

# Leverage Real-world Longitudinal Data in Large Clinical Research Networks for Alzheimer's Disease and Related Dementia (ADRD)

Rui Duan, M.S.<sup>1\*</sup>, Zhaoyi Chen, Ph.D.<sup>2\*</sup>, Jiayi Tong, B.S.<sup>3\*</sup>, Chongliang Luo, Ph.D.<sup>3</sup>,  
Tianchen Lyu, M.S.<sup>4</sup>, Cui Tao, Ph.D.<sup>5</sup>, Demetrius Maraganore, M.D.<sup>6</sup>, Jiang Bian, Ph.D.<sup>4+</sup>,  
Yong Chen, Ph.D.<sup>3+</sup>

<sup>1</sup>Department of Biostatistics, Harvard T.H. Chan School of Public Health, Boston, MA, USA

<sup>2</sup>Department of Epidemiology, College of Medicine & College of Public Health and Health Professions, University of Florida, Gainesville, FL, USA

<sup>3</sup>Department of Biostatistics, Epidemiology and Informatics, Perelman School of Medicine, The University of Pennsylvania, Philadelphia, PA, USA

<sup>4</sup>Department of Health Outcomes and Biomedical Informatics, College of Medicine, University of Florida, Gainesville, FL, USA

<sup>5</sup>School of Biomedical Informatics, The University of Texas Health Science Center at Houston, Houston, TX, USA

<sup>6</sup>Department of Neurology, College of Medicine, University of Florida, Gainesville, FL, USA

## Abstract

*With vast amounts of patients' medical information, electronic health records (EHRs) are becoming one of the most important data sources in biomedical and health care research. Effectively integrating data from multiple clinical sites can help provide more generalized real-world evidence that is clinically meaningful. To analyze the clinical data from multiple sites, distributed algorithms are developed to protect patient privacy without sharing individual-level medical information. In this paper, we applied the One-shot Distributed Algorithm for Cox proportional hazard model (ODAC) to the longitudinal data from the OneFlorida Clinical Research Consortium to demonstrate the feasibility of implementing the distributed algorithms in large research networks. We studied the associations between the clinical risk factors and Alzheimer's disease and related dementia (ADRD) onsets to advance clinical research on our understanding of the complex risk factors of ADRD and ultimately improve the care of ADRD patients.*

## Introduction

Over the last few decades, there has been increased adoption of electronic health records (EHR) systems in the United States (US). Containing rich medical information, EHR data can potentially improve the efficiency and effectiveness of health care and biomedical research<sup>1,2</sup>. EHR data are considered as real-world data (RWD) that are collected outside of traditional clinical research settings (e.g., randomized controlled trials). The U.S. Food and Drug Administration (FDA) recently launched the real-world evidence (RWE) program to encourage the use of RWD such as EHRs, administrative claims, and billing data among others, to support the development, evaluation, and monitoring of drug products<sup>3</sup>, where methods that can reliably measure disease progression and the impact of the drugs are urgently needed. The detailed temporal information contained in EHR, including conditions and diagnoses, procedures, medications, laboratory test results, disease status, and treatment outcomes, create opportunities for analyzing time-to-event outcomes, which can provide a better understanding of the disease progression and impact of the treatments or risk factors on the timing of outcomes. Among many methods developed for analyzing time-to-event data<sup>4</sup>, the Cox proportional hazards model<sup>3</sup> is one of the commonly used methods and has been widely applied in biomedical research.

\*: first three authors contributed equally

+: correspondence

To better utilize EHR data collected at different clinical sites, increasing numbers of data consortia were founded over the last few decades<sup>5-7</sup>. The national Patient-Centered Clinical Research Network (PCORnet), which is funded by Patient-Centered Outcomes Research Institute (PCORI), covers more than 100 million patients all over the United States<sup>6</sup>. As one of the nation's 9 clinical data research networks (CDRNs), the OneFlorida Clinical Research Consortium (OneFlorida CRC) was designed to accelerate the translation of promising research findings into improved patient care, with a focus on comparative effectiveness research, programmatic clinical trials, and patient-centered outcomes studies. The centerpiece of the OneFlorida network is its Data Trust—a repository of statewide health care data containing robust longitudinal and linked patient-level RWD of around 15 million (>50%) Floridians, including data from Medicaid claims, cancer registries, vital statistics, and EHRs from its clinical partners<sup>8</sup>. Currently, there are 12 various types of healthcare organizations contributing to the OneFlorida data repository: 1) four academic health centers (i.e., University of Florida Health, University of Miami Health System, Florida State University and regional campus practice partners, and University of South Florida Health), 2) seven healthcare systems including Tallahassee Memorial Healthcare (affiliated with Florida State University), Orlando Health, Adventist Health, Nicklaus Children's Hospital, Bond Community Health, Capital Health Plan, and Tampa General Hospital, and 3) CommunityHealth IT—a rural health network in Florida. Covering all 67 Florida counties<sup>9</sup>, the OneFlorida CRC provides care for more than 50% of Floridians through 4,100 physicians, 914 clinical practices, and 22 hospitals, leading to the increasing amount of longitudinal and robust patient-level records of around 15 million Floridians and over 561.1 million encounters, 1.16 billion diagnoses, 1 billion prescribing records, and 1.44 billion procedures as of December 2019.

