Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

HHS Public Access

Author manuscript

J Alzheimers Dis. Author manuscript; available in PMC 2023 September 27.

Published in final edited form as:

J Alzheimers Dis. 2023 ; 95(2): 535–548. doi:10.3233/JAD-230208.

Predicting Progression to Clinical Alzheimer's Disease Dementia Using the Random Survival Forest

Shangchen Song, MS[a], **Breton Asken, PhD, ATC**[b,e,f], **Melissa J. Armstrong, MD, MSc**[c,e], **Yang Yang, PhD**[d,*], **Zhigang Li, PhD**[a,*],

Alzheimer's Disease Neuroimaging Initiative[1]

[a]Department of Biostatistics, University of Florida College of Public Health & Health Professions and College of Medicine, Gainesville, Florida, 32611, USA

[b]Department of Clinical and Health Psychology, University of Florida College of Public Health & Health Professions, Gainesville, FL, 32611, USA

[c]Departments of Neurology and Health Outcomes & Biomedical Informatics, University of Florida College of Medicine, Gainesville, FL, 32611, USA

[d]Department of Statistics, University of Georgia Franklin College of Arts and Sciences, Athens, GA, 30602, USA

[e]Norman Fixel Institute for Neurological Diseases, University of Florida, Gainesville, FL, 32608, USA

[f]University of Florida Center for Cognitive Aging and Memory, McKnight Brain Institute, Gainesville, FL, 32610, USA

Abstract

BACKGROUND: Assessing the risk of developing clinical Alzheimer's disease (AD) dementia, by machine learning survival analysis approaches, among participants registered in Alzheimer's Disease Centers is important for AD dementia management.

OBJECTIVE: To construct a prediction model for the onset time of clinical AD dementia using the National Alzheimer Coordinating Center (NACC) and the Alzheimer's Disease Neuroimaging Initiative (ADNI) registered cohorts.

METHODS: A model was constructed using the Random Survival Forest (RSF) approach and internally and externally validated on the NACC cohort and the ADNI cohort. An R package and a Shiny app were provided for accessing the model.

---

[1]Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: http://adni.loni.usc.edu/wpcontent/uploads/how_to_apply/ADNI_Acknowledgement_List.pdf

JAD website: https://www.j-alz.com/; This version complies with a CC-BY copyright statement.

Corresponding author: Zhigang Li, PhD, Department of Biostatistics, 2004 MOWRY RD FL 5, Gainesville, Florida, 32611, USA, Tel 352-294-5915, Fax 352-294-5915, zhigang.li@ufl.edu.

*Senior authors

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RESULTS: We built a predictive model having the six predictors: delayed logical memory score (story recall), CDR® Dementia Staging Instrument - Sum of Boxes, general orientation in CDR®, ability to remember dates and ability to pay bills in the Functional Activities Questionnaire, and patient age. The C indices of the model were 90.82% (SE = 0.71%) and 86.51% (SE = 0.75%) in NACC and ADNI respectively. The time-dependent AUC and accuracy at 48 months were 92.48% (SE = 1.12%) and 88.66% (SE = 1.00%) respectively in NACC, and 90.16% (SE = 1.12%) and 85.00% (SE = 1.14%) respectively in ADNI.

CONCLUSION: The model showed good prediction performance and the six predictors were easy to obtain, cost-effective and non-invasive. The model could be used to inform clinicians and patients on the probability of developing clinical AD dementia in 4 years with high accuracy.

Keywords

Dementia; machine learning; Alzheimer's disease; survival analysis

INTRODUCTION:

Individuals presenting for clinical evaluation with memory concerns can show no clear decrements on cognitive testing (i.e., "subjective cognitive decline") or perform objectively low on testing but maintain their independence with completing daily activities (i.e., mild cognitive impairment or MCI). These patients do not inevitably progress to dementia, defined as objective evidence of cognitive impairment sufficient to impact the independent execution of daily activities. Identifying patients who are likely to progress to dementia and predicting the onset time of dementia could inform clinical management, such as prioritizing high-risk patients to receive certain interventions. It would also assist patients and their families with long-term care planning.

Recently, many prediction models were proposed for large-scale automatic screening [1], for example, Support Vector Machine, Logistic Regression, Cox Regression, Random Forest, and Neural Networks. However, most of them used classification prediction methods [2–4] that do not intrinsically model the time and censored cases (dementia onset was not observed on those subjects by the end of the study). Because cognitive changes can progress gradually, it is not uncommon that individuals in longitudinal aging studies never reach the outcome of dementia. Excluding these cases biases results. Other approaches have been limited to classical survival statistical methods (e.g. Cox proportional hazards model) [5] which are not appropriate for large complex clinical datasets with potential multicollinearity and high-order interactions.

Machine learning methods (ML) are considered state-of-the-art tools to model complex clinical datasets, especially those in heterogeneous, high-dimensional, and non-linear relationship structures [6]. The Random Survival Forest (RSF) [7] model is an ML-based survival analysis method that can model the censored data and output the time-to-event probabilities. It benefits from directly tackling the time and censoring as its main outcomes, which utilizes more information compared to classification methods and avoids biased results due to censoring. It is also able to output the probability of the event occurrence at

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

every event time observed in samples, whereas a classification model can only output the probability of event occurrence at a pre-specified time.

Predicting progression to clinically-defined dementia diagnosis is critical. With rapid advances in technology, biomarkers like structural or positron emission tomography (PET) neuroimaging, cerebrospinal fluid (CSF) assays have been increasingly available, and the accuracy of laboratory tests has significantly improved to identify the etiology of symptoms and likelihood of symptom progression in patients with neurodegenerative conditions like Alzheimer disease (AD). However, such tests are not ubiquitous, and several remain cost-prohibitive in many clinical settings. For example, AD-conversion survival analyses based on neuroimaging data are often limited in sample size which inevitably compromises their generalizability [8–13]. Prediction models that accurately classify patients at greatest risk of progressing to dementia by leveraging readily available clinical data remain essential and would have widespread utility, especially in settings without access to advanced biomarker collection tools.

In this study, we utilized two large datasets that included patients with and without cognitive impairment who were followed longitudinally to track symptom progression – the National Alzheimer's Coordinating Center Uniform Dataset (NACC-UDS) and the Alzheimer's Disease Neuroimaging Initiative (ADNI). Our overarching aim was to identify readily accessible variables that are highly predictive of patients who were free of dementia at baseline but progressed to clinical AD dementia during the study. We built and validated an RSF model that uses only a small set of accessible variables. A secondary aim was to provide the model via an R package and a Shiny app to facilitate its wide use by the clinical research community.

METHODS:

DATA SOURCE

The data were from two sources: NACC and ADNI. NACC was launched by the National Institute on Aging in 1999 and collected data from the NIA-funded Alzheimer's Disease Centers (ADCs) across the United States (https://naccdata.org). By February 2022, the NACC-UDS had three versions and contained records for over 45,000 patients ranging from cognitively normal to dementia in 45 ADCs. The other dataset used for external validation is the ADNI-MERGE dataset from ADNI database (http://adni.loni.usc.edu). ADNI, established in 2003, is another longitudinal multicenter study. Its main goal is to develop clinical, imaging, genetic, and biochemical biomarkers for the early detection and tracking of MCI and AD dementia.

ETHICAL CONSIDERATIONS

Both NACC and ADNI obtained approvals from the Institutional Review Boards of all participating institutions. In NACC, ethical approval was obtained from each site's institutional review board, and all participants provided informed written consent. Informed written consent was also acquired from participants or their proxies in ADNI. The University of Florida IRB deemed the current analysis to be exempt (IRB202202769).

