Title: Cohort Discovery and Risk Stratification for AD: An EHR-based Approach

Authors & Affiliations:

Donna Tjandra[1], Raymond Q. Migrino[2] MD, Bruno Giordani, PhD[3], Jenna Wiens, PhD[1]

[1]Computer Science and Engineering, University of Michigan, Ann Arbor, MI, USA;

[2]Phoenix Veterans Affairs Health Care System, Phoenix, AZ 85012;

[3]Neuropsychology Program, Department of Psychiatry, University of Michigan Ann Arbor, MI, USA

Corresponding Author:

Jenna Wiens, PhD

2260 Hayward Street

University of Michigan,

Ann Arbor, MI 48109

email: wiensj@umich.edu

Key Words: machine learning, electronic health record, cohort discovery, early prediction

Word Count: 3,485/3,500

ABSTRACT

Background

We sought to leverage data routinely collected in electronic health records (EHRs), with the goal of developing patient risk stratification tools for predicting risk of developing AD.

Method

Using EHR data from the University of Michigan hospitals (UM) and consensus-based diagnoses from the Michigan Alzheimer's Disease Research Center, we developed and validated a cohort discovery tool for identifying patients as AD. Applied to all UM patients, these labels were used to train an EHR-based machine learning model for predicting AD onset within 10 years.

Results

Applied to a test cohort of 1,697 UM patients, the model achieved an area under the receiver operating characteristics curve of 0.70 (95% CI=0.63-0.77). Important predictive factors included cardiovascular factors and laboratory blood testing.

Conclusion

Routinely collected EHR data can be used to predict AD onset with modest accuracy. Mining routinely collected data could shed light on early indicators of AD appearance and progression.

Word Count: 150/150

1. INTRODUCTION

Research in Context (149/150)

Systematic Review: We searched the literature for reports on predictive modeling and cohort discovery in AD. Previous research has analyzed data not routinely collected in clinical care, has focused on relatively short prediction horizons (e.g., 3 years) or is limited in the scope of EHR (electronic health record) data considered.

Interpretation: We developed and validated an EHR-based cohort discovery tool for AD patients. This tool facilitates analyses of EHR data without requiring manual chart review. Using this tool, we developed and validated an EHR-based model for predicting AD onset up to 10 years. Covariates associated with the outcome align in part with the AD literature. Novel associations included forms of healthcare utilization and urine tests. Such findings can be used to stimulate hypothesis generation and/or aid in longitudinal study recruitment.

Future Directions: Associations identified by our model require further investigation. Model performance could be improved with additional longitudinal data.

Alzheimer's disease (AD), the most common form of dementia[1], affects approximately 5.8 million Americans[1], and that number is expected to more than double by 2050[1]. The physiological changes in the brain associated with AD, including β-amyloid and tau buildup,

are currently suspected to take place at least 20 years before symptom onset[1]. Earlier identification of at-risk individuals could lead to earlier and more effective treatment.

Predictive modeling for AD risk has focused on AD-specific biomarkers such as cerebrospinal fluid (CSF), neuropsychological test scores, and complex medical imaging[2-16]. These are not routinely collected in clinical care, and thus apply to only to a subset of individuals for which these data are available. Importantly, because collection of these biomarkers can be invasive or involve significant cost/logistics, they are rarely obtained during the pre-clinical stage, limiting current predictive ability of these biomarkers to short-term horizons (e.g., 2-4 years)[2-5,10-13]. In contrast, we aimed to leverage existing databases of routinely collected electronic health record (EHR) data to develop predictive models for AD that can identify at-risk individuals up to a decade in advance.

EHRs often contain decades of longitudinal clinical data (e.g., medications) for thousands of patients[17]. However, these data have been largely underutilized in studying pre-clinical signs of AD progression[18-21]. The ability to automatically identify signs of AD patients using available data in EHRs would increase the feasibility of downstream computational analyses on large-scale datasets without the need for labor-intensive chart review. To this end, we first developed and validated a cohort discovery tool that can be applied to EHR data for automatic classification of AD individuals. Second, we applied this tool to a large cohort of patients and used machine learning techniques to develop and validate a model for estimating patient risk of developing AD within a 10-year prediction horizon. Applied more broadly, such an approach could help in identifying risk factors that arise well in advance of clinical symptoms.

2. METHODS

We describe the inclusion/exclusion criteria that were applied to two datasets to obtain our study cohorts, one for building the cohort discovery tool and another for building the predictive model.

2.1. Study Cohorts

Our analyses relied on two study cohorts i) the cohort discovery tool–cohort and ii) the risk stratification model–cohort. These cohorts were extracted from the Michigan Alzheimer's Disease Research Center (Michigan-ADRC) and the University of Michigan's Research Data Warehouse (RDW). The Michigan-ADRC, which focuses on memory and aging research, contains data for 789 participants from ~2005–2019. All participants received a consensus-based clinical diagnosis using the National Alzheimer's Coordinating Center Uniform Dataset criteria[22,23]. The RDW contains records of patient encounters (defined as inpatient and outpatient visits) with Michigan Medicine for over 4 million patients dating from ~2000–2019. These data consist of all clinical data associated with the encounter (e.g., medications). This study was approved by the Institutional Review Board at the University of Michigan.

The first *cohort discovery tool-cohort*, included all Michigan-ADRC participants with at least one RDW encounter at or after the age of 65 years. Only this age group was considered, since most cases of AD occur in that population[1]. Our second cohort, the *risk stratification model-cohort*, included patients with at least one RDW encounter between the ages of 68-72 years who had at least 10 years of follow-up or who converted to AD within 10 years. This age range allowed for a relatively large study population. We excluded patients with an AD

diagnosis before 68 years. Here, AD refers to *probable* AD, since AD cannot be officially diagnosed until after death and because this diagnosis was commonly used throughout this period.