The large-scale clinical data networks like the OneFlorida CRC allow researchers to study the diverse range of risk factors, from clinical characteristics to social determinants of health, associated with the fatal degenerative diseases: Alzheimer's disease (AD) and AD-related dementia (ADRD). Alzheimer's disease (AD) is the most common cause of dementia. In 2019, 5.8 million Americans will live with AD, among which 97% are aged  $\geq 65$  and 81% are  $\geq 75$ . By 2050, people living with AD in the US may grow to 13.8 million, fueled by the aging baby boomers<sup>10</sup>. In 2017, 121,404 deaths from AD were recorded, making AD the 6th leading cause of death and the 5th leading cause of death among Americans aged  $\geq 65$ <sup>11</sup>. The progression of AD/ADRD usually starts with normal cognition during the preclinical period, slowly advancing to mild cognitive impairment (MCI) and then gradually progresses to mild and moderate AD and eventually severe AD. It usually takes an individual with MCI 7 years to progress to mild AD, but some individuals may experience a rapid progression which took significantly less time to develop into AD<sup>12</sup>. Therefore, modeling the disease progression in an at-risk population is important.

Combining RWD from multiple clinical sites can provide a larger sample size, which can lead to more generalizable findings, and the ability to better evaluate rare risk factors of ADRD. Nevertheless, directly sharing individual patient-level data can be challenging because of the privacy concerns over protected patient health information<sup>14</sup>. The state-of-the-art method for multicenter study, without sharing patient-level information, is the meta-analysis, where each site fits separate analysis and all local estimates are synthesized using a weighted average<sup>15</sup>. In addition to meta-analysis, several distributed learning algorithms, in which only aggregate information is allowed to be shared across institutions, have been developed to overcome the privacy issue in a clinical research network<sup>16-19</sup>. Among the existing methods, Duan et al<sup>20</sup> proposed a privacy-preserving and communication-efficient distributed algorithm, named ODAC, to fit the multi-center Cox proportional hazards model. Utilizing a surrogate likelihood approach<sup>21</sup> and without iterative communication across the sites, the ODAC algorithm allows efficient and accurate identification of the risk factors associated with the time-to-event outcome of interest<sup>20</sup>. This algorithm is shown to achieve high accuracy in the sense that the results of ODAC is close to the pooled analysis in which a Cox model is fitted on the combined dataset.

In this paper, with the multicenter RWD data from OneFlorida, we evaluated the ODAC algorithm using ADRD as a use case. By studying the associations between ADRD and several clinical risk factors, 7 significant risk factors were identified by ODAC, and the 7 risk factors were consistent with previous findings on ADRD. In addition, compared with the commonly used meta-analysis, ODAC provided estimates of the effect sizes with smaller bias, leading to the conclusion that ODAC is a reliable and efficient, privacy-preserving distributed learning algorithm extremely suitable for working with real-world data from multicenter clinical research networks.

## Materials and Methods

### *Data Source and Study Population*

The OneFlorida data is a HIPAA (i.e., Health Insurance Portability and Accountability Act) limited data set (i.e., dates are not shifted; and 9-digit zip codes of patients' residencies are available, where other 16 types of patient identifiers

were removed) that contains detailed patient and clinical variables, including demographics, encounters, diagnoses, procedures, vitals, medications, and labs, following the PCORnet Common Data Model (CDM)<sup>6</sup>. The OneFlorida data undergo rigorous quality checks at its data coordinating center (i.e., University of Florida [UF]), and a privacy-preserving record linkage process is used to deduplicate records of same patients coming from different health care systems within the network<sup>22</sup>.

