

Explainable survival analytics for dementia risk management: Comparing Cox regression and DeepSurv using the national health and aging trends study

Ashrafe Alam^{*1}, Md Golam Rabbani², Gahangir Hossain³ and Sharad Sharma⁴

^{1 2}Department of Information Science, University of North Texas, Denton, Texas, USA

^{3 4}Department of Data Science, University of North Texas, Denton, Texas, USA

Email: MdRabbani@my.unt.edu²; Gahangir.Hossain@unt.edu³; Sharad.Sharma@unt.edu⁴

*Correspondence: ashrafealam@my.unt.edu

Abstract

This study develops an explainable survival-analysis framework for predicting time to first observed dementia and evaluates whether DeepSurv, a neural-network-based survival model, offers predictive and decision-support advantages over conventional Cox proportional hazards (Cox PH) regression. We analyzed longitudinal data from Rounds 1–14 of the National Health and Aging Trends Study (NHATS), including 5,527 older adults and 1,454 observed dementia events. We examined 25 demographic, cognitive, clinical, psychological, behavioral, social, and technology-related predictors. We evaluated model performance using Harrell and Antolini concordance indices, time-dependent AUC, integrated Brier score, paired bootstrap validation, and identical-fold five-fold cross-validation. In identical-fold five-fold cross-validation, Cox PH had a higher mean C-index and lower variability than DeepSurv (0.7532 ± 0.0121 vs. 0.7361 ± 0.0440). On the independent test set, Cox PH and DeepSurv showed comparable discrimination (Harrell C-index = 0.7879 and 0.7833, respectively), and the paired bootstrap difference was small (DeepSurv – Cox PH = -0.0048 , 95% CI $[-0.0112, 0.0021]$). Cox PH showed slightly higher time-dependent AUCs across time horizons, whereas DeepSurv had a marginally lower integrated Brier score (0.1279 vs. 0.1290). Age, self-rated memory, digital communication, online daily activity, and social participation emerged as influential predictors across both modeling approaches. For the highest-risk test participant, a combined favorable scenario was associated with a 35.8% reduction in the DeepSurv risk score; this was a model-based scenario, not a causal effect. The findings underscore that healthcare organizations should balance predictive performance, interpretability, and implementation complexity when evaluating AI-based decision-support models for dementia risk management.

Keywords: Dementia Risk, DeepSurv, Cox Proportional Hazards, Explainable Artificial Intelligence, Healthcare Management, Survival Analysis, NHATS, Decision Support

JEL Classification: I11, M15, O33

Copyright: © 2026 by the authors. Licensee IJBM IEISS, New Zealand. This article is an open access article distributed under the terms and conditions of the [Creative Commons Attribution \(CC BY\) license](https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Dementia is placing real strain on health care systems, aging-services organizations, families, clinicians, and caregivers alike. As the population ages, more people worldwide are developing dementia, driving up demand for screening, treatment, caregiving support, and community resources (Livingston et al., 2020; Nichols et al., 2022). Dementia has long been treated as a clinical or public health issue, but its growing prevalence is now creating a management challenge as well: organizations serving older adults must decide how to allocate limited resources, identify which residents need closer monitoring, and plan for the services those residents will eventually need. Planning is harder than it sounds because decline doesn't happen all at once or along a single track. An older adult's cognitive status may remain stable for a while even as functional limitations or other health conditions worsen, and once dementia emerges, it compounds those problems. This is why we turn to longitudinal predictive modeling: we're not just interested in whether an adverse event happens, but when. Survival models are built for exactly this kind of data, since they can handle follow-up time and account for people who drop out of a study before an event occurs. The U.S. National Health and Aging Trends Study (NHATS) is a longitudinal study of Medicare beneficiaries that tracks cognitive and physical function, chronic conditions, activities of daily living, social engagement, technology use, and health status over time (Freedman & Kasper, 2019). Earlier research using NHATS data has already highlighted the importance of multivariable measures of cognitive and functional decline, along with technology use, for understanding dementia risk among older adults (Ankuda et al., 2021; Drabo et al., 2025; Peng et al., 2024).

