Integrating Social Determinants of Health in a Multi-Modal Deep Clustering Survival Model for Injury-Risk in Alzheimer's and Related Dementia Patients

Kazi Noshin*[1], Mary Regina Boland*[3], Bojian Hou[2], Weiqing He[2], Victoria Lu[1], Li Shen†[2], Aidong Zhang†[1]

[1]Department of Computer Science, University of Virginia
[2]Department of Biostatistics, Epidemiology and Informatics, University of Pennsylvania, Philadelphia, PA 19104, USA
[3]Data Science Program, Department of Mathematics, Saint Vincent College, Latrobe, PA 15650, USA
epw9kz@virginia.edu, mary.boland@stvincent.edu, Bojian.Hou@Pennmedicine.upenn.edu, weiqingh@sas.upenn.edu, Li.Shen@pennmedicine.upenn.edu, aidong@virginia.edu

Abstract

As our population ages, the prevalence of Alzheimer's Disease and Related Dementias (ADRD) and its associated burdens continue to rise. Social Determinants of Health (SDOH) significantly influence both ADRD development and progression. Using Electronic Health Records (EHR) from a quaternary care academic medical center in a diverse urban setting, we investigated SDOH's impact on multi-modal deep clustering survival machines. Our findings revealed that SDOH improved model performance across feature selection methods (DeepCox roll-out vs. SHAP DeepExplainer) and EHR clinical modalities (medication vs. laboratory). Additionally, Laboratory features proved more informative than medications for predicting injury-fall risk. Our results highlight SDOH's crucial role in ADRD progression, particularly regarding injury-fall risk. We found that feature importance varied by selection method when analyzing multi-modality EHR data, with education emerging as a key SDOH factor among our top 10 features, underscoring its significance in ADRD progression.

Introduction

The population is aging rapidly. According to the World Health Organization (WHO) one in six people in the world will be 60+ years by 2030 (WHO 2024). In the United States of America (USA), the fifth leading cause of death among the elderly (65+) is Alzheimer's Disease and Related Dementias (ADRD) (ALZ.org 2024). Therefore, ADRD will be a major issue globally as the overall global population ages. NeuroDegenerative Disorders (NDD) is a larger family of conditions, which includes ADRD and additional disorders including Parkinson's Disease (PD), and Amyotrophic Lateral Sclerosis (ALS). NDD are even more common than ADRD, but despite their collective commonality, little is known about risk factors for these patients in their communities. Diversity is often lacking in many published studies on ADRD and NDD despite the commonality of these disorders in diverse populations (including Racial and Ethnic diversity) (Gilmore-Bykovskyi et al. 2019).

Electronic Health Records (EHR) data contains rich information from routine clinical care that can enable research on larger cohorts of patient populations. Additionally, EHR data enables research on diverse populations (Boland, Elhadad, and Pratt 2022) including exploring the role of racial (Canelón, Butts, and Boland 2021), socioeconomic (Boland et al. 2021), geographic (Boland et al. 2021; Bushong et al. 2022), and intersectional (e.g., combinations of racial, socioeconomic and geographical disparities) on health outcomes (Meeker et al. 2022, 2021; Kellier et al. 2024a,b). Prior work has shown that utilizing multi-modal features from EHR data improves feature prediction for ADRD using structured data (Noshin et al. 2025a,b).

Social Determinants of Health (SDOH) are attributes of an individual's environment that can play a role on their health. SDOH includes attributes of individuals' environment outside of their immediate control, including economic, educational, environmental, healthcare and social (Agarwal et al. 2023). They also include behavioral health risk factors, including alcohol use (Rajput, Aziz, and Siddiqui 2019; Chai, Tan, and Dong 2024), tobacco use (Chai, Tan, and Dong 2024), depression (Prather 2020; Remes, Mendes, and Templeton 2021), economic stress (Prather 2020), and physical activity (Chai, Tan, and Dong 2024). SDOH has been implicated in ADRD (Majoka and Schimming 2021). Higher levels of education is protective against ADRD (Majoka and Schimming 2021). Lower socioeconomic status, lower levels of food access and security increase the risk of ADRD (Majoka and Schimming 2021). Body mass index (BMI) also played a role in ADRD (Majoka and Schimming 2021). Transportation, housing, pollution, greening, environment, literacy, isolation, and healthcare also were important SDOH in ADRD (Adkins-Jackson et al. 2023). Neighborhood deprivation (a measure of overall SDOH) has also been implicated in ADRD (Powell et al. 2020).

Combining the study of EHR clinical data with the study of SDOH when investigating factors that increase injury-risk among ADRD patients is a much needed area. However, it raises some additional complexities due to multi-modal nature of EHR data. The purpose of this study is to integrate SDOH information in a multi-modal deep clustering sur-

*These authors contributed equally.
†Correspondence to Aidong Zhang or Li Shen

vival model (DCSM) (Hou et al. 2023, 2024) for injury-risk in ADRD patients from an academic medical center in the United States. Specifically, we adapt DCSM into a multimodal version by injecting different combinations of several sets of features including SODH, demographics, medications etc. to our data. We conduct feature selection by utilizing two methods for interpretability study, i.e., DeepCox (Katzman et al. 2018) roll-out and SHAP Deep Explainer (Lundberg 2017). With different combinations of feature sets (different modalities), we find that the SDOH modality plays an important roles no matter what feature selection methods we are using. In other words, adding SDOH to the features, the final clustering performance in terms of different risks of developing injury will be significantly improved.

Materials and Methods

Dataset

Cohort Selection Our dataset consists of EHR clinical data collected from a quaternary care academic medical center in the USA. This academic medical center is located in a densely populated, urban area with rich diversity in terms of Race, Ethnicity and Socioeconomic status (among other attributes). Our methods involve the use of survival analysis techniques and therefore, we constructed our dataset to have our outcome indicator be a fall/injury (binary outcome variable). To identify patients with a fall or injury, we used a publicly accessible set of PheCodes (Boland 2024).

We want to study the time from individual diagnosis of either Mild Cognitive Impairment (MCI) or a more serious NDD diagnosis to injury-fall. To study this, we identified patients having one of five NDDs: Alzheimer's Disease (AD), Parkinson's Disease (PD), Vascular Dementia (VD), Other Dementias (OD) and Lewy Body (LB) using a combination of PheCodes and direct extraction of relevant International Classification of Diseases (ICD) version 9 (ICD-9) and 10 (ICD-10) codes. We also extract MCI using well known and established ICD-9 and ICD-10 codes (Mao et al. 2023).