**Table 1.** Breakdown of the study populations by site.

Site	ADRD	At-risk population	prevalence
Site 1	128	3,292	5.34%
Site 2	10	983	1.02%
Site 3	711	12,181	5.84%
<b>All sources</b>	<b>849</b>	<b>16,456</b>	<b>5.16%</b>

Based on the OneFlorida data, individuals who were 65 years of age, and had no ADRD diagnosis before 2014/03/01 (i.e., the index date) were considered as “at-risk population” and were included in our analysis. We chose 2014/03/01 as the baseline to ensure a 5-year follow-up period for each individual. The outcome of interest is the time to the first diagnosis of ADRD. Conditions that were considered as ADRD include mild cognitive impairment (ICD-9: 331.81, 294.9; ICD-10: G31.83, F09), Alzheimer’s disease (ICD-9: 331.0; ICD-10: G30.0, G30.1, G30.8, G30.9), vascular dementia (ICD-9: 290.40, 290.41; ICD-10: F01.50, F01.51), Lewy body dementia (ICD-9: 331.82; ICD-10: G31.83), Frontotemporal Dementia (ICD-9: 331.19; ICD-10: G31.09), and primary progressive aphasia (ICD-9: 331.11; ICD-10: G31.01). A total of 178,251 patients who were at risk at baseline were identified in OneFlorida. After excluding patients with missing data, a total of 16,456 individuals were included in our final analysis. **Table 1** shows the prevalence of the included population across the different OneFlorida clinical sites.

### **Risk Factors**

We identified a set of risk factors from the literature and extracted the factors from patients’ medical records in OneFlorida. All records before the index date (2014/03/01) were taken into consideration in this analysis. Factors such as demographic variables (age, race, gender, and insurance type), vital signs (body mass index [BMI], lipid panel), smoking status, selected clinical diagnoses, and medications were included. Since the laboratory test results (e.g., complete blood count) had high frequencies of missing values (>50% in the total study population), these factors were removed. Besides, the clinical diagnoses that were made in <1% of the total study population were removed to minimize potential bias introduced by the small sample size. Patients who had missing values in any risk factors were removed as the current ODAC algorithm is unable to handle missing values. A total of 12 risk factors/predictors were included in the analysis. In **Table 2**, we present the summary statistics of these predictors. From the table, we observed that the patients who were diagnosed with ADRD were older, more males, and tended to have Medicare insurance. In addition, ADRD patients were more likely to have more comorbidities and higher statin use.

### **Statistical Analysis**

To help introducing the ODAC algorithm, we first define some basic notation. Suppose we have  $K$  sites in the clinical network, where the  $j$ -th site has  $n_j$  samples, with  $N = \sum_{j=1}^K n_j$ . For the  $i$ -th subject in the  $j$ -th site, we observe  $\{T_{ij}, \delta_{ij}, x_{ij}\}$ , where  $x_{ij}$  is a vector denoting  $p$  risk factors,  $T_{ij}$  is the time-to-event for the outcome of interest, and  $\delta_{ij}$  is the event indicator with  $\delta_{ij} = 1$  indicating an event, and  $\delta_{ij} = 0$  indicating an censored. We apply the distributed algorithm ODAC developed in Duan et al.<sup>20</sup>, and the detailed algorithm is shown below with the definition of each quantity in **Table 3**. The main idea of the algorithm is to fit Cox model at each site first and combine the results through a fixed-effect meta-analysis to obtain initial values for regression parameters, and then apply the surrogate likelihood approach proposed by Jordan et al.<sup>21</sup> which requires each site to calculate summary-level quantities for calculating the first and second-order derivatives of the combined log-partial likelihood function. A surrogate likelihood can serve as a good proxy of the combined log-partial likelihood function, and the final estimator is obtained by maximizing the surrogate function. **Figure 1** provides a schematic illustration of the above algorithm,

and we refer to Duan et al.<sup>20</sup> for more technical details and properties of the ODAC method. In addition to ODAC, we also applied the meta-analysis and the pooled analysis to the multicenter EHR data we extracted from OneFlorida.

