2348–2355. [PubMed: 16844704]

12. Duda, RO., Hart, PE., Stork, P. DG: Pattern Classification. Wiley & Sons; New York: 2001.

13. Kruskal WH, Wallis WA. Use of ranks in one-criterion variance analysis. Journal of the American statistical Association. 1952; 47(260):583–621.

14. Li J, Cheng K, Wang S, et al. Feature selection: A data perspective. arXiv preprint arXiv: 160107996. 2016

15. Fan R-E, Chang K-W, Hsieh C-J, Wang X-R, Lin C-J. LIBLINEAR: A Library for Large Linear Classification. J Mach Learn Res. 2008; 9:1871–1874.

16. Marson DC, Sawrie SM, Snyder S, et al. Assessing financial capacity in patients with Alzheimer disease: A conceptual model and prototype instrument. Archives of neurology. 2000; 57(6):877–884. [PubMed: 10867786]

17. Marson D. Conceptual Models and Guidelines for Clinical Assessment of Financial Capacity. Archives of clinical neuropsychology: the official journal of the National Academy of Neuropsychologists. 2016; 31(6):541–553. [PubMed: 27506235]

18. Lichtenberg PA, Stoltman J, Ficker LJ, Iris M, Mast B. A Person-Centered Approach to Financial Capacity Assessment: Preliminary Development of a New Rating Scale. Clinical gerontologist. 2015; 38(1):49–67. [PubMed: 25866438]

19. Lichtenberg PA. New Approaches to Preventing Financial Exploitation: A Focus on the Banks. The Public policy and aging report. 2016; 26(1):15–17. [PubMed: 27076776]

20. Verghese J, Robbins M, Holtzer R, et al. Gait dysfunction in mild cognitive impairment syndromes. J Am Geriatr Soc. 2008; 56(7):1244–1251. [PubMed: 18482293]

21. Dodge HH, Mattek NC, Austin D, Hayes TL, Kaye JA. In-home walking speeds and variability trajectories associated with mild cognitive impairment. Neurology. 2012; 78(24):1946–1952. [PubMed: 22689734]

22. Buracchio T, Dodge HH, Howieson D, Wasserman D, Kaye J. The trajectory of gait speed preceding mild cognitive impairment. Arch Neurol. 2010; 67(8):980–986. [PubMed: 20697049]

23. Muller ML, Frey KA, Petrou M, et al. beta-Amyloid and postural instability and gait difficulty in Parkinson's disease at risk for dementia. Movement disorders: official journal of the Movement Disorder Society. 2013; 28(3):296–301. [PubMed: 23239424]

24. Zhan Y, Chen K, Wu X, et al. Identification of Conversion from Normal Elderly Cognition to Alzheimer's Disease using Multimodal Support Vector Machine. Journal of Alzheimer's disease: JAD. 2015; 47(4):1057–1067. [PubMed: 26401783]

25. Ewers M, Brendel M, Rizk-Jackson A, et al. Reduced FDG-PET brain metabolism and executive function predict clinical progression in elderly healthy subjects. NeuroImage Clinical. 2014; 4:45–52. [PubMed: 24286024]

26. Rizk-Jackson A, Insel P, Petersen R, Aisen P, Jack C, Weiner M. Early indications of future cognitive decline: stable versus declining controls. PloS one. 2013; 8(9):e74062. [PubMed: 24040166]

27. Steenland K, Macneil J, Bartell S, Lah J. Analyses of diagnostic patterns at 30 Alzheimer's disease centers in the US. Neuroepidemiology. 2010; 35(1):19–27. [PubMed: 20357515]

28. Dodge HH, Zhu J, Woltjer R, et al. Risk of incident clinical diagnosis of Alzheimer's disease–type dementia attributable to pathology-confirmed vascular disease. Alzheimers Dement. 2017; 13:613–623. [PubMed: 28017827]

29. Monsell SE, Dodge HH, Zhou XH, Bu Y, Besser LM, Mock C, et al. Results From the NACC Uniform Data Set Neuropsychological Battery Crosswalk Study Alzheimer Dis Assoc Disord. 2016; 2:134–9.