We utilized our survival analysis dataset constructed in a previous study that has been annotated with the time from first diagnosis of an NDD or MCI to the outcome of fall. All patients that had a fall prior to diagnosis with either MCI or NDD or who were diagnosed with MCI/NDD for the first time on the same date as the fall were excluded from the study. We also excluded individuals who had only visited the academic medical center system a single time as these were lost to followup. The first date of diagnosis was the start date (for either MCI or NDD). We also excluded patients with missing information in their medications, laboratory data or demographics (e.g., age). Our final cohort used for the survival analysis experiments contained N=29,045 patients.

Feature Preprocessing

Medications We identified all distinct medication features from our de-identified EHR dataset. Previously, these medications were linked to the Observational Health Data Sciences & Informatics (OHDSI) Common Data Model (CDM) framework (Reich et al. 2024), which harmonizes various medication terminology systems to distinct concept codes. These medication concept codes were used as binary features in our models with 1 indicating presence of the medication concept.

Laboratory Results The OHDSI CDM also contains information on how to structure vital signs and laboratory results (Reich et al. 2024). This system was used for our dataset and hence the LOINC (Logical Observation Identifiers Names and Codes) terminology was used for laboratory tests, laboratory values and vital signs and these were provided to used in structured billing code format (McDonald et al. 2003).

These LOINC laboratory codes were used as binary features in our models. The binary columns for vital and laboratory code indicate whether data for the code was recorded on or before the study's start date. If data for a specific code exists, a binary value of 1 is assigned; otherwise, 0 is assigned. The final laboratory dataset contains information for the patients included in our cohort regarding their presence or absence of vital signs and laboratory tests at baseline for each patient.

Demographic Features Our dataset also includes gender and race as demographic features. We transformed the Race variable using one-hot encoding, creating binary variables for each race category: White, Black, Asian, and Other. Similarly, we converted the Hispanic and Male columns into binary variables.

SDOH Features We included 10 SDOH features, 5 derived from the Agarwal paper, namely: economic, education, environment, healthcare and social (Agarwal et al. 2023) and 5 additional SDOH features that are related to behavioral health risk factors and lifestyle, including alcohol use (Rajput, Aziz, and Siddiqui 2019; Chai, Tan, and Dong 2024), tobacco use (Chai, Tan, and Dong 2024), psychoactive substance abuse (Spooner and Hetherington 2005), self-harm (Llamocca et al. 2023) and lack of physical activity (Chai, Tan, and Dong 2024). We identified each of these 10 SDOH using ICD-9 and ICD-10 diagnosis codes.

Feature Selection

To identify key predictive features, we evaluated feature importance using two methods:

- Method 1 (DeepCox roll-out): We extracted and evaluated feature importance based on the model's learned weights by training a deep cox model. The importance score for each feature was computed by multiplying the weight matrices from all sequential layers (Montavon et al. 2019) (namely "roll-out"). The absolute values of the elements were taken to quantify the importance of each feature.
- Method 2 (SHAP DeepExplainer): SHapley Additive exPlanations (SHAP), a method based on Shapley values derived from cooperative game theory. Specifically, SHAP was applied through the deepexplainer framework (Lundberg 2017) to calculate importance scores for each feature. The absolute shapley values were used to quantify the importance of each feature.

The analysis involved 29,045 patient records, encompassing features from medication, laboratory, and combined

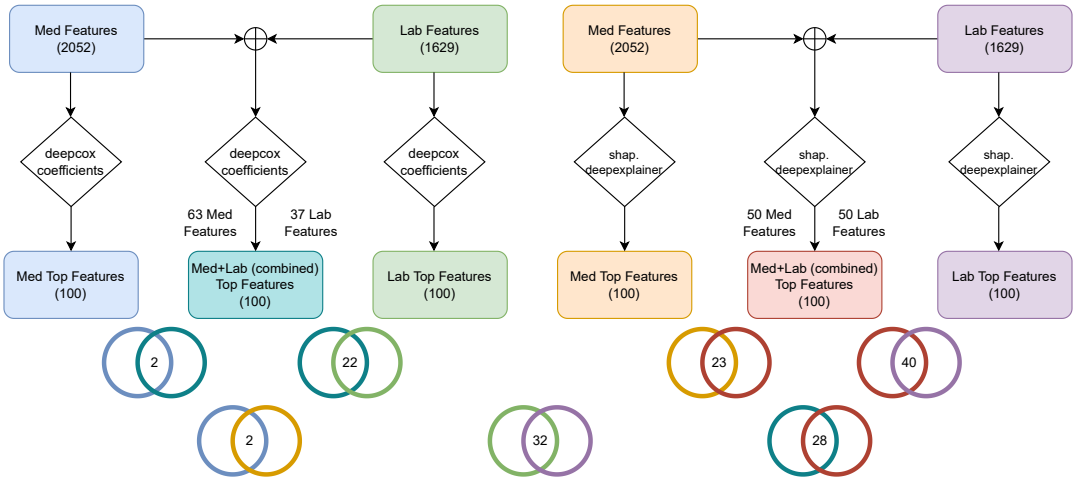

Figure 1: DeepCox roll-out and SHAP-DeepExplainer Intersection

modalities. The diagram in Figure 1 illustrates the feature selection process for medication and laboratory datasets using two methods: DeepCox roll-out and SHAP (DeepExplainer). Our dataset included 2,052 medication features and 1,629 laboratory features. Applying DeepCox roll-out, the top 100 features are selected for both datasets, resulting in 63 medication and 37 laboratory features in the combined dataset. There are 2 and 22 features in the intersection of medication and combined features, and laboratory and combined features respectively. Separately, SHAP (DeepExplainer) is applied to the same datasets, again identifying the top 100 features. From the combined features dataset, the top 100 features comprised 50 medication and 50 laboratory features. There are 23 and 40 features in the intersection of medication and combined features, and laboratory and combined features respectively. Only 2 features are common between the top 100 medication features identified by DeepCox roll-out and the top 100 medication features identified by SHAP (DeepExplainer). A total of 32 features overlap between the top 100 laboratory features identified by DeepCox roll-out and SHAP (DeepExplainer). Of the top 100 combined features selected by SHAP (DeepExplainer) and DeepCox roll-out, 28 features overlap. In summary, we worked with the following four different sets of features in this paper (shown in Figure 2):