**Table 2.** Characteristics of included risk factors in the cohort.

	<b>Control (n=15607)</b>	<b>Case (n=849)</b>	<b>Overall (n=16456)</b>
<b>Baseline_AGE</b>			
Mean (SD)	73.9 (5.94)	76.5 (6.54)	74.0 (6.00)
<b>Gender</b>			
Female	8981 (57.5%)	474 (55.8%)	9455 (57.5%)
Male	6626 (42.5%)	375 (44.2%)	7001 (42.5%)
<b>RACE_ETHNICITY</b>			
Hispanic	277 (1.8%)	20 (2.4%)	297 (1.8%)
NHB*	2324 (14.9%)	141 (16.6%)	2465 (15.0%)
NHW*	12588 (80.7%)	667 (78.6%)	13255 (80.5%)
Other	418 (2.7%)	21 (2.5%)	439 (2.7%)
<b>insurance_type</b>			
MEDICAID	140 (0.9%)	4 (0.5%)	144 (0.9%)
MEDICARE	11568 (74.1%)	694 (81.7%)	12262 (74.5%)
No payment*	48 (0.3%)	2 (0.2%)	50 (0.3%)
Other	470 (3.0%)	3 (0.4%)	473 (2.9%)
OtherGOV	161 (1.0%)	19 (2.2%)	180 (1.1%)
PRIVATE	3220 (20.6%)	127 (15.0%)	3347 (20.3%)
<b>BMI</b>			
Mean (SD)	28.9 (6.18)	27.9 (6.75)	38.5 (889)
<b>SMOKING</b>			
Current Smoker	1000 (6.4%)	46 (5.4%)	1046 (6.4%)
Former smoker	6301 (40.4%)	337 (39.7%)	6638 (40.3%)
Never smoker	8306 (53.2%)	466 (54.9%)	8772 (53.3%)
<b>depression</b>	1837 (11.8%)	175 (20.6%)	2012 (12.2%)
<b>anxiety</b>	1859 (11.9%)	145 (17.1%)	2004 (12.2%)
<b>sleep_disorder</b>	1578 (10.1%)	123 (14.5%)	1701 (10.3%)
<b>hypertension</b>	10208 (65.4%)	600 (70.7%)	10808 (65.7%)
<b>diabetes</b>	4081 (26.1%)	254 (29.9%)	4335 (26.3%)
<b>heart_disease</b>	1701 (10.9%)	136 (16.0%)	1837 (11.2%)
<b>statin</b>	4646 (29.8%)	333 (39.2%)	4979 (30.3%)

\*NHB stands for non-Hispanic black; NHW stands for non-Hispanic White; *No payment* includes self-pay, nor charge, refusal to pay/bad debt, hill burton free care, research/donor, and other.

### Algorithm ODAC

#### (1) Initialization

  In Site  $k = 1$  to  $K$ ,

**do**

    Fit a Cox model and obtain the local estimate  $\hat{\beta}_k$  and the variance estimate  $\hat{V}_k$ ;

*broadcast*  $\hat{\beta}_k, \hat{V}_k$ , and the set of unique event time points in site  $k$ .  
**end**  
**(2) Local surrogate estimator**  
 In Site  $k = 1$  to  $K$ ,  
**do**  
 obtain  $\tilde{\beta}$  and all the unique event time points across all sites  $\mathcal{T} = \{t_1 \dots t_d\}$ ;  
 calculate and broadcast the intermediate summary-level statistics  $U_j(t_s), W_j(t_s)$  and  $Z_j(t_s)$  for each  $t_s \in \mathcal{T}$ ;  
 construct the surrogate likelihood  $\tilde{L}_k(\beta)$  treating the  $k$ -th site as the local site;  
 obtain and broadcast  $\tilde{\beta}_k$  and the variance  $\tilde{V}_k$ ;  
**end**  
**(3) Evidence synthesis**  
 Obtain ODAC estimator  $\tilde{\beta}$ .  
**Return**  $\tilde{\beta}$ .