30. Roberts JS, Cupples LA, Relkin NR, Whitehouse PJ, Green RC, Group RS. Genetic risk assessment for adult children of people with Alzheimer's disease: the Risk Evaluation and Education for Alzheimer's Disease (REVEAL) study. J Geriatr Psychiatry Neurol. 2005; 18(4):250–5. [PubMed: 16306249]
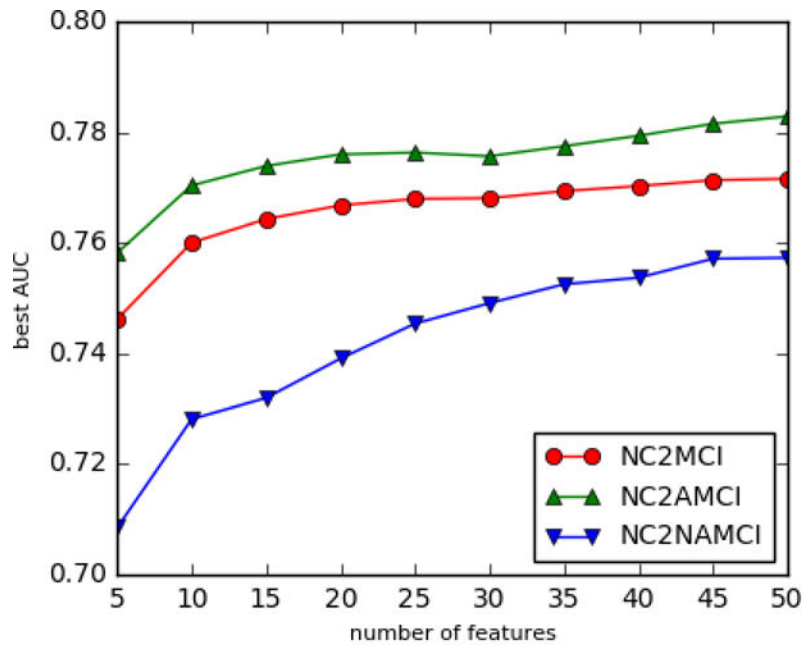
NC2MCI: AUC to discriminate incidence MCI from stable normal
NC2AMCI: AUC to discriminate incidence aMCI from stable normal
NC2NAMCI: AUC to discriminate non-amnestic MCI from stable normal

Figure 1. AUC of the best model varying with number of predictor variables (features)
[The x-axis is the number of variables]

Table 1

Demographic Characteristics at Baseline

	Normal Stayers	MCI Overall	MCI INCIDENCE				Overall Sample
					Subgroups		
			aMCI	naMCI	Dementia (MCI Skippers)*		
Frequency	**5883**	**1143**	763	253	127		**7026**
Age (mean+std)	75.2(9.9)	80.5(9.3)	81.2(9.3)	77.5(9.4)	82.4(7.5)		76.0(10.0)
Age Categories							
• **Age 90**	4.90%	13.60%	15.80%	6.80%	13.30%		6.30%
• **Age between 80 and 89**	25.60%	38.50%	38.50%	33.60%	49.10%		27.60%
• **Age between 70 and 79**	39.30%	34.30%	33.90%	36.50%	32.40%		38.50%
• **Age between 60 and 69**	23.60%	11.50%	10.00%	19.40%	4.20%		21.60%
• **Age between 50 and 59**	5.40%	1.50%	1.00%	3.40%	1.00%		4.80%
• **Age<50**	1.30%	0.60%	0.70%	0.40%	0.00%		1.20%
Women	67.20%	60.50%	61.10%	56.50%	65.40%		66.10%
Years of Education (mean+std)	15.7 (3.0)	15.3 (3.2)	15.4 (3.0)	15.2 (3.6)	14.8 (3.5)		15.7 (3.0)
Non-white	17.60%	15.40%	15.20%	17.80%	11.80%		17.20%
Missing Race	0.20%	0.30%	0.50%	0.00%	0.00%		0.20%

*: Those who received diagnosis of dementia without observing MCI incidence within 4 years from baseline. These subjects were included when outcome is overall MCI, assuming that they experienced MCI stage within 4 years, before having the diagnosis of dementia.