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

DIAGNOSIS OF CLINICAL AD DEMENTIA

The outcome of interest for the current analysis was time to a clinician diagnosis of AD dementia. In both NACC and ADNI, cognitive status and suspected etiology were determined by clinicians at each contributing site [5,14]. For the current analysis, the variable NACCUDSD was used to determine the cognitive status for each participant, and the variable NACCETPR was used for the primary etiology for cases derived from NACC [5]. In ADNI dataset, AD is the exclusive etiology, and variables DX.bl and DX at each visit were used to determine the cognitive status [14].

PARTICIPANT INCLUSION CRITERIA

In the datasets, we included those participants who (1) had at least one follow-up visit, (2) were not diagnosed with any dementia at the baseline visit, and (3) were age 55 years or older at baseline [5,15]. In NACC, we included individuals who were labeled as cognitively normal (CN), MCI, or impaired-but-not-MCI. In ADNI, we included individuals labeled as CN or MCI.

NACC VERSION CHANGES

In 2015, the NACC UDS underwent a substantial revision into version 3 [16]. Several variables in the Neuropsychological Battery in versions 1 and 2 were no longer collected in version 3, whereas alternatives of some of these variables and other new variables were added in version 3. To overcome this data version change, the NACC data were split into two parts in our analysis – the first part consisted of the participants whose first visit was measured by the UDS forms version 1 or version 2, denoted as NACC-v1v2; the second part was the cohort of participants who were evaluated by the UDS form version 3 on their first visit, denoted as NACC-v3.

MODEL CONSTRUCTION DATA

We built our initial model using NACC-v1v2 because of the large sample size and NACC's widespread use in similar dementia research [2,4,5,17]. The baseline collection date for the participants in NACC-v1v2 ranged from June 09, 2005, to June 04, 2015. From a total of 32,787 subjects and 137,544 visit records, we sequentially detected and excluded 8,834 people who did not have any follow-up visit, 7,651 people who had at least one follow-up visit but were diagnosed with dementia at the baseline visit, and 714 participants younger than 55 years old at baseline. After that, 15,588 participants were included in the NACC-v1v2 cohort.

MODEL VALIDATION DATA

NACC-v3 was used as the first external validation dataset to assess the generalizability of our derived model. NACC-v3 contained participants who had baseline measurements from March 16, 2015, to February 09, 2022. From a total of 12,313 subjects and 28,538 visit records, we sequentially excluded 4,778 people who did not have any follow-up visit, 1,719 people who had at least one follow-up visit but were diagnosed with dementia at the baseline visit, and 330 participants younger than 55 years old. After that, 5,486 participants were included in the NACC-v3 cohort.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

The second external validation dataset was ADNI, which contains individuals who were recruited in ADNI-1, ADNI-GO, ADNI-2, and ADNI-3 and had baseline visits from September 07, 2005, to May 17, 2022. The total number of participants was 2,404 with 15,945 visit records. After successively removing 265 participants who had no follow-up, 367 participants who had follow-up but were diagnosed with dementia at the baseline visit, and 3 participants who had no age information (the rest were all older than 55), the ADNI cohort consisted of 1,769 participants.

SURVIVAL ANALYSIS (TIME-TO-EVENT ANALYSIS)

Progression from CN or MCI to dementia can take several years or even decades [18]. Therefore, many participants were not diagnosed with dementia prior to the end of the study. This cannot be handled by common regression or classification methods. Our study utilized survival analysis, also known as time-to-event analysis, which is a collection of statistical methods that can appropriately handle censoring of event times.

OUTCOME VARIABLES

The primary outcome variable was the time to the clinical AD dementia. This was defined by the total months from the baseline visit to a dementia onset with a clinical primary etiology of AD or a censoring event. Censoring events were events that would essentially prevent observation of a diagnosis of dementia in the study and included dropout, death, or diagnosis of dementia with a non-AD clinical primary etiology (e.g., Lewy body disease, frontotemporal lobar degeneration).

CANDIDATE PREDICTORS

The candidate predicting variables were selected by the following variable exclusion criterion from all available 1014 variables measured at the baseline visit in NACC-v1v2. Our variable exclusion criterion consisted of (1) free text variables, such as symptom specification (2) administration variables, such as patient ID and form version; (3) cross-sectional variables that record the longitudinal results, such as baseline variables denoting if the patient finally progresses to dementia during the study; (4) variables with over 50% missing entries. After excluding these variables, 317 candidate variables remained for the subsequent variable selection procedure. A detailed data pre-processing procedure and names of candidate variables are available in the Supplementary Materials S1 and Supplementary Table 3.

PREDICTION MODEL

The Random Survival Forest was used to build the prediction model. RSF is an ensemble learning method containing a great number of deep-grown survival decision trees that can achieve low bias. The output of RSF is synthesized from the results of all the built trees. The *random* comes from two sources: 1) each tree was grown on a random bootstrap samples and 2) the splitting variables were randomly selected at each node during the tree construction. These two randomization steps decorrelate the trees and can decrease the variance introduced by bagging. With the combination of low bias and variance reduction techniques, the RSF can approximate various forms of functions while maintaining low

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

generalization errors. The RSF has advantages in handling the following complications: 1) high-dimensional data 2) a large number of categorical variables 3) existence of outliers, 4) non-linear associations and 5) other violations of assumptions for traditional regression models such as multicollinearity [19]. RSF is regarded as one of the state-of-art machine learning algorithms for heterogeneous data, such as the NACC and ADNI datasets, which consist of many categorical and continuous clinical outcome variables on different scales. The RSF was implemented by the R package, *randomForestSRC* version 3.1.1 [20]. Figure 1 displays the model development and evaluation diagram.

VARIABLE SELECTION

We used the permutation variable importance (VIMP) generated by RSF to select the best variable set [21] among the candidate variables. The VIMP was calculated by randomly permuting the values of a variable in out-of-bag samples so that the contribution of the variable to predicting the outcome is like random noise. Then the difference of the prediction errors before and after the permutation was recorded as the VIMP. If the decrease in the prediction error was high, it meant that the permutated variable was predictive of the response. We first obtained the VIMP of all the candidate 317 variables on NACC-v1v2, then these variables were ranked by the descending VIMP order. A subset of variables was accordingly selected to build the final mode considering 1) the VIMP ranking 2) availability and cost-effective consideration 3) final model complexity.

MODEL VALIDATION

Our model evaluation employed three commonly used measurements for time-to-event models. Harrell's C-index [22] assesses an overall model prediction performance regardless of a specific time point. It ranges from 0 to 1 and represents the proportion of the concordant pairs where the subjects with higher (lower) risk matched shorter (longer) survival time. Models having Harrell's C-index over 70% are generally regarded as good models [23]. The time-dependent Area under Curve (AUC) and time-dependent accuracy at 48 months were also reported for a time-specific model performance [24,25]. The standard errors (SE) of these measures were obtained by ten-fold cross-validation in the internal validation (approximately) or bootstrap sampling in the external validation. Harrell's C-index was calculated by R package *Hmisc* version 4.7–0 [26]. The time-dependent AUC and time-dependent accuracy were calculated by R package *timeROC* version 0.4 [27].

Because the prediction model was developed based on NACC-v1v2, data from NACC-v1v2 could not be used to directly test the model. Instead, a ten-fold cross-validation was employed for internal validation to prevent the over-optimistic model performance due to data leakage [19]. Specifically, the data were randomly and evenly split into ten folds; each time, one of the ten folds was treated as a testing set and the rest nine folds were used as a training set to build the inner model. There were ten different testing sets on each of which the corresponding inner model was evaluated. The average performance and standard errors (SE) of the inner models were recorded.

For the external validation using NACC-v3 and ADNI, when model variables were all available in the validation datasets, our model was directly applied to output the model

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

measurements. When model variables were not available in the validation datasets, we rescaled and replaced the NACC-v1v2 variable with the most closely related variable from the validation dataset.