1. Feature Set 1
   (a) Med_demo (114 features): Medication features (top 100) + Demographics features (14)
   (b) Med_demo_sdoh (124 features): Medication features (top 100) + Demographics features (14) + SDOH (10)
2. Feature Set 2
   (a) Lab_demo (114 features): Laboratory features (top 100) + Demographics features (14)
   (b) Lab_demo_sdoh (124 features): Laboratory features (top 100) + Demographics features (14) + SDOH (10)
3. Feature Set 3
   (a) Medlab_demo (114 features): Medication + Laboratory features (top 100) + Demographics features (14)
   (b) Medlab_demo_sdoh (124 features): Laboratory features (top 100) + Demographics features (14) + SDOH (10)
4. Feature Set 4
   (a) Combined_Medlab_demo (214 features): Medication + Laboratory features (200) + Demographics features (14)
   (b) Combined_Medlab_demo_sdoh (224 features): Laboratory features (200) + Demographics features (14) + SDOH (10)

Metrics and Settings

Our data $\mathcal{D}$ is a set of tuples $\{\mathbf{x}_i, t_i, \delta_i\}_{i=1}^{N}$ where $\mathbf{x}_i$ is the feature vector associated with the ith instance, t_i is the last-followed time, δ_i is the event indicator, and N is the number of instances. When $\delta_i = 1$ (it means the ith instance is uncensored), t_i will be the time when the event happens whereas when $\delta_i = 0$ (it means the ith instance is censored), t_i will be the time when the instance quits the study or the study ends. Denote the $\mathcal{D}_U$ as the uncensored subset where the corresponding event indicator $\delta = 1$ and $\mathcal{D}_C$ as the censored subset where $\delta = 0$.

We employed two metrics to assess the performance of the methods. For the clustering task, We used LogRank test to evaluate the model's ability to distinguish between risk groups. To assess the time-to-event prediction performance, we used the concordance index (C-index) as a measure of predictive accuracy.

We used 5 different seeds to randomly initialize our models and obtained the mean results along with std and 95% confidence intervals (CIs). We performed this study based on two modalities: medication and laboratory results. We have 4 feature sets with the information from same patients. The entire data set was split into a training set and a held-

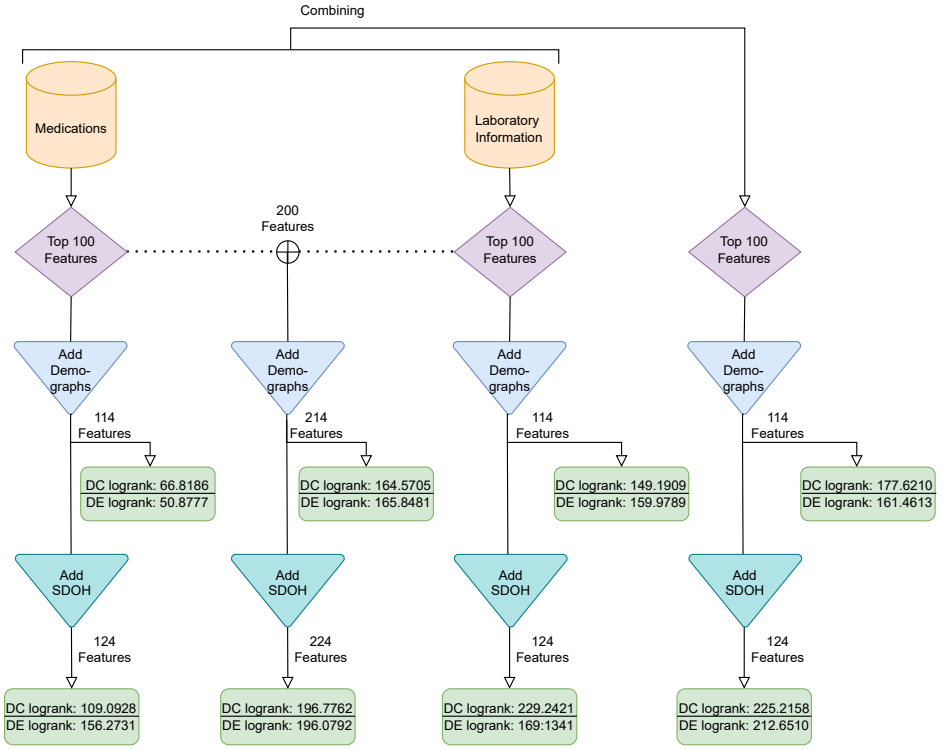

Figure 2: Analysis Flow Diagram with Logrank Results. DC: DeepCox; DE: DeepExplainer

out testing set with a ratio of 7:3. We used 'Optuna' (Akiba et al. 2019) on the training set to optimize hyperparameters for DCSM. The learning parameter step size was chosen from [1e-5, 1e-2]. The layer setting of the multiple perceptron was chosen from [[50], [100], [50, 50]] where "50" and "100" are the number of neurons in each layer. The number of distributions was chosen from [1, 5].

Method

In this paper, we adapt and apply Deep Clustering Survival Machines (DCSM) (Hou et al. 2024) to our data by considering multiple modalities. DCSM is designed to learn a conditional distribution $P(T|X = \mathbf{x})$ by optimizing the maximum likelihood estimation (MLE) of the time T. Similar to the mixture model learning paradigm, the conditional distribution $P(T|X = \mathbf{x})$ is characterized by learning a mixture over K well-defined parametric distributions, referred to as *expert distributions*. In order to use gradient-based methods to optimize MLE, the Weibull distributions are chosen as the expert distributions that are flexible to fit various distributions and have closed-form solutions for the PDF and CDF:

$$\text{PDF}(t) = \frac{\mu}{\sigma}\left(\frac{t}{\sigma}\right)^{\mu-1} e^{-\left(\frac{t}{\sigma}\right)^{\mu}},\ \text{CDF}(t) = e^{-\left(\frac{t}{\sigma}\right)^{\mu}},$$

where μ and σ are the shape and scale parameters separately.