2.2 Cohort Discovery Tool

Using diagnoses provided by the Michigan-ADRC, we investigated the accuracy of different EHR-based rules for identifying AD patients in RDW. Each rule aimed to identify RDW encounters associated with patients with an AD diagnosis and was based on EHR variables related to AD: diagnosis codes for AD, medications for AD, procedure codes for psychological/cognitive testing, and procedure codes involving moderate to high complexity medical decision making (details in Appendix 1 (A1)). For example, one rule labeled RDW encounters with a current or previous AD diagnosis code and prescription to an AD-associated medication as AD. We also evaluated an existing tool from the Phenotype Knowledge Base (PheKB)[20], which labeled those with at least five encounters with a dementia diagnosis code or prescription for an AD-associated medication as AD. Applied to a set of encounters in RDW for a patient, the first encounter that met the EHR-based criteria was labeled as 'AD' by the cohort discovery rule. Since AD is currently irreversible[1], we labeled subsequent encounters as 'AD'.

The labels produced by each EHR-based rule were compared to the Michigan-ADRC diagnoses at the patient level. Michigan-ADRC participants are followed longitudinally, and thus may have multiple timestamped diagnoses (e.g., cognitively normal, mild cognitive impairment, AD). As ground truth, we labeled the six months preceding the first AD

diagnosis in the Michigan-ADRC and anytime thereafter as AD. Prior study has shown that clinical diagnoses of AD have good diagnostic accuracy to histopathology-confirmed AD[24]. If a patient was never diagnosed with AD, then we considered them 'not AD' until six months after their last Michigan-ADRC diagnosis. Using these time frames as ground truth, comparisons to the corresponding RDW encounters were made as follows (Figure 1). Only those whose RDW and ground truth time windows overlapped were included while counting. If at least one AD-diagnosed RDW encounter was within the Michigan-ADRC-defined AD window, the patient was considered to have been correctly identified by the EHR-based rule (true positive). False positives were counted as those with at least one AD-diagnosed RDW encounter but no Michigan-ADRC diagnosis for AD within the Michigan-ADRC-defined AD time window. True negatives were counted as those not identified by the EHR-based rule and never received a Michigan-ADRC diagnosis for AD. False negatives were counted as those not identified by the EHR-based rule, but had a Michigan-ADRC diagnosis for AD.

Results were summarized by the true positive rate (sensitivity), 1-false positive rate (specificity), positive predictive value (PPV), and F1 score (F1). We measured a population-adjusted PPV since the Michigan-ADRC dataset is enriched compared to the general population (details in A2).

When evaluating EHR-based rules against each other, we prioritized maximizing the F1 score to balance the population-adjusted PPV and sensitivity. In the case of ties, we considered the adjusted PPV, unadjusted PPV, specificity, and sensitivity, in that order.

Given the rule with highest F1 score, we evaluated *when* patients received the diagnosis within RDW relative to the Michigan-ADRC, by measuring the time from the first AD Michigan-ADRC diagnosis to the first AD-labeled encounter in RDW. We also examined our ability to identify AD at the encounter level. Using the ground truth labels outlined earlier, a confusion matrix was constructed to show the number of encounters (AD/not AD) that were correctly and incorrectly identified by the EHR-based rule. Results are reported as the median with an empirical 95% confidence interval over 1,000 bootstrapped samples.

2.3 Predictive Model

In the following sections, we frame the problem of predicting AD over a 10-year horizon using EHR-extracted data. We describe feature engineering, including which EHR components were used, and model training. We then describe model evaluation in terms of predictive performance and influential features.

2.3.1 Outcome

To control for the effect of age on risk of developing AD, we aligned patients in our risk stratification cohort (Section 2.1) based on their earliest visit between 68-72 years. Patients were labeled according to the cohort discovery tool (Section 2.2) as converting to AD within 10 years or not. The date of conversion was defined as the date of the first encounter meeting the cohort discovery tool's criteria. Patients were labeled positive if they converted within 10 years of alignment and negative otherwise.

2.3.1 Variable Extraction

Given the 'alignment' visit, each patient was represented by a high-dimensional feature vector summarizing all encounters in the 1,000 days before alignment. A look-back period of 1,000 days was chosen based on the median length of available history. We extracted data pertaining to diagnoses (ICD9 (international classification of disease) codes), procedures (CPT (current procedural terminology) codes), medications (medication name, ingredient name, and VA class code), laboratory results (LOINC (Logical Observation Identifiers Names and Codes) and result values), vital sign data (e.g., temperature), healthcare utilization (e.g., encounter types), and demographic information (e.g., race). Features were categorized as "time-invariant" or "time-dependent". Time-invariant features were patient characteristics that do not change over time (e.g., race), and time-dependent features were those associated with a specific encounter or timestamp (e.g., diagnoses). Data were pre-processed with FIDDLE (FlexIble Data-Driven pipeLinE)[25], using a time window of 250 days, a pre- and post-filter threshold of 0.0003, and a frequency threshold of 1.0. Feature vectors for each patient were formed from FIDDLE's output by concatenating their time invariant data to their time-dependent data up to 1,000 days before alignment.

2.3.2 Model Training

Data were split using an 80%-20% training-test random stratified split. Using the training data, we performed model selection. Minimizing the L2-regularized hinge loss, we trained a linear-support vector machine to predict AD onset for patients aligned between 68-72 years over a 10-year horizon. The amount of regularization was tuned using five-fold cross-validation on the training set, sweeping C=[0.001-1,000] on a logarithmic scale. Analyses were performed in Python 3.6 using SciKitlearn[26].

2.3.3 Model Evaluation

Overall performance of our predictive model was measured using the area under the receiver operating characteristics curve (AUROC) and a confusion matrix measuring sensitivity, specificity, positive predictive value, and accuracy based on a threshold at the 65th percentile on the held-out test set. We measured model calibration using the Brier score[27] (details in A3). Additionally, we examined the model's ability to classify AD converters among patients with memory impairments, reporting the AUROC and confusion matrix (details and results in A9). We report all model evaluation results as empirical 95% confidence intervals generated using 1,000 bootstrapped samples unless otherwise stated.