Table 2

Top 10 Models Selected for Each Outcome.

Classifier	Feature Selection	Accuracy	Sensitivity	AUC±std	Specificity±std
Table 2a. Overall MCI					
LR	Fisher	0.627	0.793	0.764±0.01	0.599±0.01
SVM	Fisher	0.710	0.676	0.764±0.01	0.716±0.02
LR	Gini	0.628	0.789	0.764±0.01	0.601±0.01
SVM	Gini	0.710	0.676	0.763±0.01	0.716±0.02
LR	mRMR	0.627	0.789	0.763±0.01	0.600±0.01
LR	Chi2	0.625	0.790	0.763±0.01	0.598±0.01
SVM	*mRMR*	*0.710*	*0.677*	*0.762±0.01*	*0.717±0.02*
SVM	Chi2	0.709	0.679	0.762±0.01	0.715±0.01
LR	InfoGain	0.625	0.789	0.762±0.01	0.598±0.01
SVM	InfoGain	0.709	0.678	0.761±0.01	0.715±0.01
LR	LASSO	0.689	0.700	0.765±0.01	0.687±0.02
Table 2b. aMCI					
SVM	*Fisher*	*0.710*	*0.711*	*0.774±0.02*	*0.710±0.01*
SVM	Chi2	0.709	0.715	0.773±0.02	0.709±0.01
LR	Fisher	0.595	0.832	0.773±0.02	0.574±0.02
SVM	Gini	0.709	0.714	0.773±0.02	0.708±0.01
SVM	InfoGain	0.710	0.714	0.773±0.02	0.709±0.01
SVM	mRMR	0.709	0.715	0.773±0.02	0.708±0.01
LR	Gini	0.594	0.832	0.772±0.01	0.575±0.02
LR	Chi2	0.594	0.829	0.772±0.01	0.574±0.02
LR	mRMR	0.595	0.832	0.772±0.01	0.574±0.02
LR	InfoGain	0.594	0.829	0.772±0.01	0.574±0.02
LR	LASSO	0.666	0.752	0.773±0.02	0.654±0.01
Table 2c. naMCI					
RF	Chi2	0.692	0.658	0.732±0.03	0.710±0.03
RF	Gini	0.691	0.656	0.730±0.03	0.706±0.03
RF	InfoGain	0.692	0.655	0.730±0.03	0.709±0.03

Classifier	Feature Selection	Accuracy	Sensitivity	AUC±std	Specificity±std
RF	mRMR	0.693	0.652	0.728±0.03	0.712±0.03
LR	Chi2	0.590	0.734	0.728±0.03	0.605±0.03
SVM	Gini	0.745	0.579	0.727±0.03	0.752±0.03
SVM	Chi2	0.744	0.574	0.727±0.03	0.752±0.03
LR	Gini	0.592	0.734	0.727±0.03	0.607±0.03
SVM	*InfoGain*	*0.745*	*0.573*	*0.726±0.03*	*0.753±0.03*
LR	InfoGain	0.593	0.727	0.725±0.03	0.608±0.03
LR	LASSO	0.622	0.704	0.722±0.02	0.619±0.02

The best model (high AUC with the highest specificity) is marked bold fonts. For AUC and specificity we report the standard deviation over the 100 trials. The 95% confidence interval can be obtained by the confidence interval formula on page 7.