According to the framework of MLE, our goal is to maximize the likelihood with respect to the timing information T conditioned on $\mathbf{x}$. Given that the likelihood functions are different for uncensored and censored data, we calculate them separately. For the uncensored data, the log-likelihood of T is computed as follows, where **ELBO** is the lower bound of the likelihood derived by Jensen's Inequality:

$$\begin{aligned}
&\ln \mathbb{P}(\mathcal{D}_U|\Theta) = \ln\left(\Pi_{i=1}^{|\mathcal{D}_U|}\mathbb{P}(T=t_i|X=\mathbf{x}_i,\Theta)\right)\\
&=\sum_{i=1}^{|\mathcal{D}_U|}\ln\left(\sum_{k=1}^{K}\mathbb{P}(T=t_i|\alpha_k,\mu_k,\sigma_k)\mathbb{P}(\alpha_k|X=\mathbf{x}_i,\boldsymbol{w})\right)\\
&=\sum_{i=1}^{|\mathcal{D}_U|}\ln\left(\mathbb{E}_{\alpha_k}\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})[\mathbb{P}(T=t_i|\alpha_k,\mu_k,\sigma_k)]\right) \qquad (1)\\
&\geq\sum_{i=1}^{|\mathcal{D}_U|}\left(\mathbb{E}_{\alpha_k}\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})[\ln\mathbb{P}(T=t_i|\alpha_k,\mu_k,\sigma_k)]\right)\\
&=\sum_{i=1}^{|\mathcal{D}_U|}\left(\text{softmax}_K(\ln\text{PDF}(t_i|\mu_k,\sigma_k))\right)=\textbf{ELBO}_U(\Theta).
\end{aligned}$$

Similarly, the log-likelihood of T for the censored data is:

$$\begin{aligned}
&\ln \mathbb{P}(\mathcal{D}_C|\Theta) = \ln\left(\Pi_{i=1}^{|\mathcal{D}_C|}\mathbb{P}(T>t_i|X=\mathbf{x}_i,\Theta)\right)\\
&\geq\sum_{i=1}^{|\mathcal{D}_C|}\left(\mathbb{E}_{\alpha_k}\sim(\cdot|\mathbf{x}_i,\boldsymbol{w})[\ln\mathbb{P}(T>t_i|\alpha_k,\mu_k,\sigma_k)]\right)\\
&=\sum_{i=1}^{|\mathcal{D}_C|}\left(\text{softmax}_K(\ln\text{CDF}(t_i|\mu_k,\sigma_k))\right)=\textbf{ELBO}_C(\Theta).
\end{aligned} \qquad (2)$$

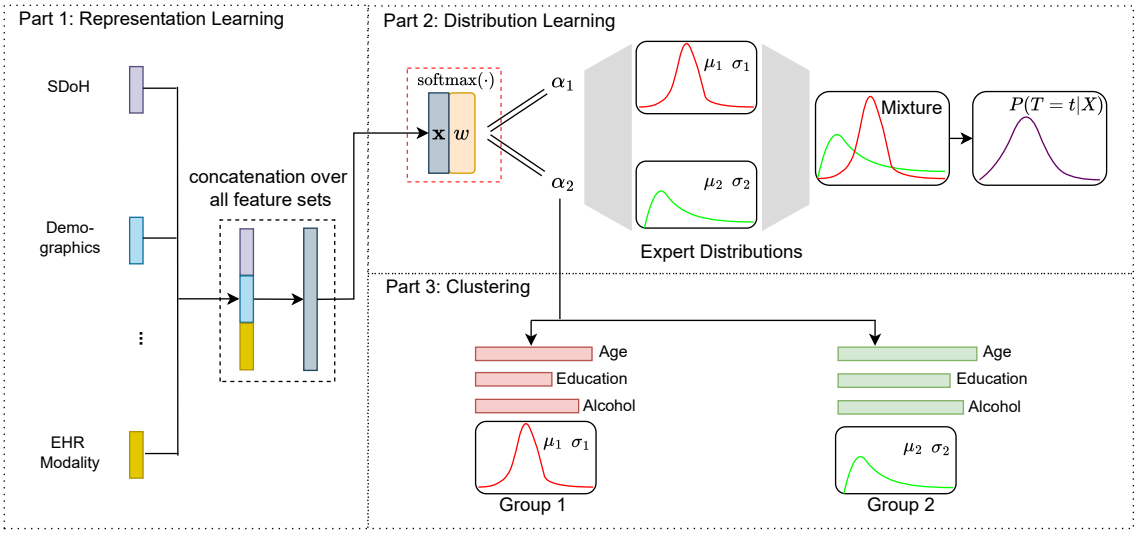

Figure 3: The mechanism of DCSM with multiple modalities. In Part 1, we use several modalities including SDOH, Demographics and EHR etc. We concatenate them together into one single vector. In Part 2, DCSM learns the survival distribution of each patient by weighted combination of two expert distributions. In Part 3, the cluster label for each patient is the index of the largest weight assigned to the expert distribution.

In addition, to stabilize the performance, we incorporate prior knowledge for μ_k and σ_k. Specifically, we minimize the prior loss L_{prior} to make them as close as possible to the μ and σ from the prior model:

$$L_{prior} = \sum_{k=1}^{K} \|\mu_k - \mu\|_2^2 + \|\sigma_k - \sigma\|_2^2. \quad (3)$$

where the prior model is learned by the same MLE framework with a single expert distribution that is still Weill distribution. The final objective L_{all} is the sum of the negative of the log-likelihoods of both the uncensored and censored data in addition to the prior loss where λ is a trade-off hyperparameter:

$$L_{all} = L_{prior} - \mathbf{ELBO}_U(\Theta) - \lambda \cdot \mathbf{ELBO}_C(\Theta). \quad (4)$$

In our case, we only consider two clusters since we mainly want to identify two groups of patient with high risk and low risk respectively. At the same time, we will consider several sets of features (modalities) including SDOH, Demographics, EHR Modalities etc. and concatenate them into one single input vector. Eventually, we have a set of parameters $\Theta = \{\boldsymbol{\theta}, \boldsymbol{w}, \{\mu_k, \sigma_k\}_{k=1}^K\}$ to learn during the training process. Because μ_k and σ_k are the same for different input instances, we clustered each instance/subject according to the weight α_k that is allocated to each expert distribution. Specifically, we assigned an subgroup/cluster indicator k to each instance when the instance's corresponding weight α_k is the largest among all K weights, where $K = 2$ in our case. The mechanism of DCSM is illustrated in Figure 3.

Results

Figure 2 illustrates the feature selection and evaluation process in survival analysis, integrating data from Medications, Laboratory Information, demographic features, and SDOH. The logrank and c-index results for our various feature sets are shown in Figure 4 and Figure 5 respectively.

The inclusion of SDOH consistently improves the logrank scores across all feature sets and for both DeepCox roll-out and SHAP (DeepExplainer) methods. For the medication dataset (top 100 features), the logrank score for DeepCox roll-out increases significantly from 66.8 to 109.1 upon SDOH inclusion, while SHAP (DeepExplainer) exhibits an even larger improvement, rising from 50.9 to 156.3. A similar trend is observed for the laboratory dataset (top 100 features), where DeepCox roll-out's logrank score increases from 149.2 to 229.2, and SHAP's (DeepExplainer) score rises from 160.0 to 169.1. The combination of medication and laboratory datasets (Feature Set 3) further highlights the value of SDOH, with DeepCox roll-out improving from 177.6 to 225.2 and SHAP (DeepExplainer) increasing from 161.5 to 212.7. For the combined dataset (Feature Set 4), which includes all top features from both modalities, the logrank score for DeepCox roll-out rises from 164.6 to 196.8, while SHAP's (DeepExplainer) score improves from 165.8 to 196.1. In most cases, Deepcox roll-out outperforms SHAP (DeepExplainer) in terms of logrank resuls on inclusion of SDOH. These results further underscore the benefit of including SDOH in predictive modeling and the effectiveness of both feature selection methods across all feature sets.