**Table 3.** Descriptions and definitions of quantities for ODAC.

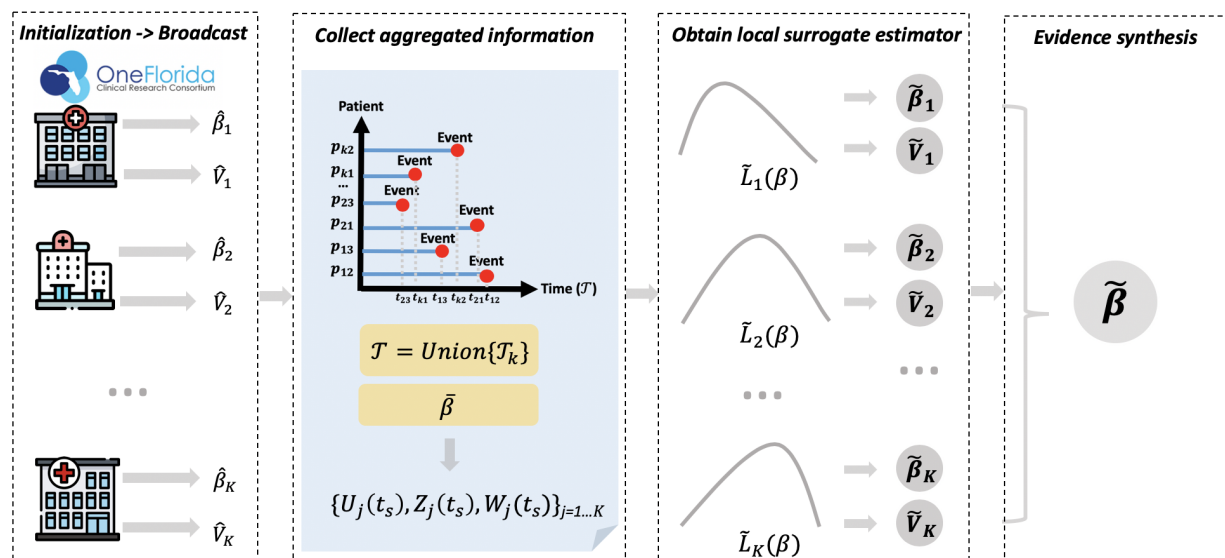
Quantity	Description	Definition
$R_j(t)$	Risk set at time $t$ in the $j$ -th site	$R_j(t) = \{i; T_{ij} \geq t\}$
$L_j(\beta)$	Log partial likelihood function at $j$ -th site	$\frac{1}{n_j} \sum_{i=1}^{n_j} \delta_{ij} \log \frac{\exp(\beta^T x_{ij})}{\sum_{s \in R_j(T_{ij})} \exp(\beta^T x_{sj})}$
$L(\beta)$	Combined log partial likelihood function	$L(\beta) = \sum_{j=1}^K \frac{n_j}{N} L_j(\beta)$
$\hat{\beta}_j$	Local estimate at $j$ -th site	$\operatorname{argmax}_{\beta} L_j(\beta)$
$\hat{V}_j$	Variance of $\hat{\beta}_j$	$-\{n_j \nabla^2 L_j(\hat{\beta}_j)\}^{-1}$
$\tilde{\beta}$	Initial value for $\beta$	$(\sum_{j=1}^K \hat{V}_j^{-1})^{-1} \sum_{j=1}^K \hat{V}_j^{-1} \hat{\beta}_j$
$\mathcal{T}$	Collection of all time points	$\{T_{ij}; \delta_{ij} = 1\}$
$U_j(t)$	Intermediate results	$\sum_{i \in R_k(t)} \exp(\beta^T x_{ik})$
$W_j(t_s)$	Intermediate results	$\sum_{i \in R_k(t)} \exp(\beta^T x_{ik}) x_{ik}$
$Z_j(t_s)$	Intermediate results	$\sum_{i \in R_k(t)} \exp(\beta^T x_{ik}) x_{ik} x_{ik}^T$
$\tilde{L}_j(\beta)$	Surrogate likelihood function at $j$ -th site	$L_j(\beta) + \langle \nabla L(\tilde{\beta}) - \nabla L_j(\tilde{\beta}), \beta \rangle + \frac{1}{2} (\beta - \tilde{\beta})^T \{\nabla^2 L(\tilde{\beta}) - \nabla^2 L_j(\tilde{\beta})\} (\beta - \tilde{\beta})$
$\tilde{\beta}_k$	Surrogate estimator in $j$ -th site	$\operatorname{argmax}_{\beta} \tilde{L}_j(\beta)$
$\tilde{V}_k$	Estimated variance of $\tilde{\beta}_k$	$\tilde{V}_j = \frac{1}{N} \{-\nabla^2 L_j(\tilde{\beta}_j)\}^{-1}$
$\tilde{\beta}$	ODAC estimator	$(\sum_{j=1}^K \tilde{V}_j^{-1})^{-1} \sum_{j=1}^K \tilde{V}_j^{-1} \tilde{\beta}_j$

## Results

**Figure 2** shows the estimated log hazard ratio and 95% confidence interval for each risk factor using the three different approaches: the pooled analysis, meta-analysis, and ODAC. The pooled analysis is considered as the gold standard, since it essentially requires sharing all the patient-level data across sites. From the plot, we observe that ODAC provides more accurate estimates for the associations between outcome and exposures than meta-analysis, in the sense that the bias to the gold standard pooled estimator is smaller. To be more specific, ODAC estimates are closer to the

pooled estimates for 11 out of the 13 risk factors we considered in the Cox regression model. The relative bias for the rest of two risk factors are below 6%. The bias for meta-analysis is up to 185%. For risk factors such as BMI and statin, the conclusion regarding the significance of the risk factor from meta-analysis is not consistent with the pooled analysis, while ODAC provides the same conclusion as the pooled analysis. The confidence intervals of ODAC is observed to be slightly larger than the other two methods, which might be caused by the heterogeneity of the data across sites.

**Figure 1.** Illustration of the ODAC algorithm. The first step is initialization, in which each site reports the local estimate of the log hazard ratio ( $\hat{\beta}_j$ ) and the variance ( $\hat{V}_j$ ) where  $j = 1, \dots, K$ . Then with these initial values broadcasted across the sites, the average of all local estimates and the union of all event times can be obtained. With these results, each site calculates and shares the intermediate results  $\{U_j(t_s), Z_j(t_s), W_j(t_s)\}$ . Next, within each site, the intermediate results and individual patient-level data are used to construct the local surrogate likelihood function. By optimizing the surrogate likelihood function, we obtain the local surrogate estimates. The final step is to synthesize the surrogate estimates from all the sites.