We also assessed the model's ability to predict over the 10-year horizon by examining the number of correctly predicted converters with respect to their time to conversion (time between alignment and first AD diagnosis). Since the model outputs a continuous risk score, we classified patients as 'high risk' if their risk score was above the 65th percentile and as low-risk otherwise. We examined five non-overlapping conversion windows, reporting the sensitivity for each.

Beyond model performance, we examined which categories of EHR information (e.g., diagnoses vs. procedures) were the most informative for prediction by comparing the AUROCs on models trained with different subsets of features (e.g., training only on diagnosis features or training only on procedural features).

We also analyzed the model's most important features using permutation importance[28], where any decrease in AUROC was measured by permuting all patient values within a feature or group of correlated features (R≥|0.7|). The most important features were identified as those with the largest drop in AUROC, taken as the median over 100 permutations and whose lower bound on an empirical 95% confidence interval was above zero.

3. RESULTS

In the following sections, we identify the best EHR-based rule for cohort discovery. We then summarize performance of the predictive model in terms of AUROC, calibration, and learned risk factors.

3.1. Cohort Discovery Tool

From 789 Michigan-ADRC volunteers, 624 (79%) volunteers 65 years and older had encounters with Michigan Medicine (details in A4). 24.8% of the 624 volunteers converted to AD.

Among several cohort discovery rules (Figure 2), the one that best identified AD patients included those with a diagnosis code for AD (Table A.1) (median F1-score=0.73 (95% CI=0.68-0.78), median adjusted PPV=0.77 (95% CI=0.71-0.82), median sensitivity=0.70 (95% CI=0.65-0.74)). The PheKB tool[20] performed significantly worse in terms of median F1-

score=0.55 (95% CI=0.48-0.62 p<0.05) and median sensitivity=0.45 (95% CI=0.31-0.51 p<0.05). Significance was determined by whether the upper bound of the 95% CI for the F1 score was below the lower bound F1 score of the best rule.

Among the true positives identified by our best rule, the first RDW diagnosis occurred 177 days before (95% CI=278 before-68 days after) the first AD Michigan-ADRC diagnosis. At the encounter level, this rule yielded a median PPV of 0.59 (95% CI=0.56-0.63) and a median sensitivity of 0.82 (95% CI=0.72-0.83) (details in A5).

3.2. Predictive Model

Applying the cohort-discovery rule with the highest F1-score to RDW (Figure 3) yielded a study population of 8,474 patients, where 4.14% converted to AD within 10 years from alignment (Table 1). 268 time-invariant features and 3,963 time-dependent features per time window across four time windows were used for prediction (feature breakdown in A6). The training and test sets consisted of 6,777 and 1,697 patients, respectively.

On the test set, we achieved an AUROC of 0.70 (95% CI=0.63-0.77) (Figure A2.a) and a Brier score of 0.028 (95% CI=0.025-0.029) (Figure A.1). Thresholding at the 65th percentile, we achieved a sensitivity of 0.62 (95% CI=0.60-0.63), a specificity of 0.66 (95% CI=0.65-0.66), and a positive predictive value of 0.07 (95% CI=0.05-0.09), for an overall accuracy of 0.66 (95% CI=0.65-0.66) (Table A.5).

The model predicted AD onset over long and short prediction horizons with high sensitivity (Figure A.3); though, performance generally decreased as the prediction horizon increased. 87% patients who converted within 2.5 years of alignment were correctly predicted, while only 53% of patients who converted within 8.4-10 years of alignment were correctly predicted. The distribution of time to conversion was left skewed, with most patients converting more than 6 years post-alignment.

Overall, data on laboratory test results, procedures, and healthcare utilization had the most predictive power (Figure 4a, Figure A.2b). Predicting using laboratory test results alone was able to achieve an AUROC of 0.62 (95% CI=0.55-0.69). However, the best performance was achieved when all categories were combined. Using longitudinal data from all previous encounters up to 1,000 days prior to alignment also improved performance, compared to when data from only the encounter of alignment was used AUROC=0.54 (95% CI=0.47-0.61) (Figure 4b). The top 10 important features pertained to healthcare utilization, procedures involving laboratory blood testing, and cardiovascular risk factors (Figure 4c, Table 2), with the median drop in AUROC between 0.002-0.040.

4. DISCUSSION

Research in predicting AD risk[2-16] has focused on datasets specifically curated for the purpose of studying AD (e.g., ADNI)[29]. While such studies can be used to identify predictors of disease progression, many of the studied variables, e.g., CSF composition, are not collected during routine clinical care, especially in the decades before symptom onset. Moreover, because of the costs associated with such data collection, study populations are

relatively small (~1,700 patients) and prediction horizons relatively short (2-4 years). In contrast, EHR data consist of routinely collected data, have been collected for over a decade at some institutions, and are available for a large portion of the population, as highlighted by Stephan et al.[30]. Given this potential, we sought to explore the utility of EHRs in modeling the progression of AD 10 years before clinical diagnosis. We developed an automated EHR-based cohort discovery tool for identifying AD patients and then applied this tool to a large cohort of patients aligned between 68-72 years. Using these data and machine learning techniques, we developed a model for predicting AD conversion within 10-years.

While EHR data have been leveraged to model other conditions[31-34], they have been largely underutilized in modeling AD progression. Most related studies focus on cohort discovery[18,20,35], characterizing the incidence of AD[19], and modeling the risk of dementia more generally while controlling for age to a lesser extent[36,37]. We differ from previous work in that we focus on only AD while prior work has focused on AD and related dementias. We chose to focus on AD alone, since it is the most common form of dementia. Previously proposed identification rules required at least five encounters with a dementia diagnosis code or AD associated medication[20]. On RDW, this rule had a lower F1 score than our best labeling rule. In addition we differ from previous risk stratification models in that we consider AD specifically[36,37], use a 10 year horizon instead of five years or less[21,36], and focus on a wider set of input covariates or potential risk factors[21,36,37]. We also control for age to a larger extent, as it has been demonstrated that previous models performed similarly to predicting on age alone[37,38].