The inclusion of SDOH also results in consistent improvements in C-index values across all feature sets, reflecting enhanced predictive concordance for survival outcomes. For the medication dataset (top 100 features), the C-index for DeepCox roll-out increases from 0.58 (without SDOH) to 0.61 (with SDOH), while SHAP (DeepExplainer) improves

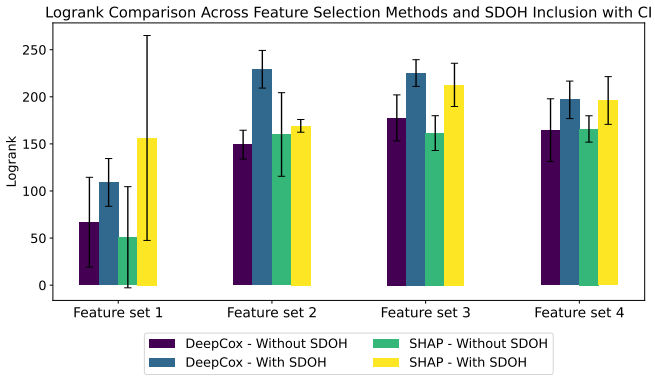

Figure 4: LogRank Comparison Across Feature Selection Methods and SDOH Inclusion with CI

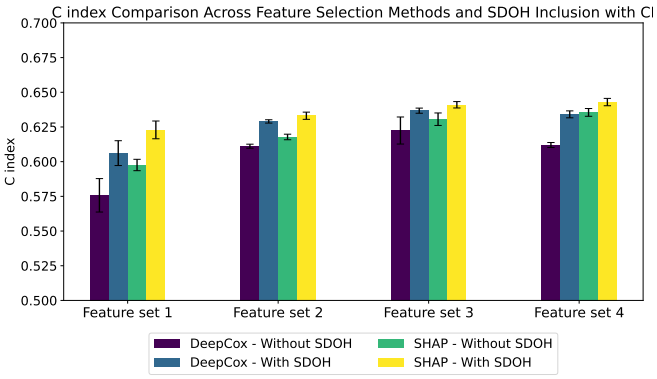

Figure 5: C-index Comparison Across Feature Selection Methods and SDOH Inclusion with CI

from 0.60 to 0.62. Similarly, for the laboratory dataset (top 100 features), DeepCox roll-out's C-index increases from 0.61 to 0.63, and SHAP (DeepExplainer) improves from 0.62 to 0.64. Feature Set 3 achieves further improvements, with DeepCox roll-out increasing from 0.62 to 0.64 and SHAP (DeepExplainer) rising from 0.63 to 0.64. Finally, for the combined dataset (Feature Set 4), DeepCox roll-out's C-index improves from 0.61 to 0.63, while SHAP (DeepExplainer) achieves the value increasing from 0.63 to 0.64. Across all feature sets, SHAP (DeepExplainer) consistently outperforms DeepCox roll-out in C-index, demonstrating its superior ability to leverage both SDOH and diverse feature types for survival prediction.

We extracted the top 10 features from the model with the highest logrank result among the five results using five different seeds for each type shown in Figure 6. Here we used the DCSM"roll-out" method (i.e., multiplying the matrices from all the layers of DCSM) to calculate the feature importance similar to what we did in the feature selection for training. Results are shown before and after inclusion of SDOH. Note that demographic features that appeared in the top 10 are colored in light green and SDOH features that appeared in the top 10 feature results are colored in light orange in Figure 6. Treatment length was a popular demographic feature that appeared in the top 10 feature along with 3 NDD types, namely AD, OD and VD. The only SDOH feature (of the 10 SDOH features included) that appeared in the top 10 was Education (Figure 6). Feature selection method did not alter the selection of Education as one of the top 10 features for the Medications (1b) models, but it did matter in the combined approach (4b) where SHAP (DeepExplainer) selected education as a top 10 feature whereas Deepcox roll-out did not.

The Kaplan-Meier plots are shown in Figure 7 and Weibull plots are shown in Figure 8. The orange curves indicate high-risk groups whereas blue curves indicate low-risk groups. Each row in both figures represents datasets with and without SDOH inclusion, while each column represents different dataset modalities.. Each subplot in Figure 7 shows the LogRank score, indicating the separation between the survival curves of Cluster 0 and Cluster 1. Figure 8 displays Weibull cumulative distribution function (CDF) plots illustrating survival probabilities over time for two expert distributions (labelled as "Expert Distribution 0" and "Expert Distribution 1").

A high injury-risk cluster of NDD patients can be detected using the DCSM algorithm along with a lower injury-risk cluster of patients (Figure 7). While the curves vary somewhat by EHR modality (e.g., medications versus laboratory results) the model still is able to extract a higher-risk cluster of patients and a lower risk cluster of patients as visualized in Figure 7. The inclusion of SDOH consistently improves the LogRank scores across all datasets, demonstrating better differentiation between risk groups. The clusters are more pronounced when SDOH is incorporated, reflecting improved model precision in capturing survival differences between clusters. For instance, the curves for the Laboraty feature sets with SDOH show a sharper decline for blue curve, indicating stronger survival differentiation. Together, these figures underscore the critical role of SDOH in improving survival predictions and the robustness of the underlying survival modeling approach.

Discussion

As the world is aging, more models are needed to understand the role of ADRD and NDD progression. Injury-fall risk is one sign of NDD progression indicating a worsening of the disease state. Methods that can identify features that are predictive of adverse injury-fall risk among those with NDD are needed. Our method that included SDOH features found improvement in the model above what was gained by demographic information alone. This indicates the importance of SDOH in NDD progression.

Overall, we found that adding SDOH features to our DCSM model improved the logrank performance across EHR modalities (laboratory values and medications) and regardless of feature selection method (Deepcox roll-out and SHAP (DeepExplainer)). This indicates the overall need to include SDOH in ADRD studies that utilize EHR data.