VARIABLE PARTIAL EFFECT

We further investigated the role each variable played in the model. The partial dependence plot (PDP) [28] can show the average model prediction given one variable adjusting for all other variables in a dataset, empirically. We in effect fix the selected variable at one of its possible values each time while letting the rest of model variables take values across all data points in the dataset. The mathematical formula is,

$$\widetilde{f}_t(x) = \frac{1}{n}\sum_{i=1}^{n} \widehat{f}_t\left(x, x_{i,o}\right)$$

where $\widetilde{f}_t$ is the partial dependence function at time t,n is the sample size, $\widehat{f}_t$ is the model prediction function at time t,x is the target variable value, $x_{i,o}$ is the rest of model variables of the i th subject. Since the model prediction also involves time point for time-to-event data as denoted, we focused on the PDPs at 48 months.

SUBGROUP ANALYSIS OF BIOMARKER-VERIFIED COHORT

We conducted a subgroup analysis on subjects in the NACC-v1v2 and NACC-v3 cohorts who had an etiological diagnosis of AD supported by a positive AD biomarker. The criterion to identify AD biomarker evidence was based on abnormally elevated amyloid on PET (variable AMYLPET in NACC) or abnormally low amyloid in CSF (variable AMYLCSF in NACC) suggested by White et al. (2022) [29].

SENSITIVITY ANALYSIS FOR MISSING DATA DUE TO IMPAIRMENT

In NACC data, missing data are coded based on reasons for missingness. One of the missing data codes (“96”) reflects data missing because of cognitive/behavior problems, likely reflecting severe clinical impairment. To use as much information as possible, in the main analysis, 96 was recoded as the lowest score of the test. For example, for a test scoring from 0 to 10, if a person had missing code 96, then his/her score was marked to 0. A sensitivity analysis was performed in which the missing code 96 was treated as non-informative missing values.

MODEL COMPARISON TO XGBOOST

Further, in place of RSF, we implemented and evaluated the XGBoost models with the same six variables selected by RSF and on the same datasets. Supplementary Materials S6 showed a brief description of XGBoost and the settings of the tuning parameters. The R package used for XGBoost is *xgboost v 1.5.2.1* [30] downloaded from CRAN.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

RESULTS:

NACC-V1V2 COHORT BASELINE CHARACTERISTICS

The NACC-v1v2 cohort included 15,588 subjects (mean age at baseline = 73.2±8.6 years, 59.5% female, mean education = 15.5±3.2 years; see Table 1). The participants were on average followed for 65.6 (SD = 46.9) months prior to event or censoring. When comparing participants that progressed to clinical AD dementia (N = 2,866, 18.4%) with the non-progressed participants (N = 12,722. 81.6%), the “progressed” group was significantly older at baseline, more likely to be male, and had fewer years of education. The “progressed” group also had shorter follow-up time, higher baseline Clinical Dementia Rating® scale Sum of Boxes scores (CDR-SB), and lower baseline Mini Mental State Examination (MMSE) scores, and was more likely to have baseline MCI and more likely to report subjective cognitive decline among its CN group at baseline.

SELECTION OF THE MODEL VARIABLES

We collected the average VIMPs of all 317 candidate variables in the NACC-v1v2 cohort by independently running the VIMP algorithm with ten different random seeds. The top 6 variables were selected (Figure 2) consisting of delayed logical memory score (MEMUNITS), subject age (NACCAGE), Functional Activities Questionnaire (REMDATES, BILLS), Clinical Dementia Rating (CDR-SB, ORIENT). All are available in both NACC-v1v2 and ADNI, whereas in NACC-v3, MEMUNITS was not collected. NACC-v3 implemented the Craft Story in lieu of Logical Memory and similarly includes an immediate and delayed recall or story components. Therefore, we rescaled and used delayed verbatim recall score from the Craft Story (CRAFTDVR) in place of MEMUNITS for the NACC-v3 external validation. Table 2 presented the names, the description, and the value ranges of these 6 variables in NACC-v1v2, NACC-v3, and ADNI.

NACC-V3 COHORT BASELINE CHARACTERISTICS

There were 5,486 subjects in the NACC-v3 cohort (mean age at baseline = 71.0±7.5 years, 60.2% female, mean education = 16.2±2.9 years; see Table 3). The participants were on average followed for 30.7±16.5 months prior to event or censoring. When comparing the progressed participants (N = 448, 8.2%) with the non-progressed participants (N = 5,038, 91.8%), the “progressed” group was older at baseline and more likely to be male. The “progressed” group again had significantly lower follow-up time and had higher baseline CDR-SB and lower MMSE scores. This group was more likely to have MCI at baseline but did not significantly report more subjective cognitive decline among its CN group at baseline, possibly relating to sample size.

ADNI COHORT BASELINE CHARACTERISTICS

The ADNI cohort included 1,769 participants (Table 4). They were on average followed for 52.1±40.9 months prior to event or censoring. Compared to the non-progressed participants (N = 1,364. 77.1%), the “progressed” group (N = 405. 22.9%) was older at baseline and more likely to be male. The “progressed” group also had significantly shorter follow-up

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

time, fewer years of education, higher baseline CDR-SB, and lower baseline MMSE. Individuals who progressed were significantly more likely to have MCI at baseline.

MODEL PERFORMANCE

The internal model evaluation on NACC-v1v2 resulted in a C index 85.21% (SE = 0.80%). The time-dependent AUC and the time-dependent accuracy at 48 months were respectively 90.26% (SE = 0.94%) and 90.06% (SE = 0.42%). The predictive performance in the testing sets across the ten-fold cross-validation table was shown in Table 5. The external validation on NACC-v3 showed a better performance in that the C index was 90.82% (SE = 0.71%). Additionally, the time-dependent AUC and accuracy at 48 months were 92.48% (SE = 1.12%) and 88.66% (SE = 1.00%), respectively. For the ADNI external validation, the C index was 86.51% (SE = 0.75%) and the time-dependent AUC and accuracy at 48 months were 90.16% (SE = 1.12%). and 85.00% (SE = 1.14%), respectively.

VARIABLE PARTIAL EFFECT

We displayed the PDPs at 48 months of the six variables based our model and NACC-v1v2 cohort (Figure 3). For convenience, we dubbed the y-axis values in PDPs as the "average predicted probability" which is interpreted as the average predicted probability of progressing to clinical AD dementia at 48 months in the NACC-v1v2 cohort at the fixed variable values, i.e., the x-axis values in the PDPs. For BILLS, the average predicted probability went up from 15% to 30% as the BILLS took value from 0 to 3. REMDATES had a similar PDP pattern as BILLS. The average predicted probability rose from 14% to 17% as ORIENT increased from 0 to 1 and leveled off for larger values of ORIENT. For CDR-SB, the average predicted probability increased sharply from 5% to 55% as the CDR-SB increased from 0 to 7.5. The flat trend at the end of the curve was due to few cases with the very high CDR-SB. For MEMUNITS, when it varied from 0 to 10, the average predicted probability dropped from 35% to 10%, which means that the more units the subject recalled, the lower average predicted probability was. After MEMUNITS scores over 10, the average predicted probability concentrated at 5% to 10%. For NACCAGE, 55–65 age group had a horizontal average predicted probability at 10%; 65–90 age group showed a steady increase from 10% to 30%; 90–110 leveled off at 30% which is possibly due to limited cases in the oldest age group. In these PDPs, all variables except for ORIENT showed prominent effects on changing the risk of progression to AD dementia at 48 months, especially for CDRSUM.