The increasing digitization of health care is also important for understanding dementia risk. In health care, patient portals, telemedicine appointments, digital communication, and online health resources and messaging are becoming increasingly essential to health organizations and patient relations. However, evidence shows that mental and social stimulation benefits cognition and brain health in older adults (Floud et al., 2021; Piolatto et al., 2022), and a recent longitudinal study found that digital exclusion is associated with increased dementia risk (Yu et al., 2026). This suggests that digital interaction, internet use, and community engagement may be useful signals for assessing dementia risk and informing data-driven clinical care strategies; however, care must be taken to avoid confusing association with causation. Cox proportional hazards (Cox PH) regression is one of the most widely used statistical techniques for analyzing time-to-event data (Cox, 1972). A key appeal of Cox PH regression is its interpretability: the model can report hazard ratios, which describe the magnitude and direction of the association between predictors and the event hazard. Unfortunately, Cox PH uses a log-linear risk function, which cannot represent complex relationships that do not fit linear and interaction terms. DeepSurv addresses some of these Cox PH limitations by using neural networks in place of the linear predictor while still using the partial likelihood loss function to model complex predictor relationships (Katzman et al., 2018).

The availability of more sophisticated machine-learning methods does not necessarily translate into superior predictive or organizational value. Recent evidence indicates that deep-learning survival models do not consistently outperform established statistical approaches, particularly on structured clinical datasets, where conventional models may already capture much of the relevant predictive information (Wiegrefe et al., 2024). This matters for healthcare organizations because advanced AI models require additional work in model development, computational resources, validation, interpretability, monitoring, and governance. Consequently, the key question is not whether a complex model can be developed, but whether its added complexity yields enough predictive or decision-support value to justify these requirements. Explainability is especially important when predictive models support high-stakes healthcare decisions. Although neural networks can capture complex relationships, their internal decision processes are less directly interpretable than conventional regression coefficients. SHapley Additive exPlanations (SHAP) provide a mechanism for examining how individual predictors contribute to machine-learning predictions at both global and individual levels (Lundberg & Lee, 2017). Nevertheless, post hoc explanations describe model behavior rather than causal relationships and should not substitute for transparent models or professional judgment in high-stakes applications (Rudin, 2019). Comparing DeepSurv SHAP explanations with Cox hazard ratios therefore offers an opportunity to assess whether the two modeling approaches identify consistent risk signals and to identify where their interpretations diverge.

These considerations are also relevant to the broader management literature on artificial intelligence. Organizational value from AI depends not only on technical capability but also on how effectively analytical systems are integrated into organizational processes, governance mechanisms, and decision-making routines. Recent management research emphasizes that sustainable AI value requires organizational capabilities and integration, not technology adoption alone (Jenne, 2026). Similarly, studies of AI-assisted decision-making have highlighted potential gains in predictive performance and operational efficiency, alongside challenges related to data quality, skills, implementation costs, cybersecurity, and regulatory requirements (Bhadauria et al., 2026). For healthcare systems, these considerations require evaluating complex predictive models against transparent benchmarks to determine whether greater algorithmic sophistication provides sufficient incremental managerial value.

In contrast, this study uses data from the longitudinal NHATS Rounds 1–14 to compare Cox PH and DeepSurv in predicting time to first observed dementia. The model includes 25 predictors, such as demographic, cognitive, clinical, psychological, behavioral, social, and technology indicators. Beyond comparing predictive discrimination, the study examines the convergence of Cox hazard ratios with DeepSurv SHAP-based explanations and uses model-based scenario analysis and longitudinal visualizations to explore the potential decision-support value of

explainable survival analytics. This design allows us to assess not only technical performance but also the broader clinical and health-management implications of model complexity.

Accordingly, the study addresses the following research questions:

RQ1: Does DeepSurv provide better predictive discrimination than Cox PH regression for longitudinal dementia-risk prediction?

RQ2: To what extent do Cox PH and DeepSurv identify consistent influential predictors of time to first observed dementia?

RQ3: What potential decision-support implications arise when an explainable deep-survival model performs comparably to a transparent statistical benchmark?

The study makes three main contributions. First, it directly compares a transparent statistical survival model with an explainable deep-learning survival model using the same longitudinal cohort, predictors, and independent test data. Second, it extends model comparison beyond predictive performance by integrating Cox hazard ratios, SHAP-based explanations, model-based scenarios, and longitudinal risk visualization. Third, it contributes to healthcare-management research by introducing a complexity–value alignment perspective, under which greater algorithmic complexity is justified when its incremental predictive or decision-support benefits outweigh the additional requirements for implementation, interpretation, validation, and governance. This perspective positions explainable survival analytics not merely as a technical prediction exercise but as a potential component of responsible, evidence-based healthcare decision support.