Compared to curated datasets like ADNI, EHR data present additional challenges. In the context of AD, EHRs do not have a set of ground truth diagnoses. We relied on the fact that a subset of individuals in RDW were also volunteers in the Michigan-ADRC for whom we had ground truth diagnoses. In addition, data from prospective studies such as ADNI are collected at fixed time intervals, while EHR data are irregularly sampled.

Despite these challenges, there are many advantages in working with EHR data. First, EHR data may contain more longitudinal data per patient than ADNI. For example, 25% of ADNI participants had more than 10 encounters, compared to over 50% in our study population. This allowed us to predict AD onset over longer horizons (10 years) with modest performance. Approximately half of the patients who converted between 8.4-10 years after alignment were correctly predicted to convert, demonstrating the possibility of early detection. The ability to predict over longer horizons could be crucial, as the physiological changes in the brain are suspected to take place at least 20 years before symptom onset[1]. Over time, as more EHR data are collected, we may be able to improve model performance and investigate longer time horizons. Second, study populations from ADNI are highly enriched with AD individuals and AD-specific data, while EHR-derived study populations are more likely to represent the general population and the types of data routinely available. Our predictive model was able to identify laboratory tests and procedures associated with AD onset up to 10 years in advance. While identification of EHR variables known to be associated with AD for model building is useful, EHR variables with no known association to AD could lead to the discovery of unknown biological mechanisms, interactions and novel biomarkers. Similarly, an EHR-based predictive tool may be used in a cost-effective strategy

to screen which at-risk patients should undergo testing using more invasive (e.g., CSF fluid) or imaging-based established biomarkers, at earlier time-points than currently practiced.

Many of the features identified as important matched the literature. However, causal relationships between them and AD progression cannot be inferred from our study alone. In particular, features related to healthcare utilization appeared to be strong predictors. This has been supported by previous work that showed an increase in healthcare utilization one year before AD diagnosis[39,40]. In addition, many of the important features related to laboratory blood tests have been previously associated with AD. Specifically, Chen et al. and Winchester et al. found that changes in blood cell composition may be associated with AD development[41,42]. Wang et al. found an association between vitamin B12 and AD development[43]. In line with Cao et al. and Le Page et al., we identified immune system biomarkers as beneficial in early detection[44,45]. In terms of comorbidities we identified as associated with increased risk, hypertension has previously been identified[46]. In addition, urine tests are associated with diabetes testing[47], another related risk factor[48]. In terms of procedures, mastoid procedures could act as a possible surrogate for hearing loss, which has been suspected to be associated with AD[49]. Correlated to mastoid procedures were vaccinations. This may because at-risk patients are in an overall poorer state of health, making them more susceptible to infection and disease. As a result, clinicians may have been more inclined to vaccinate.

Our study is not without limitation. We relied on imperfect labels from our cohort discovery tool. As a result, the model may not generalize to predicting the full spectrum of patients that convert to AD. Even with perfect performance from the model (AUROC=1), reliance on the

cohort discovery tool limits us to an AUROC of 0.85 relative to the ground truth. In addition, inaccuracies in labeling the date of AD onset may introduce additional noise to the labels. Another limitation stems from our decision to exclude censored patients. We excluded censored patients since they did not have sufficient follow-up to assign a label. Going forward, approaches for incorporating censored patients could increase the size of the study population. Furthermore, although we aligned patients between 68-72 years to control for the effects of age on our prediction task, age appeared as an important predictor. Though aligning patients at a single age (e.g., 68 years) could have mitigated this effect, this ultimately would have decreased the size of the study population.

In summary, we demonstrated the potential for EHRs as a novel source of data for developing models that characterize AD progression. Going forward, such analyses could be applied to other EHRs to generate hypotheses regarding novel early predictors and mechanisms of AD. In addition, longitudinal clinical studies involving early interventions may selectively target recruitment efforts toward 'at-risk' patients well before symptom onset.

5. ACKNOWLEDGMENTS

This research program is supported by the NIH/NIA funded Michigan Alzheimer's Disease Center (5P30AG053760) and the National Science Foundation (IIS-1553146). The views and conclusions in this document are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of the NSF or the NIH. The authors have no conflicts of interest to report.

6. REFERENCES

1. Alzheimer's Association. 2019 Alzheimer's Disease Facts and Figures. *Alzheimers Dement* 2019; **15**(3) 321-87.

2. Dukart J, Sambataro F and Bertolino A. Accurate Prediction of Conversion to Alzheimer's Disease using Imaging, Genetic, and Neuropsychological Biomarkers. *Alzheimers Dement.* 2016; **49** 1143–1159.

3. Llano DA, Bundela S, Mudar RA and Devanarayan V. A multivariate predictive modeling approach reveals a novel CSF peptide signature for both Alzheimer's Disease state classification and for predicting future disease progression. *PLOS One.* 2017.

4. Grassi M, Perna G, Caldirola D, Schruers K, Duara R and Loewenstein DA. A Clinically Translatable Machine Learning Algorithm for the Prediction of Alzheimer's Disease Conversion in Individuals with Mild and Premild Cognitive Impairment. *Alzheimers Dement.* 2018; **61** 1555–1573.

5. Young J, Modat M, Cardoso MJ, Mendelson A, Cash D and Ourselin S. Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. *NeuroImage: Clinical* 2013; **2** 735–745.

6. Thung KH, Yap PT, Adeli E, Lee SW and Shen D. Conversion and time-to-

conversion predictions of mild cognitive impairment using low-rank affinity pursuit denoising and matrix completion. *Medical Image Analysis* 2018; **45** 68–82.

7. Li Y, Yang T, Zhou J and Ye J. Multi-Task Learning based Survival Analysis for Predicting Alzheimer's Disease Progression with Multi-Source Block-wise Missing Data. *In Proceedings of the 2018 SIAM International Conference on Data Mining* 2018; 288-296.