MODEL IMPLEMENTATION AND VISUALIZATION

We provided a direct access to download and implement the model via an R package (https://github.com/songuf/NACCADNI) and a web browser visualization tool via an R Shiny App (https://shangchensong.shinyapps.io/NACC-ADNI-shiny) that can generate the predicted probabilities and the time-to-event curve for each input subject. A screenshot of the Shiny was included as the Supplementary Figure 1.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

SUBGROUP AND SENSITIVITY ANALYSES

The subgroup analysis was conducted for the subjects in the NACC cohort who had an etiology confirmation by biomarkers (n = 309 for NACC-v1v2 and n = 806 for NACC-v3). The model measurements in this subgroup were still considered good (Supplementary Table 1). The SEs were larger than the main analyses since the sample size was significantly reduced. The sensitivity analysis where the code 96 in cognitive test variables was treated as missing values shows good performance for the model as well (Supplementary Table 2).

MODEL COMPARISON TO XGBOOST

The performance of XGBoost was evaluated in the 48-month time-dependent AUCs, which were 90.60% (SE = 0.62%) in the internal validation, 89.50% (SE = 0.93%) in NACC-v3 external validation, and 84.76% (SE = 1.13%) in the ADNI external validation. All these metrics were no better or comparable than those of RSF.

DISCUSSION:

This study employed a novel RSF approach with its build-in VIMP algorithm to construct a parsimonious prediction model using the NACC-v1v2 dataset. Only six readily available and easily collected variables were required for the model prediction. The model accurately predicted clinical AD dementia onset time of participants who did not have dementia at baseline (CN, impaired-but-not-MCI or MCI) across data sets. These findings can inform targeted evaluations of older adults at risk for dementia. Earlier identification of older adults at elevated risk of progressing to dementia is important for intervention and care planning [31–33].

Previous research rarely considered intrinsic censorship in dementia data [2–4,34] but instead handled it as a classification problem. For a comprehensive review of studies focusing on classification methods, we refer readers to Kumar et al. (2021) [35], Battineni et al. (2022) [36], and Javeed et al. (2023) [37]. The classification methods typically dichotomize the response by whether or not a subject experiences the event by a specified time point. This practice may raise two issues. First, given that event onset information at only one time point is modelled, the prediction output merely reflects that one time point. Second, if a participant was lost to follow-up before the cut time point, he/she is discarded due to the incomplete observation. Our model resolves these two issues by directly considering dementia onset as a time-to-event outcome.

Although large longitudinal datasets provide abundant sources of potential risk factors for dementia progression, in the traditional model construction, the variables are often pre-specified by experts or/and evidence from the existing literature. This does not commonly cover all available variables and may miss relevant factors. The VIMP algorithm used in this study is able to screen a large number of variables while taking into account all possible interactions among the variables, providing an efficient tool to sieve out the most influential features [38].

Our model is based on RSF, which is considered the state-of-art machine learning survival model for complex clinical data. Compared to classical time-to-event methods, the RSF

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

can appropriately model high-dimensional and non-linear complex relationships with tree structures. To the best of our knowledge, this is the first survival analysis of non-dementia cohorts based on the ADNI or NACC data that utilized solely clinical variables for prediction. As a parsimonious model, our method outperformed the three most relevant historical studies that also utilized NACC or ADNI populations. Wang et al. (2022) [5] proposed a Cox proportional hazard model with 7 clinical variables to predict the time-to-any-cause dementia onset in the MCI cohort of NACC and ADNI, reporting C indices of 69% and 72%, respectively [5]. Khajehpiri et al. (2021) [14] developed an XGBoost model using 14 MRI-based and clinical variables in ADNI to predict the onset time of clinical AD dementia among MCI participants, achieving a C index of 84.5% [8]. Using a deep survival analysis with 63 features from MRI-based, genetic, and clinical variables, Mirabnahrazama et al. (2022) [13] predicted the time-to-AD-dementia in 401 subjects from either a CN or MCI state within the ADNI dataset with a C index of 83.1%. Mirabnahrazama et al. (2022) acknowledged the small sample size of their study and recommended external validation on NACC data, which is exactly what we did in this study.

Prior studies using NACC data usually did not consider the changes from UDS version 1 and version 2 to the version 3 updated in 2015 [2,4] and external validation using multiple datasets is uncommon. The three validations in our study strengthen the findings. Additionally, all variables in our model (CDR-SB, delayed story recall, item-level responses from the FAQ [ability to remember dates or pay bills], general orientation, and patient age) are easy to obtain, cost-effective, non-invasive and showed considerable predictive utility in other research. An earlier study indicated that delayed story recall was among the strongest predictors of cognitive decline in their models [3]. The rest of variables selected in our models overlap with those in the Wang and Khajehpiri models using NACC and ADNI previously, including CDR-SB, Functional Activities Questionnaire, and subject baseline age.

Dementia refers to the presence of objective cognitive and/or behavioral decline of sufficient severity to impact independent execution of activities of daily living, irrespective of etiology. Multiple neurodegenerative diseases can cause dementia and symptoms alone may have limited specificity to pinpoint the underlying disease. In recent years, academic research centers and some clinical settings have begun more routinely performing CSF analyses and/or PET imaging that are well-validated for in vivo detection of AD pathology. Incorporating AD-specific biomarkers improves accuracy in identifying the etiology of a patient’s cognitive decline. We therefore evaluated a subgroup of participants with biomarker-confirmed AD to see whether our model performed differently in patients with biomarker-supported AD compared to a clinical AD dementia cohort. The subgroup analysis among biomarker-confirmed AD participants showed that the six easily collected clinical variables are highly predictive of progress to AD-dementia in both types of cohorts, suggesting our model has a wide applicability.

Our method can be applied to develop predictive models for other types of dementia as well. For instance, we may treat Frontotemporal Lobar Degeneration (FTLD) as the event of interest and consider other dementia types as censoring events. To illustrate, we conducted the internal model training and testing for FTLD on the NACC-v1v2 data and selected

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

six most important variables by the VImp algorithm: (1) primary progressive aphasia (PPHPHIF), (2) motor disturbance severity (MOTSEV), (3) primary progressive aphasia subtype (NACCPPME), (4) the subject completed FTLD module visits or not (NACCFTD), (5) presumptive etiologic diagnosis of primary progressive aphasia (PPA) and (6) Total MMSE score (NACCMMSE). With this model, the ten-fold cross validation results showed both great C-index (93.21%, SD: 5.51%) and 48-month time-dependent AUC (93.69%, SD: 7.63%).

Our study has several limitations. First, the NACC and ADNI databases are referral-, volunteer-based databases that might not translate to population-based samples. NACC has known biases relating to recruitment of largely white non-Hispanic participants, and the generalizability of our method to other ethnic groups is to be investigated in the future when data become available. Additionally, we assumed that data were missing completely at random and the incomplete cases for the given six variables are removed during the model construction. Third, like many other survival analysis methods, the RSF relies on the assumption of non-informative censoring, which may not be necessarily true in the investigated datasets. Fourth, a carefully designed method may be needed to represent MEMUNITS with CRAFTDVR when the former is unavailable, although the current rescaling approach works reasonably well. Moreover, only publicly available clinical variables were utilized in our model; other modalities of data such as neuroimages, if available with a reasonable sample size, could be incorporated in our model to further improve its predictive performance. Finally, when high multicollinearity exists among variables, the Accumulated Local Effects Plot (ALE) was suggested in literature as an alternative to PDP [39]. However, the ALE has not been implemented in the *randomforestSRC* or any other R packages.