The C-index results in Figure 5 reveal that SHAP

Models without SDOH				Models with SDOH			
Deep Cox Feature Selection Method							
1a Meds	2 Labs	3a Meds + Labs	4a Combined	1b Meds	2b Labs	3b Meds + Labs	4b Combined
OD	12227-5	475342	17819-4	6826	12227-5	2339-0	17819-4
AD	2339-0	2867-0	12227-5	400008	17819-4	475342	5693-7
VD	17819-4	2864-7	39354-6	279645	39354-6	2030-5	2098-2
400008	5693-7	20625-0	5693-7	1423803	55147-3	300195	39354-6
3288	55147-3	29943-8	2098-2	VD	2098-2	2864-7	55147-3
6069	2098-2	2873-8	55147-3	3364	5693-7	29943-8	6826
9785	39354-6	2876-1	6826	OD	1992-7	2876-1	1992-7
2176	1992-7	2870-4	1992-7	Education	3330-8	2867-0	2864-7
4077	Treatment Length	12480-0	2867-0	AD	29943-8	12480-0	2867-0
5992	3330-8	2030-5	12480-0	Treatment Length	2876-1	2873-8	2873-8
Deep Explainer Feature Selection Method							
37617	32623-1	12227-5	66869	VD	33037-3	2339-0	7994
66869	12227-5	2021-4	37617	7781	12227-5	475342	2021-4
1539753	Treatment Length	7994	12227-5	37617	Treatment Length	2030-5	1927-3
595060	33037-3	1927-3	7781	66869	OD	300195	12227-5
VD	735-1	58413-6	70727	9071	32623-1	2864-7	14338-8
7781	6690-2	32623-1	595060	Education	3141-9	29943-8	Education
24942	1989-3	33037-3	1539753	1539753	2339-0	2876-1	58413-6
42612	20482-6	8782	2713-6	595060	1994-3	2867-0	8782
10391	5804-0	Treatment Length	71836-1	24942	1989-3	12480-0	32623-1
1908	2028-9	1994-3	735-1	10391	71836-1	2873-8	33037-3

Figure 6: Top 10 Features from the Model with Highest LogRank. Cells are colored orange to indicate SDOH features and green to indicate Demographic features

(DeepExplainer) consistently outperforms DeepCox roll-out across all feature sets, regardless of the inclusion or exclusion of SDOH. This suggests that SHAP (DeepExplainer) exhibits superior concordance in ranking survival times, reflecting its strength in predictive discrimination. Conversely, the Logrank test results in Figure 4 show that DeepCox roll-out generally outperforms SHAP (DeepExplainer) in assessing survival group separation upon the inclusion of SDOH features, except for Feature Set 1. This demonstrates that DeepCox roll-out achieves better stratification of risk groups in the presence of SDOH in our case, which may be critical for capturing more nuanced aspects of survival modeling. The findings suggest that the choice between these methods may depend on the specific requirements of the survival analysis task, whether prioritizing discrimination accuracy or group stratification capability.

The observed inferior performance of Feature Set 1, as seen in both C-index and Logrank results, could be attributed to the limitations in the features included in this set. Medication features probably lacks sufficient diversity or relevance to fully capture the underlying complexities of survival outcomes.

Regarding the SDOH features that were found important by our methods; specifically, we found that education was the SDOH feature that appeared the most among the top 10 features (Figure 6) underscoring the importance of education in ADRD progression (Majoka and Schimming 2021; Butler Jr et al. 2024). Education was the only SDOH feature that appeared in the top 10 feature results. Many studies have found that education is important in altering ADRD development and progression. One study found that higher education levels reduce the risk of ADRD (Majoka and Schimming 2021). Another study found that poor education access and quality increased the risk of ADRD (Joshi and Tampi 2024). A key indicator is highschool completion rate (Butler Jr et al. 2024).

We also found that EHR modality was important with many of the top 10 features including laboratory results and medications (Figure 6). This underscores the importance of medication exposure in injury-fall risk, which has been reported in the literature in several studies both in non-NDD settings (Lavsa et al. 2010) and in ADRD settings (Epstein et al. 2014). Our findings also highlight the importance of laboratory results as indicators of increased injury-fall risk(Suttanon et al. 2013).

Limitations of our study include that we utilized only one academic medical health system's EHR data. While this institution included many hospitals and outpatient clinics, it was all from one major academic medical health system. Therefore, this could introduce various biases due to our EHR vendor and other characteristics that are impossible to assess when utilizing one academic medical health system's