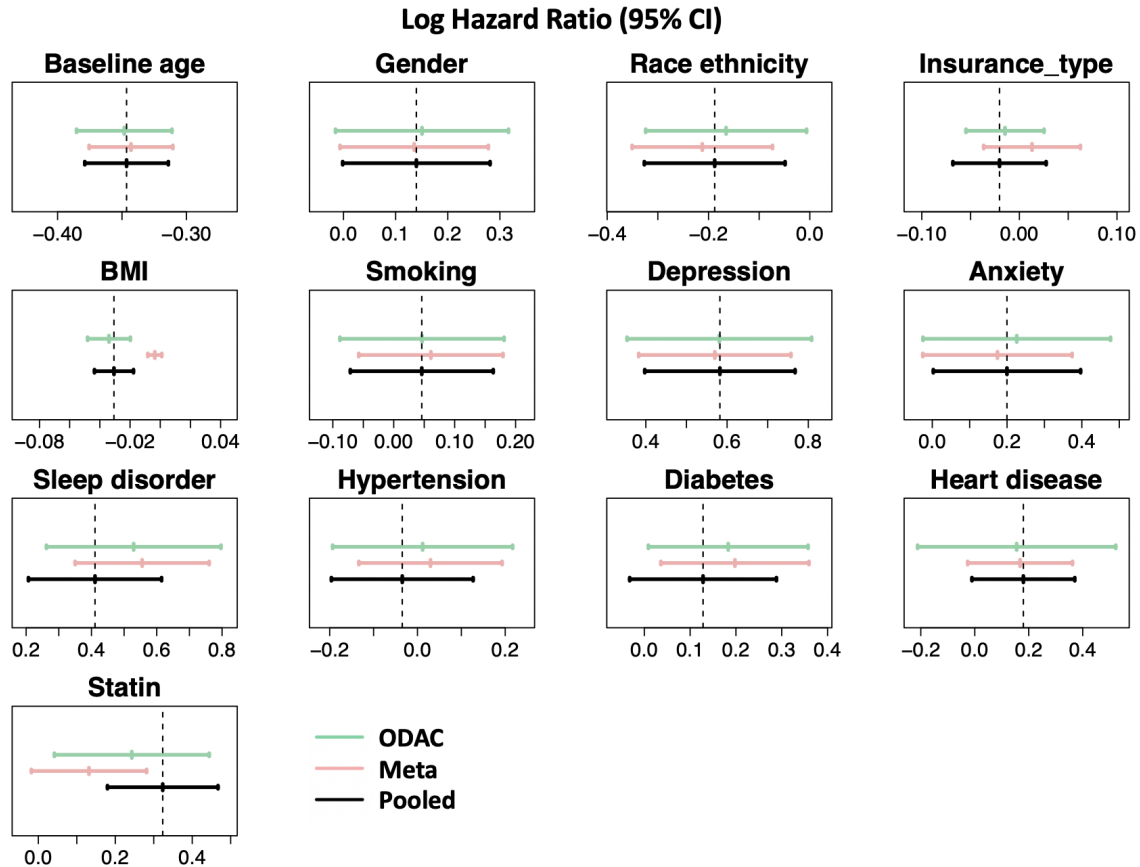
From the analysis, we identified 7 significant risk factors: anxiety, BMI, sleep disorder, statin, age, depression, and race. Most of our findings are consistent with previous reports on risk factors of ADRD. For example, demographic variables such as female, being non-Hispanic Blacks were more likely to have higher risks of ADRD. Cardiovascular conditions, including hypertension, heart diseases could also increase a person’s risk of having ADRD<sup>23–25</sup>. Mental health conditions, such as anxiety, depression, or sleep disorder, have also been demonstrated to be associated with ADRD, as patients who have ADRD tend to have symptoms that lead to decline in mental functions<sup>23,26</sup>. Interestingly, statins use were shown to be positively associated with the occurrence of ADRD in our data, which is contradicted to previous reports<sup>27</sup> that statin may be protective against the progression of AD, this is possibly due to the heterogeneity in the ADRD population<sup>28,29</sup> and potential misclassifications of how medication prescriptions are recorded in the data.