2. Literature Review and Conceptual Background

This literature review examines the use of survival analytics and explainable artificial intelligence for dementia risk prediction. It also focuses on the intersection of dementia risk management, digital and social engagement, predictive analytics, and explainable AI in healthcare decision-making. Additionally, it compares conventional Cox proportional hazards regression with deep-learning survival models and examines the organizational implications of adopting increasingly complex predictive technologies. Thus, the review establishes the conceptual basis for evaluating predictive performance, interpretability, and managerial value within a framework of complexity–value alignment.

2.1 Dementia Risk Prediction and Healthcare Management

Dementia is not only a clinical and public health challenge but also a long-term management challenge for organizations serving older adults. Rising prevalence increases demand for screening, care coordination, caregiver support, community services, and long-term resource planning (Livingston et al., 2020; Nichols et al., 2022). Longitudinal risk prediction can support these functions by estimating how the probability of an adverse outcome changes over time, rather

than simply classifying whether an outcome will occur. NHATS is particularly valuable for this purpose because it repeatedly measures cognitive, functional, health, behavioral, social, and technology-related characteristics of older Medicare beneficiaries (Freedman & Kasper, 2019). Prior NHATS research has used these data to study disability, cognitive impairment, dementia risk, and technology-related behaviors (Ankuda et al., 2021; Drabo et al., 2025; Peng et al., 2024). More broadly, recent studies have used statistical and machine-learning methods to predict dementia risk, leveraging routinely available demographic, clinical, functional, and other predictors (Brain et al., 2024; Camacho et al., 2025; Alam et al., 2026). Related prognostic research has also developed time-to-event models for outcomes among community-dwelling older adults living with dementia, illustrating the broader role of longitudinal risk modeling in aging populations (Deardorff et al., 2022). From a management perspective, however, a predictive model is useful only when its performance translates into understandable, operationally feasible decisions. This makes dementia analytics a socio-technical problem involving model performance, professional judgment, organizational processes, and resource allocation.

2.2 Digital Engagement and Social Participation in Older Adults

Digital communication and social participation are increasingly relevant to managing older-adult services because healthcare organizations rely on patient portals, telehealth, online information, and remote communication. Evidence also implies that social relationships and cognitive or social activities are associated with cognitive outcomes (Floud et al., 2021; Piolatto et al., 2022), and longitudinal research has linked digital isolation to dementia risk (Yu et al., 2026). Within NHATS, researchers have examined technology-use measures in relation to dementia and aging outcomes (Peng et al., 2024). These findings support including digital communication, online daily activity, online health engagement, and social participation in the present model. Their interpretation must remain predictive rather than causal: technology use may also reflect education, socioeconomic resources, cognitive capacity, health status, or social support. For healthcare managers, the practical importance lies in identifying potential risk markers and barriers to digitally mediated care rather than determining whether technology use itself prevents dementia.

2.3 AI-Based Prediction, Explainability, and Organizational Value

Machine-learning methods have increasingly been considered for clinical prediction and decision support as healthcare data have grown larger and more complex (Obermeyer & Emanuel, 2016). However, technological capability alone does not guarantee organizational value. Recent research argues that AI creates more sustainable value when embedded in organizational routines, governance structures, quality processes, and learning capabilities, rather than being treated as an isolated tool (Jenne, 2026). Research on AI-assisted financial decision-making similarly identifies enhanced prediction and operational efficiency, along with challenges related to data quality, skills, implementation costs, cybersecurity, and regulatory compliance (Bhadauria et al., 2026). These concerns are directly relevant to healthcare analytics, where complex models must justify

the additional implementation and governance burden. Explainability is central to this research. SHAP can describe global and individual feature contributions (Lundberg & Lee, 2017), but post hoc explanations do not establish causality and should not replace transparent models for high-stakes decisions (Rudin, 2019). Consequently, the present study compares DeepSurv explanations with Cox hazard ratios as a cross-model interpretability audit.

2.4 Cox PH, DeepSurv, and Complexity-Value Alignment

Cox proportional hazards regression remains a standard survival-analysis method because it accommodates censoring and yields directly interpretable hazard ratios (Cox, 1972). Its conventional formulation, however, represents covariate effects via a log-linear risk function. DeepSurv retains the Cox partial-likelihood framework but replaces the linear predictor with a neural network to capture more complex relationships (Katzman et al., 2018). Neural-network survival methods extend conventional Cox modeling by allowing more flexible representations of covariate–risk relationships while keeping time-to-event prediction as the primary objective (Kvamme et al., 2019). Other deep-learning survival approaches have extended this framework to model complex time-to-event relationships, including those in competing-risks settings (Lee et al., 2018).