8. Zhang C, Adeli E, Zhou T, Chen X and Shen D. Multi-Layer Multi-View Classification for Alzheimer's Disease Diagnosis. *In Proceedings of The AAAI Conference on Artificial Intelligence (AAAI-18)* 2018; 2314-2321.

9. Huang J and Alexander D. Probabilistic Event Cascades for Alzheimer's disease. *In Proceedings of Advances in Neural Information Processing Systems (NeurIPS)* 2012

10. Wang T, Qiu RG, and Yu M. Predictive Modeling of the Progression of Alzheimer's Disease with Recurrent Neural Networks. *Nature Scientific Reports* 2018; **8**(9161).

11. Li J, Rong Y, Meng H, Lu Z, Kwok T and Cheng H. TATC: Predicting Alzheimer's Disease with Actigraphy Data. *In Proceedings of Knowledge Discovery and Data Mining (KDD)* 2018; 509-518.

12. Lu D, Popuri K, Ding GW, Balachandar R and Beg MF. Multimodal and Multiscale Deep Neural Networks for the Early Diagnosis of Alzheimer's Disease using structural MR and FDG-PET. *Nature Scientific Reports* 2018; **8**(5697).

13. Lee G, Nho K, Kang B, Sohn KA and Kim D. Predicting Alzheimer's disease progression using multi-modal deep learning approach. *Nature Scientific Reports* 2019; **9**(1952).

14. Preische O, Schultz SA, Apel A, Kuhle J, Kaeser SA, Barro C et. al. Serum neurofilament dynamics predicts neurodegeneration and clinical progression in presymptomatic Alzheimer's disease. *Nature Medicine* 2019; **25** 277-283.

15. Ashton NJ, Nevado-Holgado AJ, Barber IS, Lynham S, Gupta V, Chatterjee P et. al. A plasma protein classifier for predicting amyloid burden for preclinical Alzheimer's disease. *Sci. Adv.* 2019; **5**.

16. Mattsson N, Insel PS, Donohue M, Jogi J, Ossenkoppele R, Olsson T et. al. Predicting diagnosis and cognition with ^{18}F-AV-1451 tau PET and structural MRI in Alzheimer's disease. *Alzheimers Dement.* 2019; **15** 570–580.

17. Menachemi N and Collum TH. Benefits and drawbacks of electronic health record systems. Risk Manag Healthc Policy 2011; **4** 47–55.

18. Jaakkimainen RL, Bronskill SE, Tierney MC, Herrmann N, Green D, Young J *et. al.* Identification of Physician-Diagnosed Alzheimer's Disease and Related Dementias in Population-Based Administrative Data: A Validation Study Using Family Physicians' Electronic Medical Records. *Journal of Alzheimer's Disease*, 2016; **54**(1) 337-349.

19. Perera G, Pedersen L, Ansel D, Alexander M, Arrighi HM, Avillach P *et. al.* Dementia prevalence and incidence in a federation of European Electronic Health Record databases: The European Medical Informatics Framework resource. *Alzheimers Dement.* 2018; **14**(2) 130–139.

20. Carlson C. Group Health Cooperative. Dementia. PheKB; 2012 Available from: https://phekb.org/phenotype/10

21. Zhang XS, Tang F, Dodge HH, Zhou J, and Wang F. MetaPred: Meta-Learning for Clinical Risk Prediction with Limited Patient Electronic Health Records. *In Proceedings of KDD*. 2019.

22. Morris JC, Weintraub S, Chui HC, Cummings J, Decarli C, Ferris S *et. al.* The Uniform Data Set (UDS): clinical and cognitive variables and descriptive data from Alzheimer Disease Centers. *Alzheimer Dis Assoc Disord*. 2006; **20**(4) 210-216.

23. Weintraub S, Besser L, Dodge HH, Teylan M, Ferris S, Goldstein FC et. al. Version 3 of the Alzheimer Disease Centers' Neuropsychological Test Battery in the Uniform Data Set (UDS). *Alzheimer Dis Assoc Disord.* 2018; **32**(1) 10-17.

24. Galasko D, Hansen LA, Katzman R, Wiederholt W, Masliah E, Terry R et. al. Clinical-Neuropathological Correlations in Alzheimer's Disease and Related Dementias. *Arch Neurol.* 1994; **21** 888-895.

25. Sjoding M, Tang S, Davarmanesh P, Song Y, Koutra D, and Wiens J. Democratizing EHR Analyses – A Comprehensive, Generalizable Pipeline for Learning from Clinical Data. *In Proceedings of MLHC.* (Clinical Abstract). 2019

26. Scikit-learn: Machine Learning in Python. Pedregosa et al., *JMLR* 2011; (**12**) 2825-2830.

27. Brier GS. Verification of Forecasts Expressed in Terms of Probability. *Monthly Weather Review*. 1950; **78** 1–3.

28. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. *Bioinformatics 2010*; **26**(10) 1340-1347.

29. ADNI Study Information. http://www.adni-info.org/.

30. Stephan BCM, Kurth T, Matthews FE, Brayne C, and Dufouil C. Dementia risk prediction in the population: are screening models accurate? *Nat. Rev. Neurol*. 2010; (**6**) 318-326.

31. Walsh CG, Ribeiro JD and Franklin JC. Predicting Risk of Suicide Attempts Over Time Through Machine Learning. *Clinical Psychological Science* 2017; **5**(3) 457-469.

32. Zheng T, Xie W, Xu L, He X, Zhang Y, You M et. al. A machine learning-based framework to identify type 2 diabetes through electronic health records. *International Journal of Medical Informatics* 2017; **97** 120–127.

33. Zeiberg D, Prahlad T, Nallamothu BK, Iwashyna TJ, Wiens J et. al. Machine learning for patient risk stratification for acute respiratory distress syndrome. *PLOS One.* 2019.

34. Oh J, Makar M, Fusco C, McCaffrey R, Rao K, Ryan EE et. al. A Generalizable, Data-Driven Approach to Predict Daily Risk of Clostridium difficile Infection at Two Large Academic Health Centers. *Infect Control Hosp Epidemiol* 2018; **39**(4) 425-433.

