EHR Data for Alzheimer's Disease Research:

Opportunities and Challenges for Statistical Analyses

Sujuan Gao, Ph.D.

Indiana Alzheimer's Disease Research Center

Opportunities offered by EHR data

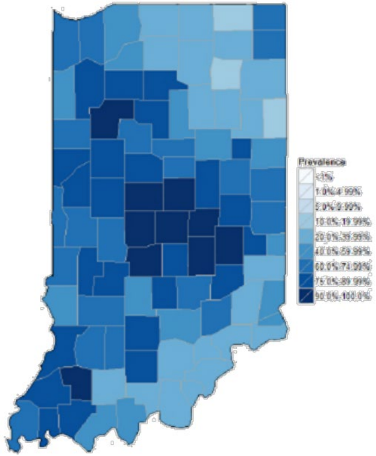
- Large sample size and cost efficient
- Longitudinal data
- Diverse data types
  - History of medical conditions (beyond those collected in UDS)
  - Laboratory results
  - Medication prescription: dose, frequency, and duration
  - Health care utilization (outpatient, ED, hospitalization)
- Population health studies
- Reduced recall bias

Challenges in EHR Diagnosis of AD

- Evaluation:
  - No systematic evaluation
  - Lack of detailed cognitive testing
- Diagnosis
  - Lack of specificity for AD
  - Potential for misdiagnosis
- Time of diagnosis
  - Delay in diagnosis
  - Impact on timely intervention
- These limitations can be overcome by linking with ADRC or AD-focused research data

EHR system at Indiana University

- The first EMR was developed in 1972 by the Regenstrief Institute at IU
  - Regenstrief Medical Record System, a physician-designed, integrated patient information system, was developed and first used in the Wishard Diabetes Clinic.
- Indiana Network for Patient Care (INPC)
  - A state-wide EHR data repository
    - 100+ hospitals, representing 38 health systems
    - 12,000+ practices with over 30,000 providers
    - 12 million+ patients
    - 9 billion clinical data elements



Integrating EHR data at the Indiana ADRC

- Merging EHR data with research data collected from the Indianapolis-Ibadan Dementia Project
    - Changes in glucose level and dementia diagnosis
    - Replication in a large EHR cohort
    - Antihypertensive medications and dementia risk

Alzheimers Dement. 2017 February ; 13(2): 111–118. doi:10.1016/j.jalz.2016.08.017.

Glucose level decline precedes dementia in elderly African Americans with diabetes

Hugh C. Hendrie, MB, ChB, DSc[a,b,c,*], Mengjie Zheng, MS[d], Wei Li, MD, PhD[e], Kathleen Lane, MS[d], Roberta Ambuehl, MS[b], Christianna Purnell, BA[a], Frederick W. Unverzagt, PhD[c], Alexia Torke, MD, MS[a,b,f], Ashok Balasubramanyam, MD[g], Chris M. Callahan, MD[a,b,f], and Sujuan Gao, PhD[d]

Alzheimers Dement. 2018 December ; 14(12): 1572–1579. doi:10.1016/j.jalz.2018.03.008.