These considerations motivate the complexity–value alignment framework proposed in this study, which distinguishes three related yet conceptually distinct dimensions: predictive performance, implementation cost, and organizational value. Predictive performance refers to a model’s empirical effectiveness in discrimination, calibration, error reduction, and consistent performance on validation samples. Implementation cost includes the required resources and the organizational factors that must be addressed when developing, computing, validating, interpreting, integrating, monitoring, and governing the model. Organizational value means that a model creates significant advantages for decision-making, such as enhanced risk stratification, personalized explanations, workflow efficiency, or resource allocation. In such cases, increased complexity is reasonable if the added value of predictions or organizational value exceeds the extra costs. If the two models have similar predictive power, the Cox PH model may be preferred for its simplicity and interpretability, which make it easier to explain, validate, and govern. At the same time, DeepSurv can add value through nonlinear modeling, personalized explanations, or scenario analysis.

3. Research Methodology

This study used a longitudinal survival-analysis framework to compare a conventional statistical model with a deep-learning approach for predicting time to dementia diagnosis. The analytical workflow included cohort construction, predictor specification, leakage-safe preprocessing, model development and hyperparameter optimization, independent test-set evaluation, internal validation, and explainability analysis. Cox proportional hazards regression model served as the transparent statistical benchmark, whereas DeepSurv captured potentially nonlinear relationships among predictors. We evaluated model performance using multiple discrimination and prediction-

error metrics and assessed predictor contributions and potential decision-support value using SHAP-based explanations and model-based scenarios.

3.1 Data Source, Validation, and Reproducibility

The analysis used longitudinal data from the National Health and Aging Trends Study (NHATS), comprising 43,345 person-round observations and 42 variables across Rounds 1–14 (National Health and Aging Trends Study [NHATS], 2024). After validating required variables and recoding NHATS-specific missing-value codes, the cleaned dataset included 7,410 unique participants with no duplicate participant-round records. Overall, cell-level missingness was 21.93%. To enhance reproducibility, we used a fixed random seed of 42 throughout the computational workflow and enabled deterministic operations wherever supported.

In addition to evaluation on the independent held-out set, we performed five-fold cross-validation on the development data to assess model stability. We selected five-fold cross-validation to balance performance estimation with the computational cost of repeated DeepSurv training. We used identical folds for Cox PH and DeepSurv to allow direct comparison. Within each fold, missing-value imputation was fitted exclusively to the training partition and then applied to the validation partition. We standardized predictors for DeepSurv within each training fold. We calculated Harrell's C-index for each validation fold for both models, allowing direct, fold-matched comparison of discrimination.

Table 1 presents results from five-fold cross-validation, providing additional evidence of comparative model stability. Cox PH achieved a mean C-index of 0.7532 ± 0.0121 , compared with 0.7361 ± 0.0440 for DeepSurv. The mean fold-wise difference (DeepSurv – Cox PH) was -0.0171 . DeepSurv performed better in folds 1 and 2, whereas Cox PH performed better in folds 3–5. DeepSurv also exhibited greater between-fold variability, with C-indices ranging from 0.6726 to 0.7788, compared with 0.7369 to 0.7709 for Cox PH.

Table 1. Five-Fold Cross-Validation Performance of Cox PH and DeepSurv Using Identical Folds

Fold	Cox PH C-index	DeepSurv C-index
1	0.7709	0.7788
2	0.7532	0.7749
3	0.7369	0.6726
4	0.7540	0.7378

5	0.7508	0.7163
Mean ± SD	0.7532 ± 0.0121	0.7361 ± 0.0440

3.2 Construction of the Time-to-Dementia Cohort

The primary endpoint was first observed dementia, not all-cause mortality. For each participant, we ordered observations by follow-up time. If dementia was observed, we set duration to the time of the first dementia-positive record and coded the event as 1; otherwise, we right-censored the participant at the last observed follow-up time and coded the event as 0. We excluded participants with nonpositive or missing survival duration. This procedure yielded 5,527 participants, 1,454 dementia events, a 26.31% event rate, and a median follow-up of 4.0 years. The notebook separately audited mortality and functional decline as potential secondary outcomes, but neither outcome was present in the supplied analysis CSV; consequently, no mortality or functional-decline models were estimated.