35. Ford E, Greenslade N, Paudyal P, Bremner S, Smith HE, Banerjee S et al. Predicting dementia from primary care records: A systematic review and meta-analysis. *PLOS One*. 2018.

36. Walters K, Hardoon S, Petersen I, Iliffe S, Omar RZ, Nazareth I et al. Predicting dementia risk in primary care: development and validation of the Dementia Risk Score using routinely collected data. *BMC Med*. 2016; 14(**6**).

37. Licher S, Leening MLG, Yilmaz P, Wolters FJ, Heeringa J, Bindels PJE et al. Development and Validation of a Dementia Risk Prediction Model in the General Population: An Analysis of Three Longitudinal Studies. *Am J* Psychiatry. 2018.

38. Licher S, Yilmaz P, Leening MLG, Wolters FJ, Vernooij MW, Stephan BCM et al. External validation of four dementia prediction models for use in the general community-dwelling population: a comparative analysis from the Rotterdam Study. *Neuroepidemiology*. 2018; (33) 645-655.

39. Eisele M, van den Bussche H, Koller D, *et. al.* Utilization Patterns of Ambulatory Medical Care before and after the Diagnosis of Dementia in Germany – Results of a Case-Control Study. *Dement Geriatr Cogn Disord* 2010; (**29**) 475–483.

40. Benjamins MR, Ellison CG, Krause NM, Marcum JP. Religion and preventive service use: do congregational support and religious beliefs explain the relationship between attendance and utilization? *J Behav Med.* 2011; **34**(6) 462-476.

41. Chen SH, Bu XL, Jin WS et. al. Altered peripheral profile of blood cells in Alzheimer disease: A hospital-based case-control study. *Medicine.* 2017; **96**(21).

42. Winchester LM, Powell J, Lovestone S, and Nevado-Holgado AJ. Red blood cell indices and anaemia as causative factors for cognitive function deficits and for Alzheimer's disease. *Genom Med.* 2018; **10**(51).

43. Wang HX, Wahlin A, Basun H, Fastbom J, Winblad B, and Fratiglioni L. Vitamin B12 and folate in relation to the development of Alzheimer's disease. *Neurology* 2001; **56**(9).

44. Cao W, and Zheng H. Peripheral immune system in aging and Alzheimer's disease. *Mol Neurodegener* 2018; **13**(51).

45. Le Page A, Dupuis G, Frost EH, *et. al.* Role of the peripheral innate immune system in the development of Alzheimer's disease. *Exp Gerontol.* 2018; (**107**) 59-66.

46. Fillit H, Nash DT, Rundek T, and Zuckerman A. Cardiovascular risk factors and dementia. *Am J Geriatr Pharmacother* 2008; **6**(2):100-18.

47. Marsden J and Pickering D. Urine testing for diabetic analysis. *Community Eye Health.* 2015; 28(92) 77.

48. Luchsinger JA, Tang MX, Stern Y, Shea S, and Mayeux R. Diabetes Mellitus and Risk of Alzheimer's Disease and Dementia with Stroke in a Multiethnic Cohort. *Am J Epidemiol.* 2001; **154**(7) 635-641.

49. Fritze T, Teipel S, Óvári A, Kilimann I, Witt G, and Doblhammer G. Hearing Impairment Affects Dementia Incidence. An Analysis Based on Longitudinal Health Claims Data in Germany. *PLOS One.* 2016.

FIGURE LEGENDS

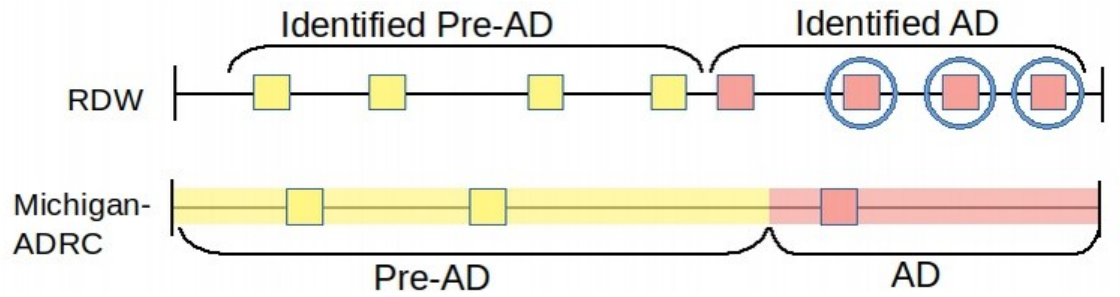


Figure 1: Comparing Michigan-ADRC and RDW encounters for a sample patient. Each row represents a timeline for the respective dataset, and encounters are indicated with squares. Shading along the Michigan-ADRC timeline indicates consensus-based diagnoses. A true positive is counted if at least one identified AD RDW encounter overlaps with the Michigan-ADRC defined AD window (e.g., the encounters in the blue circles).