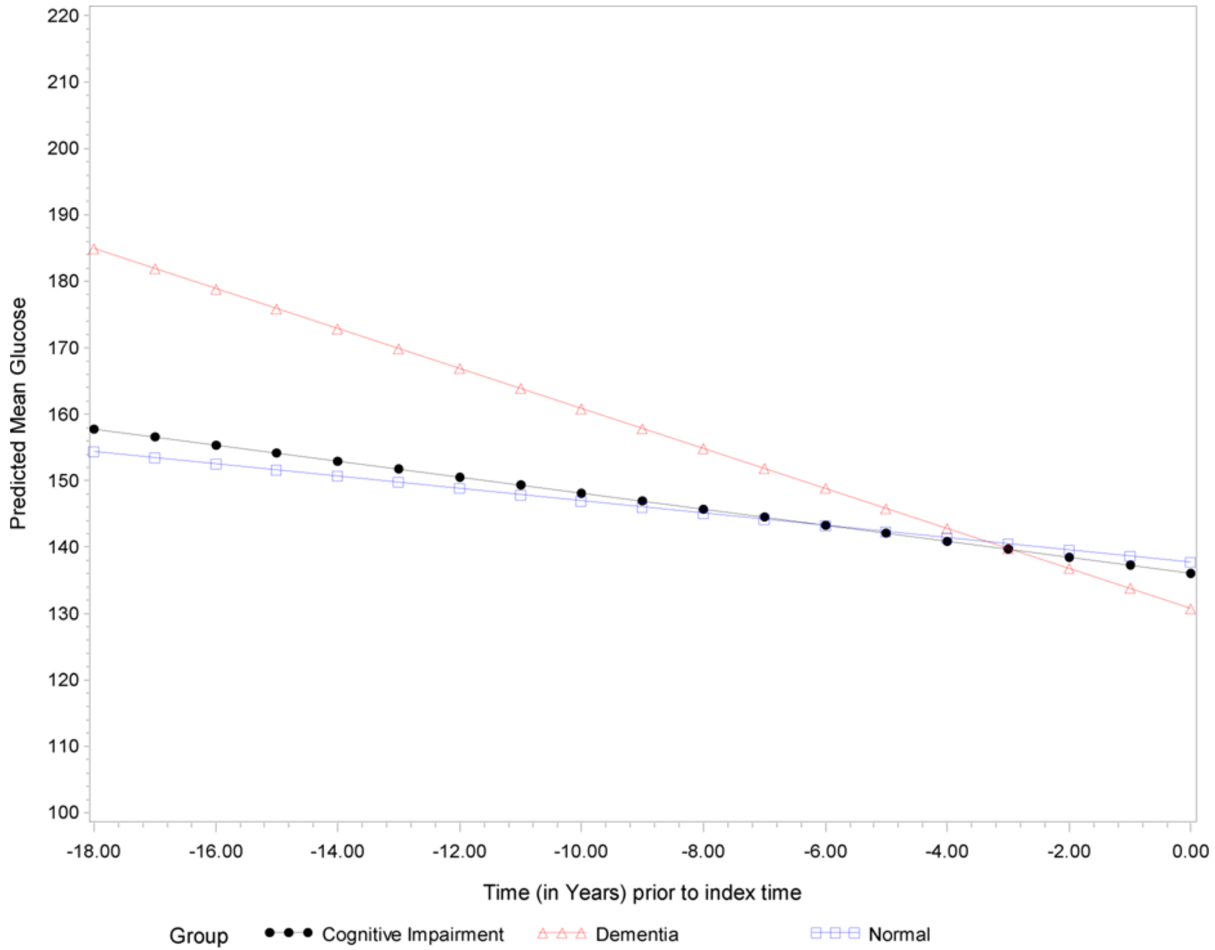
Changes of glucose levels precede dementia in African Americans with diabetes but not in Caucasians

Hugh C. Hendrie, DSc[a,b,c,*], Mengjie Zheng, MS[d], Kathleen A. Lane, MS[d], Roberta Ambuehl, MS[b], Christianna Purnell, BA[a], Shanshan Li, PhD[d], Frederick W. Unverzagt, PhD[c], Michael D. Murray, PharmD[b,e], Ashok Balasubramanyam, MD[f], Chris M. Callahan, MD[a,b,g], and Sujuan Gao, PhD[d]

J Gen Intern Med. 2018 Apr; 33(4): 455–462. PMCID: PMC5880772
Published online 2018 Jan 12. doi: 10.1007/s11606-017-4281-x PMID: 29330643

Antihypertensive Medication and Dementia Risk in Older Adult African Americans with Hypertension: A Prospective Cohort Study

Michael D. Murray, PharmD, MPH,[1,2] Hugh C. Hendrie, DSc,[1,3,4] Kathleen A. Lane, MS,[5] Mengjie Zheng, MS,[5] Roberta Ambuehl, MS,[1] Shanshan Li, PhD,[5] Frederick W. Unverzagt, PhD,[4] Christopher M. Callahan, MD,[1,3,6] and Sujuan Gao, PhD[5]