3.3 Predictor Specification and Domain Audit

Twenty-five predictors were prespecified across seven domains: demographic (age category, race, marital status, region, residential status, gender); cognitive (self-rated memory); and clinical (stroke, multiple falls, hospitalization, ADL impairment, frailty, fall risk, mobility-aid use, an alternative ADL indicator, health condition, and worry-related limitation). ADL-related measures captured limitations in activities of daily living, a construct commonly operationalized using established functional-status frameworks (Katz et al., 1963; Lawton & Brody, 1969). Missingness varied substantially. Table 2 shows that online daily activity (61.48%), online health engagement (61.50%), multiple falls (70.58%), worry-related limitation (72.53%), and mobility-aid use (73.66%) had particularly high missingness rates. We retained these variables because they represented prespecified clinical or technology domains of interest, but estimates involving them require cautious interpretation. We used leakage-safe iterative imputation to prevent validation or test observations from influencing imputation; however, this procedure does not remove uncertainty associated with extensive or potentially informative missingness.

Table 2. Predictor Missingness Requiring Particular Caution

Predictor	Domain	Missing (%)	Current handling / robustness need
Online daily activity	Technology	61.48	Iterative imputation; exclusion sensitivity recommended

Online health Technology	61.50	Iterative imputation; sensitivity recommended	exclusion
Multiple falls	Clinical 70.58	Iterative imputation; sensitivity recommended	exclusion
Worry-related limitation	Clinical 72.53	Iterative imputation; sensitivity recommended	exclusion
Mobility-aid use	Clinical 73.66	Iterative imputation; sensitivity recommended	exclusion

Note. Leakage-safe imputation prevents validation/test leakage but does not remove uncertainty caused by extensive or potentially informative missingness.

We prespecified 25 predictors across seven domains: demographic, cognitive, clinical, psychological, behavioral, social, and technological. Clinical predictors included measures of functional limitation, such as ADL and instrumental activities of daily living relevant to independent living (Katz et al., 1963; Lawton & Brody, 1969). We represented frailty using a deficit-accumulation approach, in which more accumulated health deficits indicate greater frailty (Rockwood & Mitnitski, 2007). Psychological predictors included anxiety and depressive symptoms, with depressive symptoms assessed using the two-item Patient Health Questionnaire framework (Kroenke et al., 2003). Missingness varied substantially. Several variables were nearly complete, whereas online daily activity (61.48%), online health engagement (61.50%), multiple falls (70.58%), worry-related limitation (72.53%), and mobility-aid use (73.66%) had high missingness. This pattern makes leakage-safe imputation particularly important.

3.4 Development, Validation, and Test Partitions

We partitioned the analytic cohort using stratified random sampling to maintain comparable proportions of dementia events across datasets. We reserved 30% of the cohort as an independent test set ($n = 1,659$; 436 events; event rate = 26.28%) and did not use it during model development or hyperparameter tuning. The remaining 70% constituted the development sample and was further divided into training and validation sets at an 80:20 ratio. This resulted in 3,094 participants in the training set (814 events; 26.31%) and 774 participants in the validation set (204 events; 26.36%). Overall, we allocated approximately 56% of participants to training, 14% to validation, and 30% to independent testing. The similar event rates across all three partitions indicate that the stratification procedure successfully preserved the distribution of dementia events. Table 3 summarizes the cohort partitions.

Table 3. Analytic Cohort and Data Partitions

Dataset	N	Dementia events	Event rate (%)	Analytical role
Full analytic cohort	5,527	1,454	26.31	Overall analysis
Training set	3,094	814	26.31	Model development
Validation set	774	204	26.36	Hyperparameter selection
Independent test set	1,659	436	26.28	Final model evaluation

Note. Median follow-up for the full analytic cohort was 4.0 years. Stratified partitioning maintained a similar proportion of first-observed dementia events across the training, validation, and independent test sets

3.5 Leakage-Safe Missing-Data Processing and Standardization

Missing-data handling is an important consideration in statistical inference because the mechanism underlying missingness can affect the validity of the resulting estimates (Rubin, 1976). We handled missing predictor values using scikit-learn's IterativeImputer. We fitted the imputation model only on the training data, then applied it to the validation and independent test sets to prevent information leakage across data partitions. The procedure used median initialization and a maximum of 20 iterations. After imputation, we standardized predictor variables with StandardScaler. We estimated scaling parameters from the training set and applied them unchanged to the validation and test sets. During five-fold cross-validation, imputation and standardization were performed independently within each training fold before being applied to the corresponding validation fold. This fold-specific preprocessing ensured that validation observations did not influence imputation or feature scaling during model development.