In conclusion, this study developed and validated an RSF model with good prediction accuracy progression to clinical AD dementia. The RSF model built in this study is easily implementable and can provide useful prognostic information for clinical practice. Future studies may directly conduct variable selection among cohorts with biomarker-confirmed AD dementia if sample size allows and investigate variables discarded in this study due to missingness. The RSF model will be extended to handle informative censoring to further improve its predictive performance.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

ACKNOWLEDGEMENTS:

The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADRCs: P30 AG062429 (PI James Brewer, MD, PhD), P30 AG066468 (PI Oscar Lopez, MD), P30 AG062421 (PI Bradley Hyman, MD, PhD), P30 AG066509 (PI Thomas Grabowski, MD), P30 AG066514 (PI Mary Sano, PhD), P30 AG066530 (PI Helena Chui, MD), P30 AG066507 (PI Marilyn Albert, PhD), P30 AG066444 (PI John Morris, MD), P30 AG066518 (PI Jeffrey Kaye, MD), P30 AG066512 (PI Thomas Wisniewski, MD), P30 AG066462 (PI Scott Small, MD), P30 AG072979 (PI David Wolk, MD), P30 AG072972 (PI Charles DeCarli, MD), P30 AG072976 (PI Andrew Saykin, PsyD), P30 AG072975 (PI David Bennett, MD), P30 AG072978 (PI Neil Kowall, MD), P30 AG072977 (PI Robert Vassar, PhD), P30 AG066519 (PI Frank LaFerla, PhD), P30 AG062677 (PI Ronald Petersen, MD, PhD), P30 AG079280 (PI Eric Reiman, MD), P30 AG062422 (PI Gil Rabinovici,

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

MD), P30 AG066511 (PI Allan Levey, MD, PhD), P30 AG072946 (PI Linda Van Eldik, PhD), P30 AG062715 (PI Sanjay Asthana, MD, FRCP), P30 AG072973 (PI Russell Swerdlow, MD), P30 AG066506 (PI Todd Golde, MD, PhD), P30 AG066508 (PI Stephen Strittmatter, MD, PhD), P30 AG066515 (PI Victor Henderson, MD, MS), P30 AG072947 (PI Suzanne Craft, PhD), P30 AG072931 (PI Henry Paulson, MD, PhD), P30 AG066546 (PI Sudha Seshadri, MD), P20 AG068024 (PI Erik Roberson, MD, PhD), P20 AG068053 (PI Justin Miller, PhD), P20 AG068077 (PI Gary Rosenberg, MD), P20 AG068082 (PI Angela Jefferson, PhD), P30 AG072958 (PI Heather Whitson, MD), P30 AG072959 (PI James Leverenz, MD).

ADNI Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health (www.fnih.org). The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuro Imaging at the University of Southern California.

CONFLICT OF INTEREST:

M. Armstrong receives support from the NIH (P30AG066506, R01AG068128, R01NS121099, R44AG062072), the Florida Department of Health (grant 20A08), and as the local PI of a Lewy Body Dementia Association Research Center of Excellence. She serves on the DSMBs for the Alzheimer's Therapeutic Research Institute/Alzheimer's Clinical Trial Consortium and the Alzheimer's Disease Cooperative Study. She has provided educational content for Medscape. All other authors declare that they have no conflicts of interest.

FUNDING:

M. Armstrong receives support from the NIH (P30AG066506, R01AG068128, R01NS121099, R44AG062072), the Florida Department of Health (grant 20A08). S. Song and Z. Li are supported by R01GM123014. All other authors have no funding to report.

DATA AVAILABILITY:

The NACC and ADNI data supporting the findings of this study are available upon request at and https://adni.loni.usc.edu/data-samples/access-data/, respectively. Programming codes will be made available upon request by email to the corresponding author.

REFERENCE:

[1]. Grueso S, Viejo-Sobera R (2021) Machine learning methods for predicting progression from mild cognitive impairment to Alzheimer's disease dementia: a systematic review. Alzheimers Res Ther 13, 162. [PubMed: 34583745]

[2]. James C, Ranson JM, Everson R, Llewellyn DJ (2021) Performance of Machine Learning Algorithms for Predicting Progression to Dementia in Memory Clinic Patients. JAMA Netw Open 4, e2136553. [PubMed: 34913981]

[3]. Gupta A, Kahali B (2020) Machine learning-based cognitive impairment classification with optimal combination of neuropsychological tests. Alzheimers Dement Transl Res Clin Interv 6, e12049.

[4]. Lin M, Gong P, Yang T, Ye J, Albin RL, Dodge HH (2018) Big Data Analytical Approaches to the NACC Dataset: Aiding Preclinical Trial Enrichment. Alzheimer Dis Assoc Disord 32, 18–27. [PubMed: 29227306]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[5]. Wang M, Sajobi TT, Ismail Z, Seitz D, Chekouo T, Forkert ND, Fischer K, Mackie A, Pearson D, Patry D, Cieslak A, Menon B, Barber P, McLane B, Granger R, Hogan DB, Smith EE, Initiative for the ADN (2022) A pragmatic dementia risk score for patients with mild cognitive impairment in a memory clinic population: Development and validation of a dementia risk score using routinely collected data. Alzheimers Dement Transl Res Clin Interv 8, e12301.

[6]. Spooner A, Chen E, Sowmya A, Sachdev P, Kochan NA, Trollor J, Brodaty H (2020) A comparison of machine learning methods for survival analysis of high-dimensional clinical data for dementia prediction. Sci Rep 10, 20410. [PubMed: 33230128]

[7]. Ishwaran H, Kogalur UB, Blackstone EH, Lauer MS (2008) Random survival forests. The Annals of Applied Statistics 2, 841–860.

[8]. Liu K, Chen K, Yao L, Guo X (2017) Prediction of Mild Cognitive Impairment Conversion Using a Combination of Independent Component Analysis and the Cox Model. Front Hum Neurosci 11, 33. [PubMed: 28220065]

[9]. Li Y, Wang L, Zhou J, Ye J (2019) Multi-task learning based survival analysis for multi-source block-wise missing data. Neurocomputing 364, 95–107.

[10]. Pölsterl S, Sarasua I, Gutiérrez-Becker B, Wachinger C (2020) A Wide and Deep Neural Network for Survival Analysis from Anatomical Shape and Tabular Clinical Data. In Machine Learning and Knowledge Discovery in Databases, Cellier P, Driessens K, eds. Springer International Publishing, Cham, pp. 453–464.

[11]. Nakagawa T, Ishida M, Naito J, Nagai A, Yamaguchi S, Onoda K, on behalf of the Alzheimer's Disease Neuroimaging Initiative (2020) Prediction of conversion to Alzheimer's disease using deep survival analysis of MRI images. Brain Commun 2, fcaa057. [PubMed: 32954307]

[12]. Lu P, Colliot O (2022) Multilevel Survival Modeling with Structured Penalties for Disease Prediction from Imaging Genetics data. IEEE J Biomed Health Inform 26, 798. [PubMed: 34329174]

[13]. Mirabnahrazam G, Ma D, Beaulac C, Lee S, Popuri K, Lee H, Cao J, Galvin JE, Wang L, Beg MF (2023) Predicting time-to-conversion for dementia of Alzheimer's type using multi-modal deep survival analysis. Neurobiol Aging 121, 139–156. [PubMed: 36442416]

[14]. Khajehpiri B, Moghaddam H, Forouzanfar M, Lashgari R, Ramos-Cejudo J, Osorio R, Ardekani B (2021) Survival Analysis in Cognitively Normal Subjects and in Patients with Mild Cognitive Impairment Using a Proportional Hazards Model with Extreme Gradient Boosting Regression. J Alzheimers Dis 85, 1–14.