## Discussion and conclusion

In this paper, with the longitudinal EHR data from the OneFlorida CRC, we evaluated the empirical performance of a privacy-preserving, communication-efficient distributed learning algorithm—the ODAC algorithm. We studied the associations between ADRD and 13 risk factors and identified 7 significant factors that are consistent with the previous findings on ADRD. In the evaluation, other than applying ODAC, we also conducted the meta-analysis, which is currently the most popular method for multicenter analysis. From the results, we observed that the ODAC algorithm is able to provide improved estimation of effect sizes for the risk factors compared with the meta-analysis. In ODAC, the strategy that each clinical site serves as the local site reduces the potential impact of one local site on the final estimate. Different from the other existing distributed algorithms, the ODAC is a non-iterative algorithm by constructing the surrogate likelihood function without sharing patient-level information. In a nutshell, ODAC is a

practical, robust, efficient, and accurate algorithm for modeling time-to-event outcomes.

**Figure 2** displays the estimated log hazard ratio and 95% confidence interval for each risk factor using the three methods: ODAC (green), meta-analysis (light pink), and pooled estimate (black).



Even though the pathophysiology of AD and ADRD is not well understood, there is evidence indicating the heterogeneity in AD as well as the heterogeneity in progression to AD through different intermediate disease stages. Identification of the factors contributing to different progression pathways from MCI to AD are crucial for clinical prognostication and risk stratification to guide counseling and selection of potential treatments. RWD from large clinical networks similar to OneFlorida provide the golden opportunity to examine the heterogeneity of AD. Distributed learning algorithms like ODAC, thus, become critical to leverage these RWD to generate RWE. Risk models built with ODAC can help us identify critical factors for both the primary prevention (i.e., from non-AD to AD) and secondary prevention (i.e., from MCI to AD) of AD, which ultimately lead to better care for ADRD patients.

In this study, we only used the structured data from the OneFlorida network, where some other important risk factors, especially social determinants of health (SDoH) were not readily available and thus not included in our analysis. On the other hand, clinical narratives<sup>30</sup> in EHR contain more detailed patient information, including SDoH. Further, we only were able to model ADRD onset as the outcome. To accurately study the progression of ADRD, we would need to be able to extract and model other intermediate outcomes such as neuropsychological tests (e.g., Mini-Mental State Examination and Severe Impairment Battery), that are not typically captured in structured EHR either. In future studies, advanced natural language processing (NLP) methods can be leveraged to extract additional risk factors and neuropsychological test results from clinical narratives.

In the future, the ODAC algorithm can be extended in several aspects. First, the distributed algorithms to integrate and study other types of outcomes can be considered, for example, count data and longitudinal outcomes. Secondly, ODAC is based on the fact that we treat the pooled analysis as the gold standard method, which essentially requires data to be homogenous distributed across sites. However, the effect sizes as well as the baseline hazard function at

each site can be different in practice<sup>31</sup>. To efficiently synthesize evidence from heterogeneous clinical sites in large CRN, we are planning to extend the current distributed algorithms by allowing site-specific hazard function and effect sizes in the future. Thirdly, we will work on developing distributed algorithm to handle time-varying covariates or time-varying coefficients in survival analysis. Lastly, we have been developing an open-source software to implement the ODAC algorithm within CRN to facilitate data integration and nationwide comparative effectiveness studies.

## Acknowledgement

This work was supported in part by NIH grants 1R01AI130460, 1R01LM012607, R01CA246418, R21AG061431 and UL1TR001427, and PCORI grants ME-2018C3-14754. The content is solely the responsibility of the authors and does not necessarily represent the official views of the NIH and PCORI.