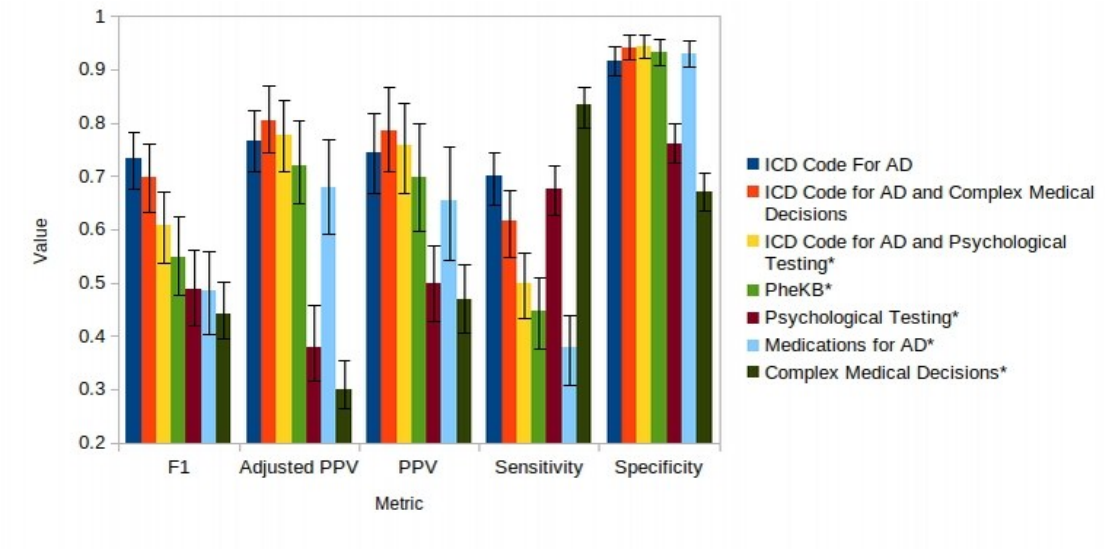


Figure 2: Cohort Discovery Results. Comparison of results from cohort discovery tools which tested a single EHR component, were previously published, or whose F1 score was greater than 0.5. Each color corresponds to the identification tool indicated in the figure legend. Complexity in medical decisions was measured by the amount and variety of patient data examined by a physician, patient risk, and treatment options. A '*' in the figure legend denotes criteria whose F1 score was significantly worse than the best cohort discovery tool.

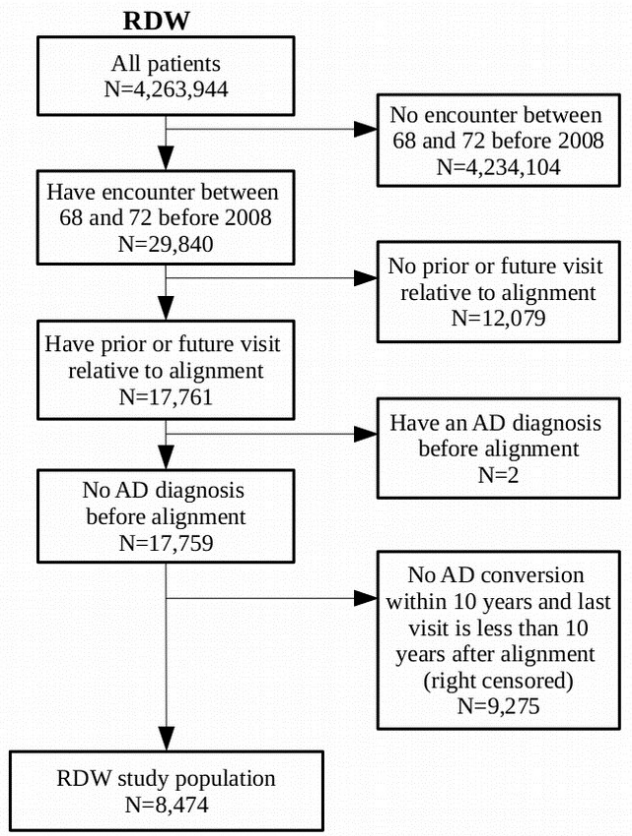


Figure 3: Applying Inclusion/Exclusion Criteria. We begin with all patients in Michigan Medicine's Research Data Warehouse (RDW). Numbers in each box correspond to the number of patients included/excluded.

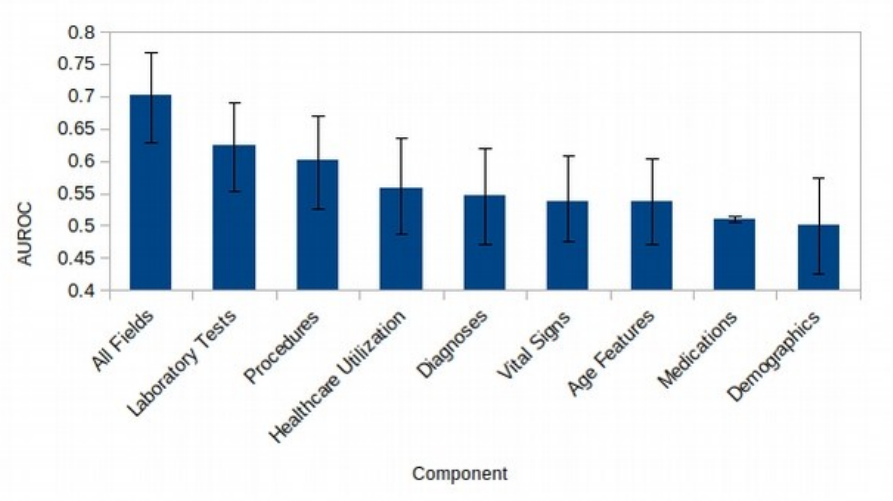
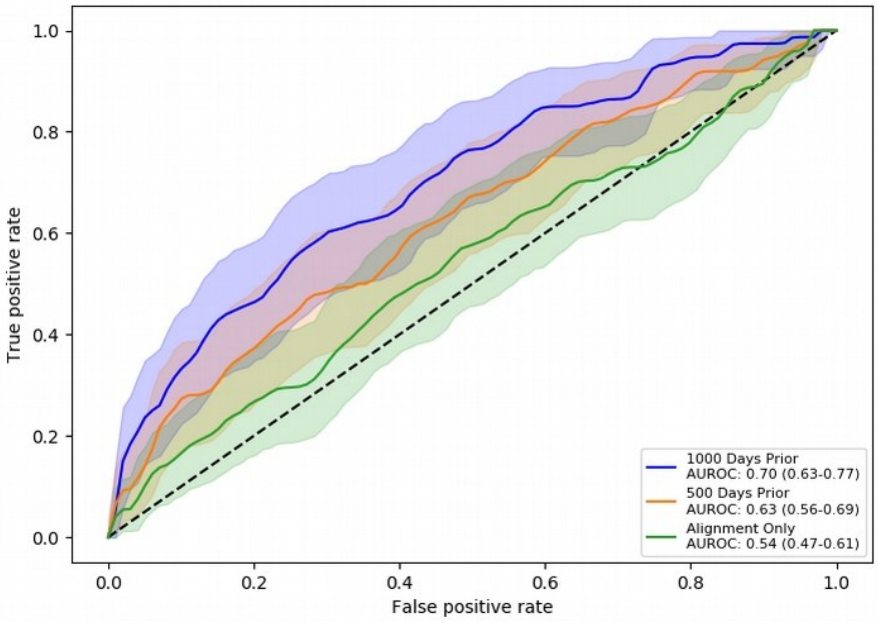
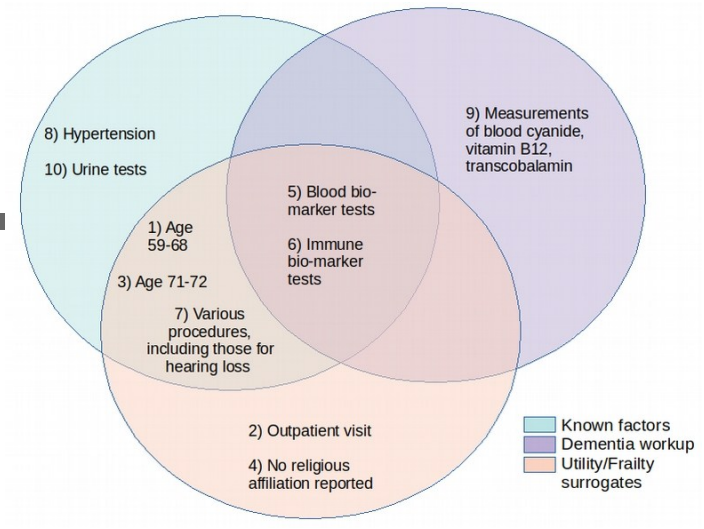