[15]. Mukadam N, Sampson EL (2011) A systematic review of the prevalence, associations and outcomes of dementia in older general hospital inpatients. Int Psychogeriatr 23, 344–355. [PubMed: 20716393]

[16]. Besser L, Kukull W, Knopman DS, Chui H, Galasko D, Weintraub S, Jicha G, Carlsson C, Burns J, Quinn J, Sweet RA, Rascovsky K, Teylan M, Beekly D, Thomas G, Bollenbeck M, Monsell S, Mock C, Zhou XH, Thomas N, Robichaud E, Dean M, Hubbard J, Jacka M, Schwabe-Fry K, Wu J, Phelps C, Morris JC (2018) Version 3 of the National Alzheimer's Coordinating Center's Uniform Data Set. Alzheimer Dis Assoc Disord 32, 351–358. [PubMed: 30376508]

[17]. Wang T, Qiu RG, Yu M (2018) Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. Sci Rep 8, 9161. [PubMed: 29907747]

[18]. Selkoe DJ (2012) Preventing Alzheimer's Disease. Science 337, 1488–1492. [PubMed: 22997326]

[19]. Hastie T, Tibshirani R, Friedman JH (2009) The elements of statistical learning: data mining, inference, and prediction, Springer, New York.

[20]. Ishwaran H, Kogalur UB (2023) Fast Unified Random Forests for Survival, Regression, and Classification (RF-SRC), manual. https://cran.r-project.org/package=randomForestSRC. Last updated June 01, 2022, Accessed on October 6, 2022.

[21]. Ishwaran H, Lu M, Kogalur UB (2022) randomForestSRC: Variable Importance (VIMP) with Subsampling Inference Vignette. http://randomforestsrc.org/articles/vimp.html. Last updated June 01, 2022, Accessed on December 6, 2022.

[22]. Harrell FE, Califf RM, Pryor DB, Lee KL, Rosati RA (1982) Evaluating the yield of medical tests. JAMA 247, 2543–2546. [PubMed: 7069920]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

[23]. Laimighofer M, Krumsiek J, Buettner F, Theis FJ (2016) Unbiased Prediction and Feature Selection in High-Dimensional Survival Regression. J Comput Biol 23, 279–290. [PubMed: 26894327]

[24]. Kamarudin AN, Cox T, Kolamunnage-Dona R (2017) Time-dependent ROC curve analysis in medical research: current methods and applications. BMC Med Res Methodol 17, 53. [PubMed: 28388943]

[25]. Zhu W, Zeng N, Wang N (2010) Sensitivity, specificity, accuracy, associated confidence interval and ROC analysis with practical SAS implementations. NESUG Proc Health Care Life Sci Baltim Md 19, 67.

[26]. Harrell FE Jr, with contributions from Charles Dupont and many others (2022) Hmisc: Harrell Miscellaneous. R package version 4.7–0 https://CRAN.R-project.org/package=Hmisc.

[27]. Blanche P, Dartigues J-F, Jacqmin-Gadda H (2013) Estimating and Comparing time-dependent areas under receiver operating characteristic curves for censored event times with competing risks. Stat Med 32, 5381–5397. [PubMed: 24027076]

[28]. Ishwaran H, Lu M, Kogalur UB (2022) randomForestSRC: Partial Plots Vignette. http://randomforestsrc.org/articles/partial.html. Last updated June 01, 2022, Accessed on December 6, 2022.

[29]. White RS, Barber JM, Harp JP, Jicha GA (2022) Examining the Effects of Formal Education Level on the Montreal Cognitive Assessment. J Am Board Fam Med JABFM jabfm.2022.AP.220093.

[30]. Chen T, He T (2023) xgboost: eXtreme Gradient Boosting. https://cran.r-project.org/web/packages/xgboost/vignettes/xgboost.pdf. Last updated March 31, 2023, Accessed on May 30, 2023.

[31]. Gauthier SG (2005) Alzheimer's disease: the benefits of early treatment. Eur J Neurol 12 **Suppl** 3, 11–16. [PubMed: 16144532] **Suppl**

[32]. Budd D, Burns LC, Guo Z, L'italien G, Lapuerta P (2011) Impact of early intervention and disease modification in patients with predementia Alzheimer's disease: a Markov model simulation. Clin Outcomes Res CEOR 3, 189–195.

[33]. Folch J, Busquets O, Ettcheto M, Sánchez-López E, Castro-Torres RD, Verdaguer E, Garcia ML, Olloquequi J, Casadesús G, Beas-Zarate C, Pelegri C, Vilaplana J, Auladell C, Camins A (2018) Memantine for the Treatment of Dementia: A Review on its Current and Future Applications. J Alzheimers Dis JAD 62, 1223–1240. [PubMed: 29254093]

[34]. Gill CJ, Mwananyanda L, MacLeod W, Kwenda G, Pieciak R, Etter L, Bridges DJ, Chikoti C, Chirwa S, Chimoga C, Forman L, Katowa B, Lapidot R, Lungu J, Matoba J, Mwinga G, Mubemba B, Mupila Z, Muleya W, Mwenda M, Ngoma B, Nkazwe R, Nzara D, Pawlak N, Pemba L, Saasa N, Simulundu E, Yankonde B, Thea DM (2022) Sustained high prevalence of COVID-19 deaths from a systematic post-mortem study in Lusaka, Zambia: one year later. medRxiv 2022.03.08.22272087.

[35]. Kumar S, Oh I, Schindler S, Lai AM, Payne PRO, Gupta A (2021) Machine learning for modeling the progression of Alzheimer disease dementia using clinical data: a systematic literature review. JAMIA Open 4, ooab052. [PubMed: 34350389]

[36]. Battineni G, Chintalapudi N, Hossain MA, Losco G, Ruocco C, Sagaro GG, Traini E, Nittari G, Amenta F (2022) Artificial Intelligence Models in the Diagnosis of Adult-Onset Dementia Disorders: A Review. Bioengineering 9, 370. [PubMed: 36004895]

[37]. Javeed A, Dallora AL, Berglund JS, Ali A, Ali L, Anderberg P (2023) Machine Learning for Dementia Prediction: A Systematic Review and Future Research Directions. J Med Syst 47, 17. [PubMed: 36720727]

[38]. Molnar C (2022) Interpretable Machine Learning: A Guide for Making Black Box Models Explainable. https://christophm.github.io/interpretable-ml-book/. Last updated December 14, 2022, Accessed on December 22, 2022.

[39]. Whetten AB, Stevens JR, Cann D (2021) The implementation of random survival forests in conflict management data: An examination of power sharing and third party mediation in post-conflict countries. PLoS ONE 16, e0250963. [PubMed: 33939757]