3.6 Cox Proportional Hazards Benchmark

We fitted a penalized Cox proportional hazards (Cox PH) model using the same 25 predictors as DeepSurv, with a ridge penalty of 0.05 to stabilize coefficients. Predictor effects were summarized as hazard ratios (HRs) with 95% confidence intervals, and discrimination was evaluated on the independent test set using Harrell's C-index (Harrell, 2015). We assessed the proportional-hazards assumption after model fitting using Schoenfeld residual-based tests with a rank-time transformation. Predictor-specific p-values were examined at a nominal .05 threshold and interpreted with caution given the multiplicity inherent in testing 25 covariates. When this assumption was violated, we treated the corresponding Cox coefficient as an average association over follow-up rather than a strictly time-constant effect (Stensrud & Hernán, 2020).

3.7 DeepSurv Architecture and Bayesian Hyperparameter Optimization

We implemented DeepSurv using torchtuples as a neural-network extension of the Cox proportional hazards model. Hyperparameters were optimized with Optuna's Tree-structured Parzen Estimator (TPE) over 30 trials (Akiba et al., 2019). The search considered network architectures, dropout rates, learning rates, batch sizes, and batch normalization. We trained models with the Adam optimizer for up to 250 epochs, with early stopping to prevent overfitting.

3.8 Predictive Evaluation, Calibration, and Internal Validation

We evaluated both Cox PH and DeepSurv on the same independent test set using Harrell's C-index, Antolini's time-dependent concordance index (Antolini et al., 2005), integrated Brier score (IBS; Graf et al., 1999), and inverse-probability-of-censoring-weighted cumulative and dynamic AUC at 1, 3, 5, and 10 years. We computed IBS over a common evaluation interval. We quantified uncertainty in the held-out C-index comparison using 500 paired bootstrap resamples (Efron, 1979). We evaluated calibration for Cox PH and DeepSurv at the four prediction horizons. We grouped participants into predicted-risk groups based on model-predicted event probabilities and compared mean predicted probabilities with Kaplan–Meier-estimated observed event probabilities. The evaluation therefore considered discrimination, prediction error, calibration, and internal validation rather than relying on a single performance metric (Collins et al., 2024).

3.9 Cross-Validation, Sensitivity Analysis, and Explainability

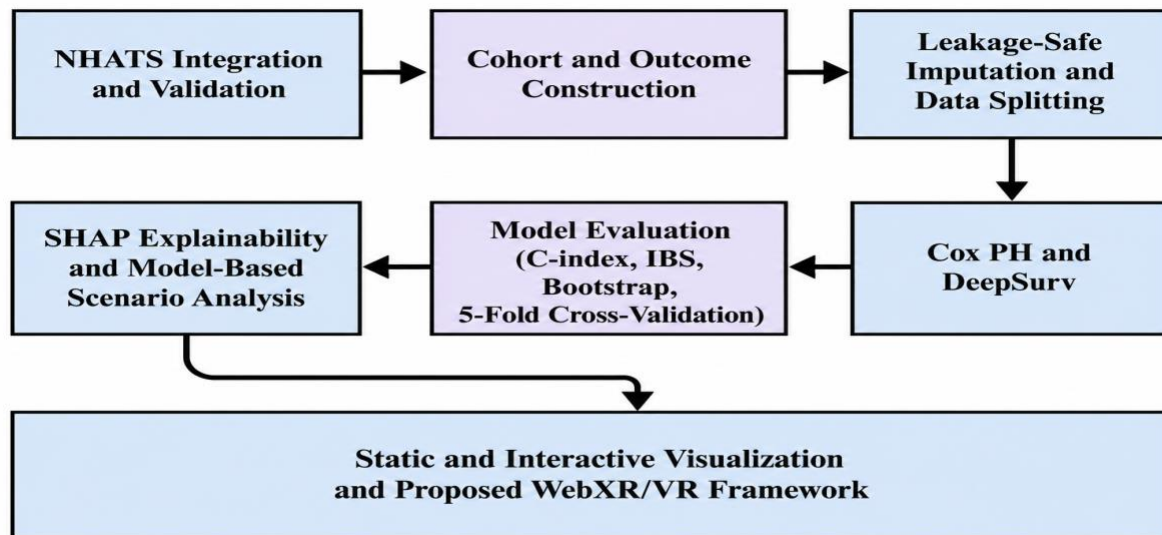
We conducted five-fold cross-validation with identical folds for Cox PH and DeepSurv, fitting imputation independently within each training fold and estimating DeepSurv standardization without access to the corresponding validation fold. We also evaluated four alternative DeepSurv network architectures to assess architectural sensitivity. We used SHAP KernelExplainer to characterize DeepSurv predictor contributions at the global and participant levels. We defined global importance as the mean absolute SHAP value. Cox PH predictors were ranked by the absolute magnitude of the fitted log-hazard coefficient, whereas DeepSurv predictors were ranked by mean absolute SHAP values. We assessed agreement in predictor importance rankings between the two models using Top-5 and Top-10 overlap.