220
210
200
190
180
170
160
150
140
130
120
110
100
Predicted Mean Glucose
-18.00
-16.00
-14.00
-12.00
-10.00
-8.00
-6.00
-4.00
-2.00
0.00
Time (in Years) prior to index time
Group
Cognitive Impairment
Dementia
Normal

Integrating EHR data at the Indiana ADRC

- Utilizing EHR data for IADRC participants
  - Medical history, labs and medications for consensus diagnoses (web-based data access)
  - Data analyses for various research projects
- Risk factors for dementia in ICU survivors using EHR data
- Pilot site of NACC’s COVID supplement

Statistical Challenges in Analyzing EHR Data

- Cohort and variable definition
  - Understanding EHR data structure
- Accurate patient identification
  - Collecting Medical Record Numbers (MRNs)
  - Handling data from multiple healthcare systems
  - Patient matching in cases without MRNs
- Data quality
  - Need for extensive data QC
- Completeness
  - Establishing an observation window for each individual
  - Establish censoring points for non-events

Statistical Challenges in Analyzing EHR Data

- Varying time intervals for longitudinal data
  - Proper alignment of time for longitudinal models
  - Assessing the robustness of findings to the number of measurements
- Selection bias
  - Healthy individuals may have fewer EHR encounters
- Confounding bias
  - Conducting a careful examination of all model variables
  - Reducing reliance on automated model selection procedure
- Causal inference
  - Inverse-probability weighting: Modeling the probability of exposure/treatment
  - Standardization: Modeling conditional mean outcomes based on confounders

OPEN

A comparison of machine learni methods for survival analysis of high-dimensional clinical data for dementia prediction

Annette Spooner[1], Emily Chen[1], Arcot Sowmya[1], Perminder Sachdev[2,3], Nicole A. Julian Trollor[2,3,4] & Henry Brodaty[2,3]