Figure 4: Comparison of EHR data contributions. (a) Analysis of Individual EHR Data Fields. Comparison of model performance when trained with specific fields of EHR data. In this experiment, all data prior to alignment were used. Error bars represent 95% confidence intervals.



(b) Analysis of Longitudinal Data. Comparison of model performance when trained on information from all encounters up to 1,000 prior to alignment versus training on information from up to 500 days before alignment and information from alignment only. In this experiment, data from all EHR components were used. Error bars represent 95% confidence intervals. The black dashed line represents the ROC curve for random predictions.



(c) Analysis of Individual Features. Broad categories in which the features from Table 2 can fall. Number correspond to those found in Table 2.

Table 1. Select characteristics of study cohort	
Patient Demographics	**RDW** **N=8,474**
Number of encounters per patient pre-alignment (IQR)	11 (4-25)
Number of encounters per patient post-alignment (IQR)	84 (36-172)
Female (%)	54.94
Clinical Characteristics	
Most common co-morbidity	Essential hypertension
Most common procedure	Laboratory tests related to hematology and coagulation
Most common medication	Morphine
AD conversion within 10 years (%)	4.14

Obtained from the inclusion/exclusion criteria in Figure 3. RDW: Michigan Medicine's Research Data Warehouse, IQR: interquartile range, AD: Alzheimer's disease

Table 2. Important features

Feature Group	Description	Drop in AUROC (95% CI)
1. Age between 59-68	• Maximum age between 59-68 • Age between 59-68	0.0400 (0.0251-0.0675)
2. Visit type – outpatient between 250-500 days before alignment	• Patient has an outpatient visit • Time between visits is in (0, 2] days	0.0180 (0.0060-0.0360)
3. Age between 71-72	• Maximum age between 71-72 • Age between 71-72	0.0070 (0.0015-0.0161)
4. Religion value NON	Patient does not report a religious association	0.0047 (0.0015-0.0128)
5. Lab test 32623-1 with value in (5.30, 7.4] 21000-5 with value in (11.099, 12.9] 4544-3 with value in (16.799, 36.8] 777-3 with value in (25.999, 190.0] 785-6 with value in (15.699, 29.5] 786-4 with value in (29.799, 33.7] 787-2 with value in (52.499, 86.3] 789-8 with value in (2.149, 4.09] between 750-1000 days of alignment	Blood measurements of • platelet mean volume • erythrocyte distribution • hematocrit • erythrocyte mean corpuscular hemoglobin	0.0041 (0.0026-0.0074)
6. Laboratory test 736-9 with value in (0.399, 16.6] 5905-5 with value in (0.099, 6.1]	Blood measurements of • lymphocytes • monocytes	0.0037 (0.0005-0.0093)

704-7 with value in (0.000, 0.7] 731-0 with value in (0.099, 1.1] 742-7 with value in (0.000, 0.4] 751-8 with value in (0.099, 3.0] between 500-750 days of alignment	• basophils • neutrophils	
7. Diagnosis code V04.8 along with procedures 9065x and G000x between 250-500 days before alignment	• Vaccines for influenza, pneumococcal disease • Revision mastoidectomy • Injection of samarium lexidrona	0.0028 (0.0006-00073)
8. Non-invasive systolic blood pressure in (127, 136] between 500-750 days before alignment	Elevated blood pressure/hypertension	0.0023 (0.0004-0.0041)
9. Procedure 8260x and lab test 2132-9 with value in (89.999, 382.8] between 0-250 days before alignment	Measurements of • blood cyanide • vitamin B12 • transcobalamin	0.0021 (0.0012-0.0031)
10. Laboratory test 50557-8 with value negative 27297-1 with value negative 50561-0 with value negative 50563-6 with value < 1mg/dl 53327-3 with value negative 53328-1 with value negative 57747-8 with value negative between 250-500 days of alignment	Urine measurements of • ketones • leukocyte esterase • protein • urobilinogen • total bilirubin • glucose • erythrocytes	0.0021 (0.0009-0.0044)

Summary of the top 10 most important feature groups, as determined by permutation importance. The letter "x" is used to denote any character. Laboratory tests, diagnoses, and procedures are represented as LOINC, ICD9, and CPT codes respectively. CI: confidence interval