3.10 Model-Based Scenario and Longitudinal Visualization Analyses

The scenario analysis deliberately selected the highest-risk test participant as an illustrative extreme-risk reference profile. For categorical predictors, a feasible scenario shifted the value by one observed level; for continuous predictors, the change was set to 20% of the observed range. The combined scenario included only changes that reduced the fitted DeepSurv risk score for this participant. Because the analysis concerns one extreme-risk profile, the resulting percentage changes are participant-specific demonstrations of model behavior and are not population-average effects, treatment effects, or evidence that the same changes would reduce dementia risk in other participants. The notebook also generated longitudinal normalized DeepSurv risk trajectories,

tertile-based risk groups, a temporal predictor heatmap, and interactive visualization outputs. The WebXR/VR component was conceptual, not deployed or user-tested.

Figure 1. Overall Analytical Workflow



To determine whether the scenario patterns occurred outside the intentionally selected high-risk case, we also analyzed a representative participant from the median-risk group. This was an observed test-set participant whose DeepSurv risk score was closest to the median predicted risk in the test set.

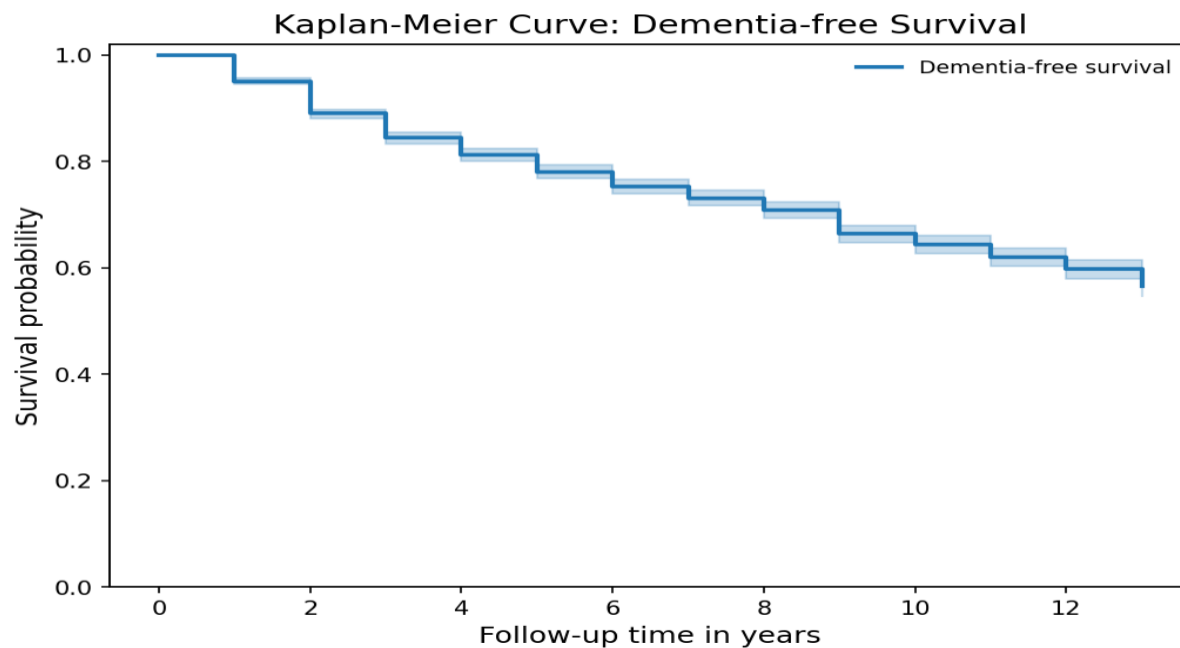
4. Results

This section presents empirical findings from the longitudinal survival analysis. The results first describe the analytic cohort and data quality, then present Cox PH estimates and compare the predictive performance of Cox PH and DeepSurv. We examine time-dependent discrimination, internal validation, and architecture sensitivity to assess model stability. Finally, SHAP-based explainability and model-based scenario analyses evaluate predictor convergence and DeepSurv's potential decision-support value. Together, these analyses address the study's three research questions on predictive performance, explanatory consistency, and incremental managerial value.

4.1 Cohort and Data Quality

The final survival cohort comprised 5,527 participants, of whom 1,454 developed first-observed dementia during follow-up (26.31%). The median follow-up was 4.0 years. The original longitudinal source contained 43,345 records from 7,410 participants, with no duplicate participant-round records. Overall missingness was 21.93%, concentrated in several technology and clinical variables. Figure 2 shows dementia-free survival over follow-up.