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

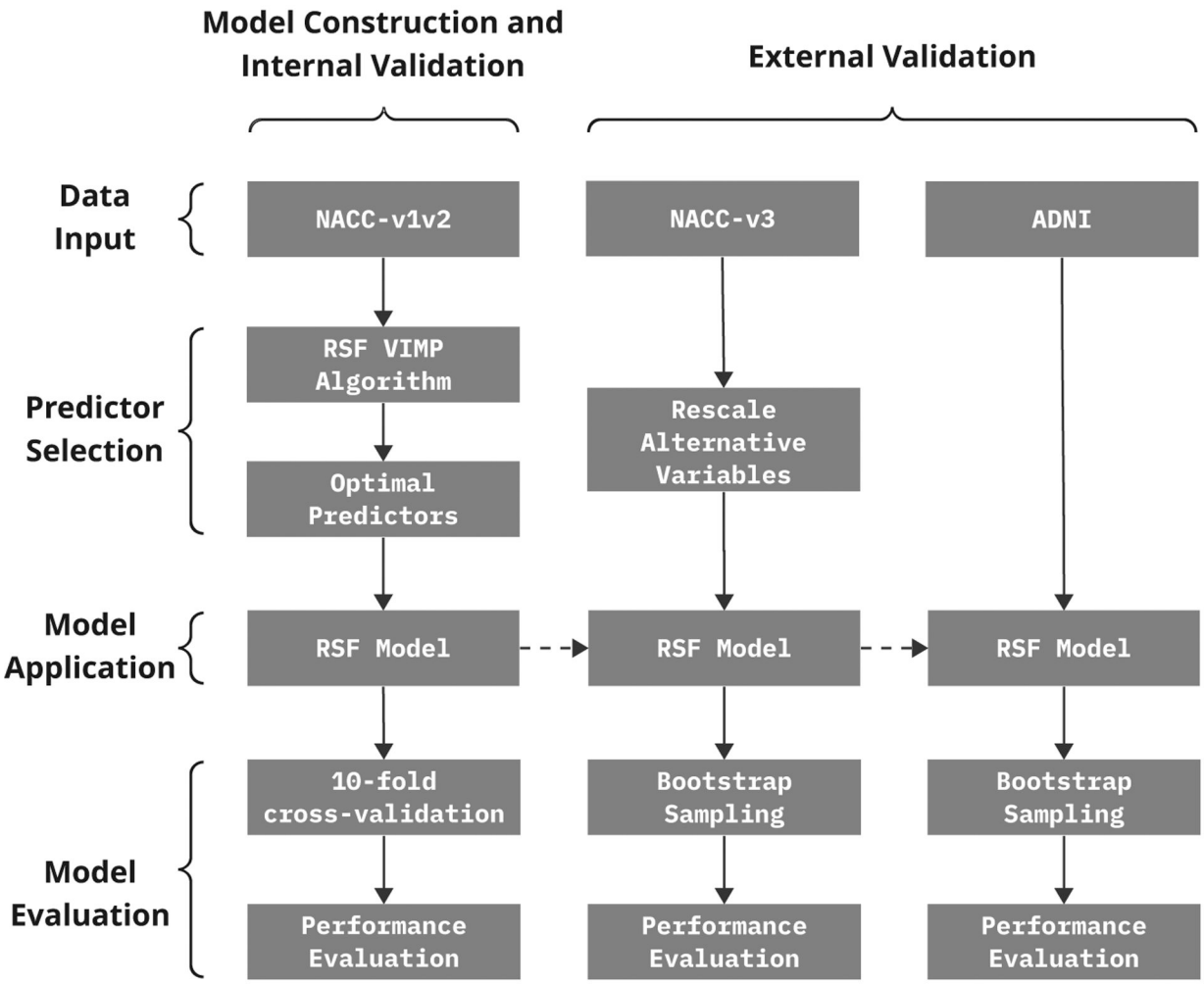

Figure 1.
Model Development and Validation Diagram

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

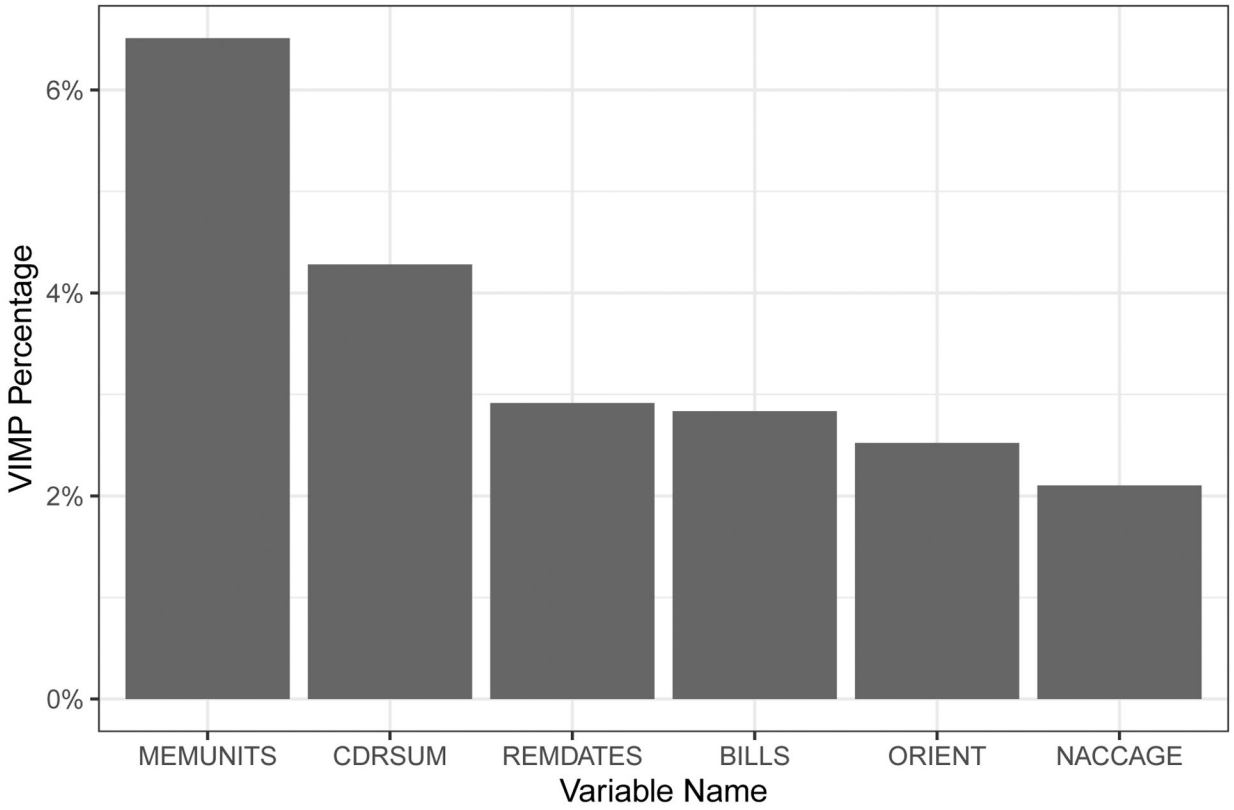

Figure 2.
Top 6 Variable Importance Percentages

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

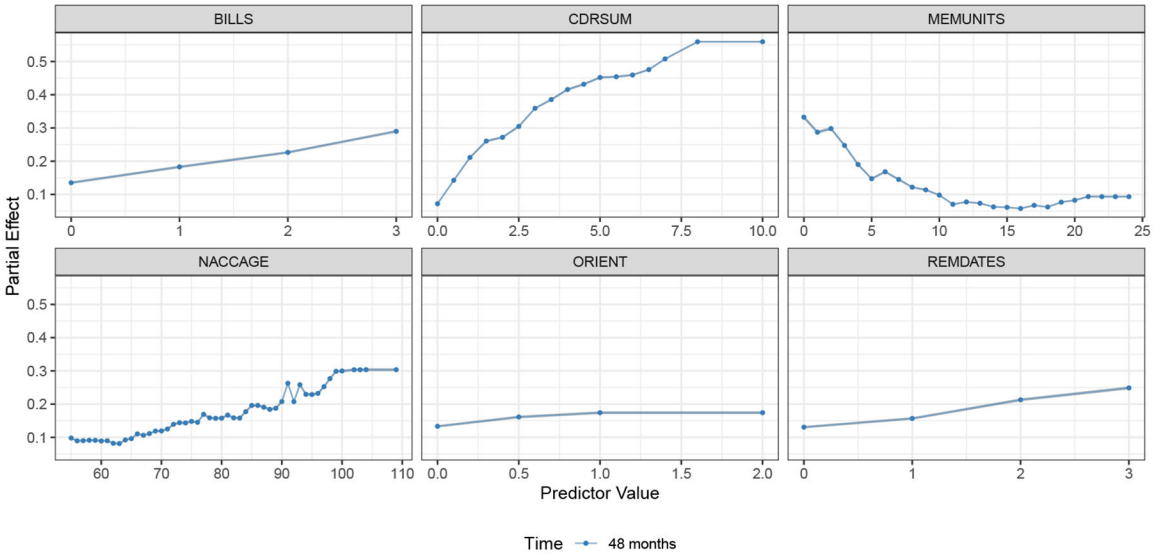

Figure 3.
Partial dependence plot

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 1

NACC-v1v2 Cohort Baseline Characteristics

Characteristics	Non-progressed (N = 12,722)	Progressed (N = 2,866)	P value	Overall (N = 15,588)
Follow-up Time *,months, mean (SD)	69.3 (47.5)	48.9 (39.8)	< 0.001	65.6 (46.9)
Age, years, mean (SD)	72.5 (8.54)	76.6 (7.88)	< 0.001	73.2 (8.60)
Sex, n (%)			< 0.001	
Male	5,022 (39.5%)	1,293 (45.1%)		6315 (40.5%)
Female	7,700 (60.5%)	1,573 (54.9%)		9273 (59.5%)
Education, years, mean (SD)	15.5 (3.20)	15.3 (3.30)	< 0.001	15.5 (3.20)
CDR-SB, mean (SD)	0.392 (0.772)	1.27 (1.24)	< 0.001	0.553 (0.940)
MMSE, mean (SD)	28.5 (1.81)	27.1 (2.43)	< 0.001	28.3 (2.02)
Subjective Cognitive Decline Among CN, n (%)			< 0.001	
Yes	2,183 (24.7%)	256 (32.4%)		2439 (25.3%)
No	6,656 (75.3%)	534 (67.6%)		7190 (74.7%)
Cognitive Status, n (%)			< 0.001	
CN	8,839 (69.5%)	790 (27.6%)		9,629 (61.8%)
Impaired-not-MCI	846 (6.6%)	184 (6.4%)		1,030 (6.6%)
MCI	3,037 (23.9%)	1,892 (66.0%)		4,929 (31.6%)

For continuous variables, the one-sided Mann-Whitney U test is used; for categorical variables, the Chi-square test is used. The NA entries were removed for continuous variables before calculating the summary statistics. The NA category for categorical variables was not considered in the Chi-square test. SD: standard deviation, CDR-SB: CDR® Dementia Staging Instrument - Sum of Boxes, MMSE: Mini Mental State Examination, CN: cognitively normal, MCI: mild cognitive impairment.

*The follow-up time is prior to event or censoring.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 2

Selected Top 6 Variables

Name in NACC-v1v2	**Name in ADNI**	**Description in NACC-v1v2**	**Designed Value Range in NACC-v1v2**
MEMUNITS	LDELTOTAL	Logical Memory IIA - Delayed - Total number of story units recalled	Integers from 0 to 25
NACCAGE	AGE	Subject age at the baseline visit	Integers from 55 to 120
CDRSUM	CDRSB	CDR® Dementia Staging Instrument - Sum of Boxes	0.0, 0.5, 1.0, 1.5, …, 18.0 (scores of 16.5 and 17.5 not possible)
REMEDATES	FAQREM N 9	In the past four weeks, did the subject have any difficulty or need help with: Remembering appointments, family occasions, holidays, medications	0 = Normal; 1 = Has difficulty, but does by self; 2 = Requires assistance; 3 = Dependent
BILLS	FAQFINAN N 1	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
ORIENT	CDORIENT	Orientation sub question in CDR	0.0, 0.5, 1.0, 2.0,3.0
Name in NACC-v1v2	**Surrogate Variable in NACC-v3**	**Description in NACC-v3**	**Designed Value Range in NACC-v3**
MEMUNITS	CRAFTDVR	Craft Story 21 Recall (Delayed) — Total story units recalled, verbatim scoring	Integers from 0 to 44

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 3

NACC-v3 Cohort Baseline Characteristics