Table 3

Variables Selected as Predictors of MCI Incidence and Characteristics at Baseline

Variable Name	Minimum Value	Maximum value	Mean (Total)	Mean (Converters)	Mean (Stable Normal)	Variable Descriptions
Table 3a. MCI overall						
NACCAGEB	21.00 *	100.00	72.43	77.32	71.48	Subject age at the initial visit (years)
CDRSUM	0.00	6.00	0.10	0.25	0.07	Standard CDR sum of boxes, 0, 0.5, 1, 1.5, ~18.0 (except scores of 16.5 and 17.5 which are not possible).
DECIN	0.00	1.00	0.12	0.25	0.09	Does the informant report a decline in subject's memory relative to previously attained abilities? 0 = No, 1 = Yes
MOSLOW	0.00	1.00	0.17	0.20	0.15	Slowness (Has the subject noticeably slowed down in walking or moving or handwriting, other than due to an injury or illness? Has his/her facial expression changed, or become more "wooden" or masked and unexpressive?) 0 = No 1 = Yes
MEMORY	0.00	1.00	0.05	0.11	0.03	Memory in Clinical Dementia Rating (CDR) 0.0 = No impairment 0.5 = Questionable impairment 1.0 = Mild impairment 2.0 = Moderate impairment 3.0 = Severe impairment
NACCZLMD	−3.02	3.84	0.37	−0.06	0.45	Age-, sex-, and education-adjusted z-score for Logical Memory 1A-Delayed total
NACCZTRB	−5.40	2.43	−0.17	−0.55	−0.10	Age-, sex-, and education-adjusted z-score for the Trail B score
NACCZVEG	−2.96	6.11	0.95	0.54	1.02	Age-, sex-, and education-adjusted z-score for Category Fluency: vegetables
SPEECH	0.00	8.00	0.09	0.16	0.06	0 = Normal. 1 = Slight loss of expression, diction and/or volume. 2 = Monotone, slurred but understandable; moderately impaired. 3 = Marked impairment,
TAXES	0,1,2,3,8		0 : 90.07%, 1 : 2.18%, 2 : 1.27%, 3 : 0.67%, 8 : 5.80%	0 : 80.84%, 1 : 5.30%, 2 : 3.91%, 3 : 2.14%, 8 : 7.81%	0 : 91.84%, 1 : 1.59%, 2 : 0.77%, 3 : 0.39%, 8 : 5.41%	Taxes in Functional Activities Questionnaire (FAQ) In the past four weeks, did the subject have any difficulty or need help with assembling tax records, business affairs, or other papers. 0 = Normal 1 = Has difficulty, but does by self 2 = Requires assistance 3 = Dependent 8 = Not applicable (e.g., never did)