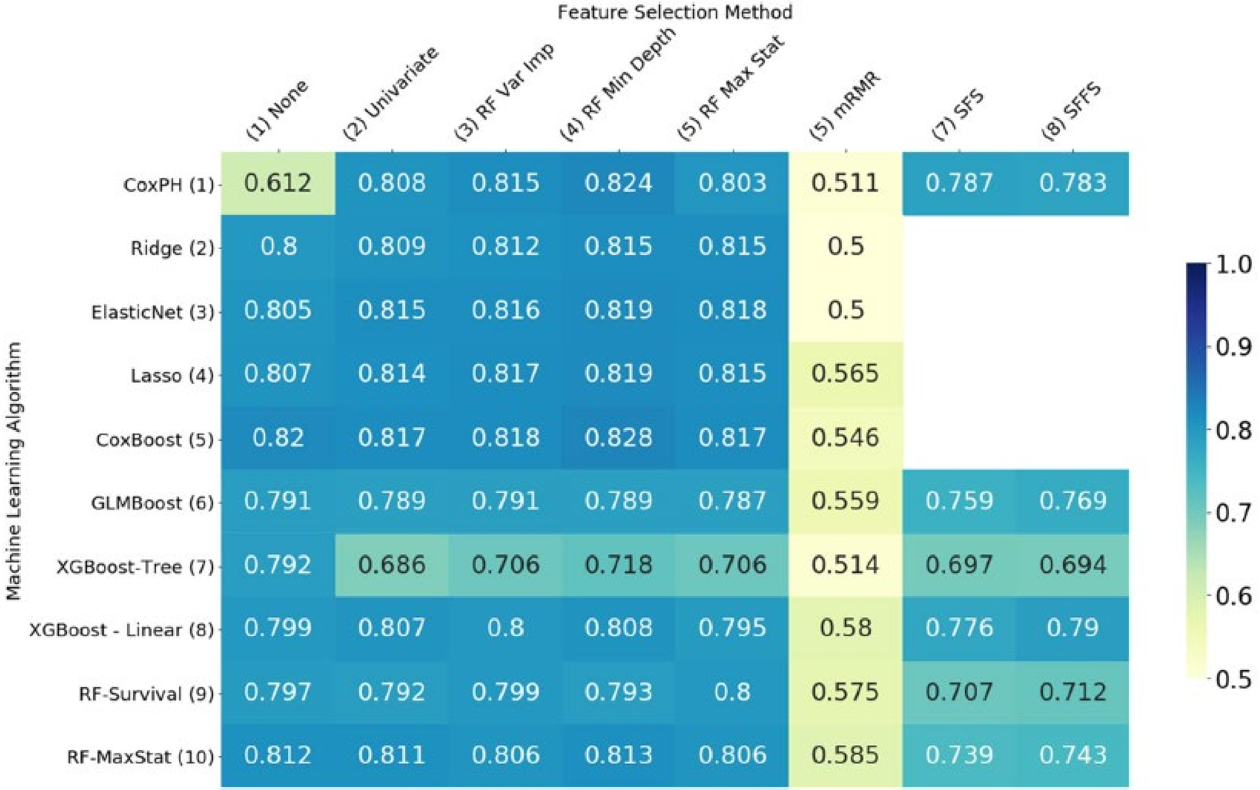


Figure 1. Heatmap illustrating the performance of each of the machine learning algorithms with each feature selection method on the MAS dataset, measured by the mean concordance index.

JAMA Cardiology | Original Investigation

Use of Machine Learning Models to Predict Death After Acute Myocardial Infarction

Rohan Khera, MD, MS; Julian Haimovich, MD; Nathan C. Hurley, BS; Robert McNamara, MD;
John A. Spertus, MD, MPH; Nihar Desai, MD, MPH; John S. Rumsfeld, MD, PhD; Frederick A. Masoudi, MD, MSPH;
Chenxi Huang, PhD; Sharon-Lise Normand, PhD; Bobak J. Mortazavi, PhD; Harlan M. Krumholz, MD, SM

Table 2. Performance Characteristics of Models for Predicting In-Hospital Mortality in Acute Myocardial Infarction

Characteristic	Logistic regression	LASSO	Neural network	XGBoost	Meta-classifier
Variables included in the model of McNamara et al[21]					
Model performance metrics					
AUROC (95% CI)	0.878 (0.875-0.881)	0.874 (0.870-0.879)	0.874 (0.870-0.878)	0.886 (0.882-0.890)	0.886 (0.882-0.890)
Precision-recall AUC	0.372	0.367	0.371	0.395	0.398
F score	0.415	0.408	0.411	0.432	0.432
Sensitivity	0.42 (0.41-0.43)	0.43 (0.42-0.45)	0.41 (0.40-0.42)	0.44 (0.43-0.45)	0.43 (0.42-0.44)
Specificity	0.97 (0.97-0.97)	0.97 (0.97-0.97)	0.97 (0.97-0.97)	0.97 (0.97-0.97)	0.98 (0.97-0.98)
PPV	0.41 (0.40-0.42)	0.38 (0.37-0.39)	0.41 (0.40-0.42)	0.42 (0.41-0.43)	0.44 (0.43-0.45)
NPV	0.97 (0.97-0.97)	0.97 (0.97-0.98)	0.97 (0.97-0.97)	0.98 (0.97-0.98)	0.97 (0.97-0.98)

Summary

1. EHR data offers valuable, longitudinal medical information that can be challenging to obtain in research projects limited by funding and time.
2. The integration of EHR data and ADRD focused research data presents a unique opportunity to explore mid-life medical conditions many years before the onset of AD.
3. Critical considerations in EHR data analysis:
   - Ensuring data accuracy
   - Utilizing appropriate analytic methods for missing data and time-varying exposures
   - Employing advanced statistical methods to mitigate potential biases
   - Performing sensitivity analyses to ensure robustness of findings