## References

1. Chaudhry B, Wang J, Wu S, et al. Systematic review: impact of health information technology on quality, efficiency, and costs of medical care. *Ann Intern Med* 2006; 144: 742–752.
2. Blumenthal D, Glaser JP. Information technology comes to medicine. *N Engl J Med* 2007; 356: 2527–2534.
3. FRAMEWORK FOR FDA'S REAL-WORLD EVIDENCE PROGRAM. U.S. Food & Drug, <https://www.fda.gov/media/120060/download>.
4. Kalbfleisch JD, Prentice RL. *The Statistical Analysis of Failure Time Data*. John Wiley & Sons, 2011.
5. Hripcsak G, Duke JD, Shah NH, et al. Observational Health Data Sciences and Informatics (OHDSI): Opportunities for Observational Researchers. *Stud Health Technol Inform* 2015; 216: 574–578.
6. Collins FS, Hudson KL, Briggs JP, et al. PCORnet: turning a dream into reality. *J Am Med Inform Assoc* 2014; 21: 576–577.
7. Forrest CB, Margolis PA, Bailey LC, et al. PEDSnet: a National Pediatric Learning Health System. *J Am Med Inform Assoc* 2014; 21: 602–606.
8. Shenkman E, Hurt M, Hogan W, et al. OneFlorida Clinical Research Consortium: Linking a Clinical and Translational Science Institute with a Community-Based Distributive Medical Education Model. *Acad Med* 2018; 93: 451–455.
9. Shenkman EA. OneFlorida Clinical Research Consortium. PCORI, <https://www.pcori.org/research-results/2015/oneflorida-clinical-research-consortium> (2019).
10. Alzheimer's Association. 2019 Alzheimer's disease facts and figures. *Alzheimers Dement* 2019; 15: 321–387.
11. Hebert LE, Weuve J, Scherr PA, et al. Alzheimer disease in the United States (2010-2050) estimated using the 2010 census. *Neurology* 2013; 80: 1778–1783.
12. Thalhauser CJ, Komarova NL. Alzheimer's disease: rapid and slow progression. *J R Soc Interface* 2012; 9: 119–126.
13. Silva MVF, Loures C de MG, Alves LCV, et al. Alzheimer's disease: risk factors and potentially protective measures. *J Biomed Sci* 2019; 26: 33.
14. Tucker K, Branson J, Dilleen M, et al. Protecting patient privacy when sharing patient-level data from clinical trials. *BMC Med Res Methodol* 2016; 16 Suppl 1: 77.
15. DerSimonian R, Laird N. Meta-analysis in clinical trials. *Controlled Clinical Trials* 1986; 7: 177–188.
16. Wu Y, Jiang X, Kim J, et al. Grid Binary LOGistic REgression (GLORE): building shared models without sharing data. *J Am Med Inform Assoc* 2012; 19: 758–764.
17. Lu C-L, Wang S, Ji Z, et al. WebDISCO: a web service for distributed cox model learning without patient-level data sharing. *J Am Med Inform Assoc* 2015; 22: 1212–1219.
18. Duan R, Boland MR, Moore JH, et al. ODAL: A one-shot distributed algorithm to perform logistic regressions on electronic health records data from multiple clinical sites. *Pac Symp Biocomput* 2019; 24: 30–41.
19. Duan R, Boland MR, Liu Z, et al. Learning from electronic health records across multiple sites: A communication-efficient and privacy-preserving distributed algorithm. *J Am Med Inform Assoc* 2020; 27: 376–385.
20. Duan R, Luo C, Schuemie MJ, et al. Learning from local to global - an efficient distributed algorithm for modeling time-to-event data. *J Am Med Inform Assoc*. Epub ahead of print 6 July 2020. DOI:10.1093/jamia/ocaa044.
21. Jordan MI, Lee JD, Yang Y. Communication-Efficient Distributed Statistical Inference. *J Am Stat Assoc* 2019; 114: 668–681.
22. Bian J, Loiacono A, Sura A, et al. Implementing a hash-based privacy-preserving record linkage tool in the OneFlorida clinical research network. *Jamia Open* 2019; 2: 562–569.



23. Stampfer MJ. Cardiovascular disease and Alzheimer's disease: common links. *J Intern Med* 2006; 260: 211–223.
24. Qiu C, Winblad B, Fratiglioni L. The age-dependent relation of blood pressure to cognitive function and dementia. *Lancet Neurol* 2005; 4: 487–499.
25. Savva GM, Stephan BC. Alzheimer's Society Vascular Dementia Systematic Review Group. Epidemiological studies of the effect of stroke on incident dementia: a systematic review. *Stroke* 2010; 41: e41–e46.
26. Edwards GA Iii, Gamez N, Escobedo G Jr, et al. Modifiable Risk Factors for Alzheimer's Disease. *Front Aging Neurosci* 2019; 11: 146.
27. Schultz BG, Patten DK, Berlau DJ. The role of statins in both cognitive impairment and protection against dementia: a tale of two mechanisms. *Transl Neurodegener* 2018; 7: 5.
28. Crum J, Wilson J, Sabbagh M. Does taking statins affect the pathological burden in autopsy-confirmed Alzheimer's dementia? *Alzheimers Res Ther* 2018; 10: 104.
29. Chu C-S, Tseng P-T, Stubbs B, et al. Use of statins and the risk of dementia and mild cognitive impairment: A systematic review and meta-analysis. *Sci Rep* 2018; 8: 5804.
30. Meystre SM, Savova GK, Kipper-Schuler KC, et al. Extracting information from textual documents in the electronic health record: a review of recent research. *Yearb Med Inform* 2008; 128–144.
31. Tong J, Duan R, Li R, et al. Robust-ODAL: Learning from heterogeneous health systems without sharing patient-level data. In *Pacific Symposium on Biocomputing*. *Pac Symp Biocomput* 2020; 25:695–706.