Variable Name	Minimum Value	Maximum value	Mean (Total)	Mean (Converters)	Mean (Stable Normal)	Variable Descriptions
NACCZLMI	−3.37	3.44	0.33	−0.04	0.40	Age-, sex-, and education-adjusted z-score for Logical Memory 1A-Immediate total number of items recalled
NACCZWAI	−4.67	5.50	0.37	0.01	0.43	Age-, sex-, and education-adjusted z-score for the WAIS-R Digit Symbol score
NACCLEVB	0.00	18.00	0.47	0.91	0.38	Levy B score for levodopa-nonresponsive symptoms
BILLS	0,1,2,3,8		0 : 93.14%, 1 : 1.84%, 2 : 0.67%, 3 : 0.51%, 8 : 3.84%	0 : 86.14%, 1 : 5.02%, 2 : 2.14%, 3 : 1.95%, 8 : 4.74%	0 : 94.48%, 1 : 1.23%, 2 : 0.39%, 3 : 0.23%, 8 : 3.67%	Bills in Functional Activities Questionnaire (FAQ) In the past four weeks, did the subject have any difficulty or need help with writing checks, paying bills, or balancing a checkbook. 0 = Normal 1 = Has difficulty, but does by self 2 = Requires assistance 3 = Dependent 8 = Not applicable (e.g., never did)
NACCZANI	−3.85	5.76	−0.01	−0.31	0.05	Age-, sex-, and education-adjusted z-score for Category Fluency: animals
Table 3b. aMCI						
NACCAGEB	21.00	100.00	72.23	77.96	71.48	Subject age at the initial visit (years)
NACCZLMD	−3.02	3.84	0.38	−0.12	0.45	Age-, sex-, and education-adjusted z-score for Logical Memory 1A-Delayed total
NACCZLMI	−3.37	3.44	0.34	−0.08	0.40	Age-, sex-, and education-adjusted z-score for Logical Memory 1A-Immediate total number of items recalled
NACCZVEG	−2.96	6.11	0.97	0.56	1.02	Age-, sex-, and education-adjusted z-score for Category Fluency: vegetables
MEMORY	0.00	1.00	0.04	0.10	0.03	Memory in Clinical Dementia Rating (CDR) 0.0 = No impairment 0.5 = Questionable impairment 1.0 = Mild impairment 2.0 = Moderate impairment 3.0 = Severe impairment
NACCZWAI	−4.67	5.50	0.39	0.06	0.43	Age-, sex-, and education-adjusted z-score for the WAIS-R Digit Symbol score
DECIN	0.00	1.00	0.11	0.23	0.09	Does the informant report a decline in subject's memory relative to previously attained abilities? 0 = No, 1 = Yes
NACCLEVB	0.00	18.00	0.44	0.88	0.38	Levy B score for levodopa-nonresponsive symptoms UDS subjects: The Unified Parkinson's Disease Rating Scale (UPDRS) items can be categorized into two groups: symptoms associated with dopaminergic deficiency and symptoms not associated with dopaminergic deficiency. The Levy B score is a summary score for the severity of UPDRS items not associated with dopaminergic deficiency: speech and axial impairment.

Variable Name	Minimum Value	Maximum value	Mean (Total)	Mean (Converters)	Mean (Stable Normal)	Variable Descriptions
						Speech (SPEECH)/Arising from a chair (ARISING)/Posture (POSTURE)/gait (GAIT)/Posture stability (POSSTAB) Possible score range: 0 –20
CDRSUM	0.00	6.00	0.08	0.22	0.07	Standard CDR sum of boxes, 0, 0.5, 1, 1.5, ~18.0 (except scores of 16.5 and 17.5 which are not possible).
DECCLIN	0.00	1.00	0.04	0.10	0.03	Based on the clinician's judgment, is the subject currently experiencing meaningful impairment in cognition? 0 = No, 1 = Yes
COGSTAT	1.00	4.00	1.68	1.95	1.64	Based on the UDS neuropsychological examination, the subject's cognitive status is deemed 1 = Better than normal for age 2 = Normal for age 3 = One or two test scores abnormal 4 = Three or more scores are abnormal or lower than expected
NACCZTRB	−5.40	2.43	−0.14	−0.48	−0.10	Age-, sex-, and education-adjusted z-score for the Trail B score
NACCZANI	−3.85	5.76	0.01	−0.28	0.05	Age-, sex-, and education-adjusted z-score for Category Fluency: animals
TRAVEL	0,1,2,3,8		0 : 96.07%, 1 : 2.16%, 2 : 0.68%, 3 : 0.35%, 8 : 0.74%	0 : 90.10%, 1 : 5.58%, 2 : 2.09%, 3 : 0.98%, 8 : 1.26%	0 : 96.83%, 1 : 1.73%, 2 : 0.50%, 3 : 0.27%, 8 : 0.68%	In the past four weeks, did the subject have any difficulty or need help with traveling out of the neighborhood, driving, or arranging to take public transportation. 0 = Normal 1 = Has difficulty, but does by self 2 = Requires assistance 3 = Dependent 8 = Not applicable (e.g., never did)
NACCMND	0.00	1.00	0.00	0.00	0.00	Motor neuron disease write-in on Form D1 0 = No write-in of motor neuron disease or ALS 1 = Write-in indicating presence of motor neuron disease or ALS This variable is designed to flag subjects for whom a clinical diagnosis of "motor neuron disease" (including "ALS" or similar indicative text), was written-in on Form D1. This variable flags the presence of a write-in only, and not whether the condition was considered to contribute to cognitive impairment.
Table 3c. naMCI						
COGSTAT	1.00	4.00	1.66	2.07	1.64	Based on the UDS neuropsychological examination, the subject's cognitive status is deemed: 1 = Better than normal for age 2 = Normal for age 3 = One or two test scores abnormal 4 = Three or more scores are abnormal or lower than expected
NACCZWAI	−4.67	5.22	0.41	−0.14	0.43	Age-, sex-, and education-adjusted z-score for the WAIS-R Digit Symbol score