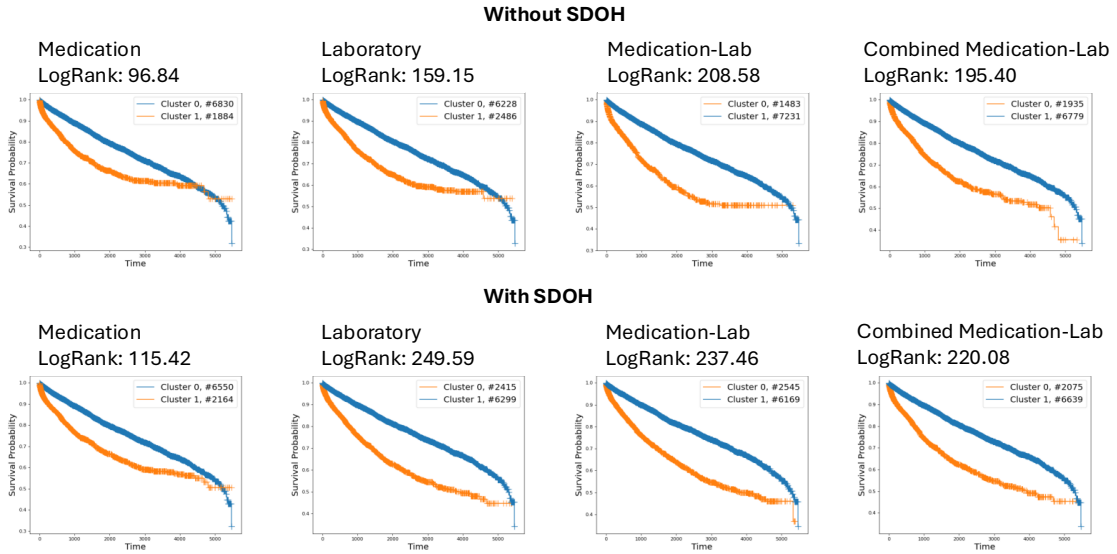

Figure 7: KM Plots from all Modalities

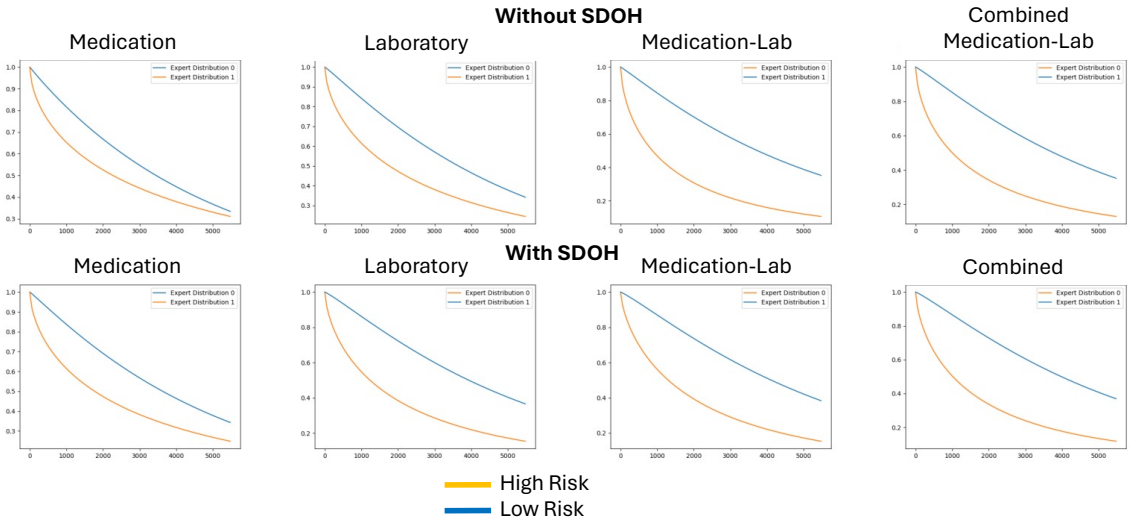

Figure 8: Weibull Plots from all Modalities

data. Another limitation is that we utilized only two EHR modalities (medications and laboratory results) and we did not use the laboratory values themselves and only the presence/absence of the test. More granular results will likely be attainable when utilizing the laboratory value result in addition to the binary information included in our study.

Conclusion

Our study explores the utility of including SDOH in multi-model deep clustering survival models to predict injury-risk among those with NDD, including Alzheimer's Disease and Related Dementias (ADRD). We found that including SDOH improved the logrank and cindex of our results for both EHR modalities (laboratory values, medications) regardless of feature selection method utilized (DeepCox and SHAP (DeepExplainer)). The contribution of SDOH to the models was in addition to patient demographics that were already included in the model. The models showed improvement when SDOH features were included indicating their importance in ADRD progression (as measured via injury-fall risk in this study). We found that education was the SDOH factor that appeared in the top 10 features from our models' results confirming the literature on the role of education in the development and progression of NDD, including ADRD. Interestingly, the literature often highlights the importance of higher education as being protective against development and progression of ADRD. Overall, our results demonstrate the utility of including SDOH along with other EHR modalities for studying ADRD progression from EHR data. We also demonstrate that these features can be utilized to separate patients into high and low risk patient clusters, which may be important for patient care planning.

Acknowledgments

Research reported in this publication/ presentation was supported by the National Institute On Aging of the National Institutes of Health under Award Numbers P30 AG073105, U01 AG066833, U01 AG068057 and R01 AG071470.

References

Adkins-Jackson, P. B.; George, K. M.; Besser, L. M.; Hyun, J.; Lamar, M.; Hill-Jarrett, T. G.; Bubu, O. M.; Flatt, J. D.; Heyn, P. C.; Cicero, E. C.; et al. 2023. The structural and social determinants of Alzheimer's disease related dementias. *Alzheimer's & Dementia*, 19(7): 3171–3185.

Agarwal, A. R.; Prichett, L.; Jain, A.; and Srikumaran, U. 2023. Assessment of Use of ICD-9 and ICD-10 Codes for Social Determinants of Health in the US, 2011-2021. *JAMA Network Open*, 6(5): e2312538–e2312538.

Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; and Koyama, M. 2019. Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

ALZ.org. 2024. Alzheimer's Disease Facts and Figures. *Accessed in July 2024*, https://www.alz.org/alzheimers-dementia/facts-figures .

Boland, M. R. 2024. Boland Lab GitHub: Alzheimer's Disease and Related Dementias (ADRD) Project. *Accessed in September 2024*, https://github.com/bolandlab/AlzheimersDiseaseandRelatedDementias\.

Boland, M. R.; Elhadad, N.; and Pratt, W. 2022. Informatics for sex-and gender-related health: understanding the problems, developing new methods, and designing new solutions.

Boland, M. R.; Liu, J.; Balocchi, C.; Meeker, J.; Bai, R.; Mellis, I.; Mowery, D. L.; and Herman, D. 2021. Association of Neighborhood-Level Factors and COVID-19 infection patterns in Philadelphia using spatial regression. *AMIA Summits on Translational Science Proceedings*, 2021: 545.

Bushong, A.; McKeon, T.; Regina Boland, M.; and Field, J. 2022. Publicly available data reveals association between asthma hospitalizations and unconventional natural gas development in Pennsylvania. *Plos one*, 17(3): e0265513.

Butler Jr, K. R.; Lafferty, D.; Naylor, S. B.; and Tarver MD, K. C. 2024. Social Determinants of Alzheimer's Disease and Dementia: The Mississippi Landscape. *Journal of Public Health in the Deep South*, 4(2): 2.

Canelón, S. P.; Butts, S.; and Boland, M. R. 2021. Evaluation of stillbirth among pregnant people with sickle cell trait. *JAMA network open*, 4(11): e2134274–e2134274.

Chai, X.; Tan, Y.; and Dong, Y. 2024. An investigation into social determinants of health lifestyles of Canadians: a nationwide cross-sectional study on smoking, physical activity, and alcohol consumption. *BMC Public Health*, 24(1): 2080.

Epstein, N. U.; Guo, R.; Farlow, M. R.; Singh, J. P.; and Fisher, M. 2014. Medication for Alzheimer's disease and associated fall hazard: a retrospective cohort study from the Alzheimer's Disease Neuroimaging Initiative. *Drugs & aging*, 31: 125–129.

Gilmore-Bykovskyi, A. L.; Jin, Y.; Gleason, C.; Flowers-Benton, S.; Block, L. M.; Dilworth-Anderson, P.; Barnes, L. L.; Shah, M. N.; and Zuelsdorff, M. 2019. Recruitment and retention of underrepresented populations in Alzheimer's disease research: a systematic review. *Alzheimer's & Dementia: Translational Research & Clinical Interventions*, 5: 751–770.

Hou, B.; Li, H.; Jiao, Z.; Zhou, Z.; Zheng, H.; and Fan, Y. 2023. Deep Clustering Survival Machines with Interpretable Expert Distributions. In *2023 IEEE 20th International Symposium on Biomedical Imaging (ISBI)*, 1–4. IEEE.