Characteristics	Non-progressed (N = 5,038)	Progressed (N = 448)	P value	Overall (N = 5,486)
Follow-up Time *, months, mean (SD)	31.3 (16.5)	23.7 (13.7)	< 0.001	30.7 (16.5)
Age, years, mean (SD)	70.7 (7.45)	73.7 (7.43)	< 0.001	71.0 (7.50)
Sex, n (%)			< 0.001	
Male	1,962 (38.9%)	222 (49.6%)		2184 (39.8%)
Female	3,076 (61.1%)	226 (50.4%)		3302 (60.2%)
Education, years,mean (SD)	16.2 (2.90)	16.1 (3.22)	0.47	16.2 (2.90)
CDR-SB, mean (SD)	0.426 (0.768)	1.80 (1.18)	< 0.001	0.538 (0.893)
MMSE, mean (SD)	27.7 (2.44)	24.4 (2.01)	< 0.001	27.5 (2.60)
Subjective Cognitive Decline Among CN, n (%)			0.47	
Yes	952 (28.3%)	12 (35.3%)		964 (28.3%)
No	2,417 (71.7%)	22 (64.7%)		2,439 (71.7%)
NA	1 (0.0%)	0 (0%)		1 (0.0%)
Cognitive Statis, n (%)			< 0.001	
CN	3,370 (66.9%)	34 (7.6%)		3,404 (62.0%)
Impaired-not-MCI	280 (5.6%)	12 (2.7%)		292 (5.3%)
MCI	1,388 (27.6%)	402 (89.7%)		1,790 (32.7%)

For continuous variables, the one-sided Mann-Whitney U test is used; for categorical variables, the Chi-square test is used. The NA entries were removed for continuous variables before calculating the summary statistics. The NA category for categorical variables was not considered in the Chi-square test. SD: standard deviation, CDR-SB: CDR® Dementia Staging Instrument - Sum of Boxes, MMSE: Mini Mental State Examination, CN: cognitively normal, MCI: mild cognitive impairment.

*The follow-up time is prior to event or censoring.

Author Manuscript Author Manuscript Author Manuscript Author Manuscript

Table 4

ADNI Cohort Baseline Characteristics

Characteristics	Non-progressed (N = 1,364)	Progressed (N = 405)	P value	Overall (N = 1,769)
Follow-up Time*, months, mean (SD)	57.1 (42.3)	35.1 (30.2)	< 0.001	52.1 (40.9)
Age, years, mean (SD)	72.6 (6.97)	74.0 (6.92)	< 0.001	72.9 (6.98)
Sex, n (%)			0.0064	
Male	701 (51.4%)	240 (59.3%)		941 (53.2%)
Female	663 (48.6%)	165 (40.7%)		828 (46.8%)
Education, years, mean (SD)	16.3 (2.68)	15.9 (2.73)	0.0025	16.2 (2.7)
CDR-SB, mean (SD)	0.599 (0.826)	1.73 (1.05)	< 0.001	0.859 (1.00)
MMSE, mean (SD)	28.6 (1.55)	27.3 (1.87)	< 0.001	28.3 (1.72)
Cognitive Status, n (%)			< 0.001	
CN	750 (55.0%)	35 (8.6%)		785 (44.4%)
MCI	614 (45.0%)	370 (91.4%)		984 (55.6%)

For continuous variables, the one-sided Mann-Whitney U test is used; for categorical variables, the Chi-square test is used. The NA entries were removed for continuous variables before calculating the summary statistics. The NA category for categorical variables was not considered in the Chi-square test. SD: standard deviation, CDR-SB: CDR® Dementia Staging Instrument - Sum of Boxes, MMSE: Mini Mental State Examination, CN: cognitively normal, MCI: mild cognitive impairment.

*The follow-up time is prior to event or censoring.

Author Manuscript

Author Manuscript

Author Manuscript

Author Manuscript

Table 5

NACC-v1v2 Cohort Baseline Characteristics

Fold	C index	Accuracy	AUC
1	84.32	89.69	89.84
2	85.06	90.34	91.31
3	84.02	89.89	89.4
4	85.35	89.72	89.77
5	86.72	90.56	91.74
6	85.81	90.18	90.67
7	85.97	89.47	91.35
8	84.59	90.68	89.30
9	84.92	90.39	90.00
10	85.35	89.69	89.21
Average	85.21	90.061	90.26

AUC: area under curve. Accuracy and AUC are both 48-month time-dependent.