Variable Name	Minimum Value	Maximum value	Mean (Total)	Mean (Converters)	Mean (Stable Normal)	Variable Descriptions
CDRSUM	0.00	5.00	0.08	0.30	0.07	Standard CDR sum of boxes, 0, 0.5, 1, 1.5, ~18.0 (except scores of 16.5 and 17.5 which are not possible).
MOSLOW	0.00	1.00	0.16	0.18	0.15	Slowness (Has the subject noticeably slowed down in walking or moving or handwriting, other than due to an injury or illness? Has his/her facial expression changed, or become more "wooden" or masked and unexpressive?) 0 = No 1 = Yes
NACCZDBL	−4.24	2.51	0.07	−0.26	0.09	Age-, sex-, and education-adjusted z-score for Digit Span Backward length
MEMORY	0.00	1.00	0.04	0.15	0.03	Memory in Clinical Dementia Rating (CDR) 0.0 = No impairment 0.5 = Questionable impairment 1.0 = Mild impairment 2.0 = Moderate impairment 3.0 = Severe impairment
NACCZBOS	−9.76	1.88	−0.31	−0.69	−0.30	Age-, sex-, and education-adjusted z-score for the Boston Naming Test score
NACCZVEG	−2.96	6.11	1.00	0.53	1.02	Age-, sex-, and education-adjusted z-score for Category Fluency: vegetables
DECIN	0.00	1.00	0.10	0.28	0.09	Does the informant report a decline in subject's memory relative to previously attained abilities? 0 = No 1 = Yes
MOFALLS	0.00	1.00	0.06	0.05	0.06	Does the subject fall more than usual? 0 = No 1 = Yes
NACCZANI	−3.85	5.76	0.03	−0.42	0.05	Age-, sex-, and education-adjusted z-score for Category Fluency: animals
NACCZTRB	−5.40	2.43	−0.12	−0.69	−0.10	Age-, sex-, and education-adjusted z-score for the Trail B score
PRESTAT	1.00	3.00	2.08	2.03	2.08	Presumed disease status at enrollment. 1 = Case/patient/proband 2 = Control/normal 3 = No presumed disease status
MOTREM	0.00	1.00	0.13	0.13	0.13	Has the subject had rhythmic shaking, especially in the hands, arms, legs, head, mouth or tongue?) 0 = No 1 = Yes
JUDGMENT	0.00	1.00	0.02	0.06	0.01	Judgment & problem solving in CDR 0.0 = No impairment 0.5 = Questionable impairment 1.0 = Mild impairment 2.0 = Moderate impairment

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Variable Name	Minimum Value	Maximum value	Mean (Total)	Mean (Converters)	Mean (Stable Normal)	Variable Descriptions
						3.0 = Severe impairment

If response is categorical with multiple responses, we created multiple dummy variables with the lowest category as a reference group. Other variables were treated as continuous variables in all models.