Hou, B.; Wen, Z.; Bao, J.; Zhang, R.; Tong, B.; Yang, S.; Wen, J.; Cui, Y.; Moore, J. H.; Saykin, A. J.; Huang, H.; Thompson, P. M.; Ritchie, M. D.; Davatzikos, C.; and Shen, L. 2024. Interpretable deep clustering survival machines for Alzheimer's disease subtype discovery. *Medical Image Analysis*, 97: 103231.

Joshi, P.; and Tampi, R. 2024. Social Determinants of Health for Alzheimer's Disease and Other Dementias. *Psychiatric Annals*, 54(7): e216–e222.

Katzman, J. L.; Shaham, U.; Cloninger, A.; Bates, J.; Jiang, T.; and Kluger, Y. 2018. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Medical Research Methodology*, 18.

Kellier, D.; Anto, M.; Boland, M. R.; Press, C.; Hughes, N.; Ostapenko, S.; Farrar, J.; and Szperka, C. 2024a. Disparities in Pediatric Outpatient Office Visits After First-time Migraine-related Emergency Visit (P1-12.007). In *Neurology*, volume 102, 6635. AAN Enterprises.

Kellier, D.; Anto, M.; Boland, M. R.; Press, C.; Hughes, N.; Ostapenko, S.; Farrar, J.; and Szperka, C. 2024b. Disparities in the Evaluation and Treatment of Pediatric Migraine in the Emergency Department Using a Language-learning Model (P2-12.007). In *Neurology*, volume 102, 3840. AAN Enterprises.

Lavsa, S. M.; Fabian, T. J.; Saul, M. I.; Corman, S. L.; and Coley, K. C. 2010. Influence of medications and diagnoses on fall risk in psychiatric inpatients. *American journal of health-system pharmacy*, 67(15): 1274–1280.

Llamocca, E. N.; Yeh, H.-H.; Miller-Matero, L. R.; Westphal, J.; Frank, C. B.; Simon, G. E.; Owen-Smith, A. A.; Rossom, R. C.; Lynch, F. L.; Beck, A. L.; et al. 2023. Association between adverse social determinants of health and suicide death. *Medical care*, 61(11): 744–749.

Lundberg, S. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874*.

Majoka, M. A.; and Schimming, C. 2021. Effect of social determinants of health on cognition and risk of Alzheimer disease and related dementias. *Clinical Therapeutics*, 43(6): 922–929.

Mao, C.; Xu, J.; Rasmussen, L.; Li, Y.; Adekkanattu, P.; Pacheco, J.; Bonakdarpour, B.; Vassar, R.; Shen, L.; Jiang, G.; et al. 2023. AD-BERT: Using pre-trained language model to predict the progression from mild cognitive impairment to Alzheimer's disease. *Journal of Biomedical Informatics*, 144: 104442.

McDonald, C. J.; Huff, S. M.; Suico, J. G.; Hill, G.; Leavelle, D.; Aller, R.; Forrey, A.; Mercer, K.; DeMoor, G.; Hook, J.; et al. 2003. LOINC, a universal standard for identifying laboratory observations: a 5-year update. *Clinical chemistry*, 49(4): 624–633.

Meeker, J. R.; Burris, H. H.; Bai, R.; Levine, L. D.; and Boland, M. R. 2022. Neighborhood deprivation increases the risk of Post-induction cesarean delivery. *Journal of the American Medical Informatics Association*, 29(2): 329–334.

Meeker, J. R.; Canelón, S. P.; Bai, R.; Levine, L. D.; and Boland, M. R. 2021. Individual-level and neighborhood-level risk factors for severe maternal morbidity. *Obstetrics & Gynecology*, 137(5): 847–854.

Montavon, G.; Binder, A.; Lapuschkin, S.; Samek, W.; and Müller, K.-R. 2019. Layer-wise relevance propagation: an overview. *Explainable AI: interpreting, explaining and visualizing deep learning*, 193–209.

Noshin, K.; Boland, M. R.; Hou, B.; He, W.; Lu, V.; Manning, C.; Shen, L.; and Zhang, A. 2025a. Determining the Importance of Clinical Modalities for NeuroDegenerative Disorders, Alzheimer's Disease, and Risk of Patient Injury Using Machine Learning and Survival Analysis. *AMIA-IS'25: AMIA Informatics Summit*, in press: 10 pages.

Noshin, K.; Boland, M. R.; Hou, B.; Lu, V.; Manning, C.; Shen, L.; and Zhang, A. 2025b. Uncovering Important Diagnostic Features for Alzheimer's, Parkinson's and Other Dementias Using Interpretable Association Mining Methods. *Pacific Symposium of Biocomputing*, 631–647.

Powell, W. R.; Buckingham, W. R.; Larson, J. L.; Vilen, L.; Yu, M.; Salamat, M. S.; Bendlin, B. B.; Rissman, R. A.; and Kind, A. J. H. 2020. Association of Neighborhood-Level Disadvantage With Alzheimer Disease Neuropathology. *JAMA Netw Open*, 3(6): e207559.

Prather, A. A. 2020. Stress is a key to understanding many social determinants of health. *Health Affairs Forefront*.

Rajput, S. A.; Aziz, M. O.; and Siddiqui, M. A. 2019. Social determinants of Health and Alcohol consumption in the UK. *Epidemiology, Biostatistics, and Public Health*, 16(3).

Reich, C.; Ostropolets, A.; Ryan, P.; Rijnbeek, P.; Schuemie, M.; Davydov, A.; Dymshyts, D.; and Hripcsak, G. 2024. OHDSI Standardized Vocabularies—a large-scale centralized reference ontology for international data harmonization. *Journal of the American Medical Informatics Association*, 31(3): 583–590.

Remes, O.; Mendes, J. F.; and Templeton, P. 2021. Biological, psychological, and social determinants of depression: a review of recent literature. *Brain sciences*, 11(12): 1633.

Spooner, C.; and Hetherington, K. 2005. *Social determinants of drug use*. National Drug and Alcohol Research Centre, University of New South Wales Sydney.

Suttanon, P.; Hill, K. D.; Said, C. M.; and Dodd, K. J. 2013. A longitudinal study of change in falls risk and balance and mobility in healthy older people and people with Alzheimer disease. *American Journal of Physical Medicine & Rehabilitation*, 92(8): 676–685.

WHO. 2024. Ageing and health. *Accessed in November 2024*, https://www.who.int/news-room/fact-sheets/detail/ageing-and-health .