*

: Seven subjects of MCI were under age 50.

Table 4

Predictor Variables Specific to aMCI or naMCI (Listed in order of from high to low weights)

Variable	Description	Direction
Table 4a. Predictor Variables Specific to aMCI		
NACCAGEB	Age at assessment	+
NACCZLMD	Age-, sex-, and education-adjusted z-score for Logical Memory 1A-Delayed total	−
DECCLIN	Does the clinician believe there has been a current meaningful decline in the subject's memory, non-memory cognitive abilities, behavior, or ability to manage his/her affairs, or have there been motor/movement changes relative to previously attained abilities? (Yes)	+
TRAVEL	In the past four weeks, did the subject have any difficulty or need help with traveling out of the neighborhood, driving, or arranging to take public transportation.	+
NACCZLMI	Age-, sex-, and education-adjusted z-score for Logical Memory 1A-Immediate total number of items recalled	−
NACCLEVB	Levy B score for levodopa-nonresponsive symptoms: UPDRS non- dopaminergic deficiency.	+
NACCMND	Subjects for whom a clinical diagnosis of "motor neuron disease" is indicated in the form (Yes)	+
Table 4b. Predictor Variables Specific to naMCI		
MOFALLS	Does the subject fall more than usual? (Yes)	+
PRESTAT	Presumed disease status at enrollment. (case or proband)	−
MOSLOW	Slowness (Has the subject noticeably slowed down in walking or moving or handwriting, other than due to an injury or illness? Has his/her facial expression changed, or become more "wooden" or masked and unexpressive?) (Yes)	+
NACCZBOS	The Boston Naming Test Age-, sex-, and education-adjusted z-score for the Boston Naming Test score	−
NACCZDBL	The length on the Digit Span Backward test Age-, sex-, and education-adjusted z-score for Digit Span Backward length	−
JUDGMENT	Judgment & problem solving in CDR	+
MOTREM	Tremor (Has the subject had rhythmic shaking, especially in the hands, arms, legs, head, mouth or tongue?) (Yes)	+
Table 4c. Predictor variables common to both aMCI and naMCI incidence (In the order of from high to low weights appeared for aMCI incidence)		
COGSTAT	Based on the UDS neuropsychological examination, the subject's cognitive status is deemed: 1 = Better than normal for age 2 = Normal for age 3 = One or two test scores abnormal 4 = Three or more scores are abnormal or lower than expected	+
DECIN	Does the informant report a decline in subject's memory relative to previously attained abilities? (Yes)	+
NACCZVEG	Total number of vegetables named in 60 seconds Age-, sex-, and education-adjusted z-score for Category 'vegetables'	−
NACCZWAI	Age-, sex-, and education-adjusted z-score for the WAIS-R Digit Symbol score	−
NACCZTRB	Age-, sex-, and education-adjusted z-score for the Trail B score	−
CDRSUM	Clinical Dementia Rating Sum of Box	+
MEMORY	Memory in CDR	+
NACCZANI	Total number of animals named in 60 seconds Age-, sex-, and education-adjusted z-score for Category 'animals'	−
Table 4d. Variables selected only when outcome is overall MCI		
SPEECH	0 = Normal. 1 = Slight loss of expression, diction and/or volume. 2 = Monotone, slurred but understandable; moderately impaired. 3 = Marked impairment, difficult to understand.	+

Variable	Description	Direction
	4 = Unintelligible.	
TAXES	In the past four weeks, did the subject have any difficulty or need help with assembling tax records, business affairs, or other papers	+
BILLS	In the past four weeks, did the subject have any difficulty or need help with writing checks, paying bills, or balancing a checkbook.	+

*: The variables in Table 4c. are in the order of from high to low weights appeared for aMCI incidence. Please check supplemental material for the full table with variable description. The symbol +/− indicates the positive/negative effects of the variables.