Figure 2. Kaplan-Meier Estimate of Dementia-Free Survival in the NHATS Analytic Cohort



4.2 Cox PH Estimates and Proportional-Hazards Diagnostics

The penalized Cox PH model achieved a held-out Harrell C-index of 0.7879. Age category showed the strongest positive association with time to first observed dementia (HR = 1.713, 95% CI 1.568–1.871, $p < 0.001$). Digital communication (HR = 0.772, 95% CI 0.708–0.841, $p < 0.001$), social participation (HR = 0.754, 95% CI 0.655–0.869, $p < 0.001$), and online daily activity (HR = 0.824, 95% CI 0.688–0.986, $p = 0.035$) were associated with a lower hazard. Poorer self-rated memory (HR = 1.227, 95% CI 1.134–1.328, $p < 0.001$), mobility-aid use (HR = 1.327, $p = 0.014$), race (HR = 1.119, $p = 0.004$), region (HR = 1.083, $p = 0.010$), residential status (HR = 1.292, $p = 0.023$), and anxiety (HR = 1.083, $p = 0.048$) were associated with a higher hazard. Schoenfeld residual diagnostics identified nominal proportional-hazards violations for gender ($p = 0.0032$) and social participation ($p = 0.0294$), whereas the remaining 23 predictors had $p \geq 0.05$. Accordingly, we interpreted the effects for the two flagged predictors as average associations over follow-up rather than strictly time-invariant effects. Table 4 shows full Cox estimates and predictor-specific PH diagnostics.

Table 4. Penalized Cox Proportional Hazards Model and Predictor-Specific Proportional-Hazards Diagnostics

Predictor	β	HR	95% CI	p-value	PH test p
Age category	0.538	1.713	1.568–1.871	<0.001	0.118
Digital communication	-0.259	0.772	0.708–0.841	<0.001	0.098
Self-rated memory	0.205	1.227	1.134–1.328	<0.001	0.273
Social participation	-0.282	0.754	0.655–0.869	<0.001	0.029
Race	0.113	1.119	1.038–1.208	0.004	0.479
Stroke	0.165	1.179	0.992–1.402	0.062	0.841
Online daily activity	-0.194	0.824	0.688–0.986	0.035	0.850
Multiple falls	0.089	1.093	0.933–1.280	0.270	0.436
Vigorous physical activity	-0.061	0.941	0.862–1.028	0.176	0.088
Anxiety	0.080	1.083	1.001–1.172	0.048	0.370
Marital status	0.068	1.070	0.960–1.192	0.220	0.462
Hospitalization	0.073	1.076	0.939–1.233	0.294	0.285
ADL impairment	0.048	1.050	0.964–1.143	0.266	0.988
Depression	0.071	1.074	0.989–1.166	0.090	0.983
Frailty	0.014	1.014	0.871–1.181	0.853	0.267
Fall risk	0.028	1.028	0.851–1.242	0.775	0.937
Mobility-aid use	0.283	1.327	1.059–1.663	0.014	0.888
Online health engagement	-0.145	0.865	0.650–1.151	0.321	0.978

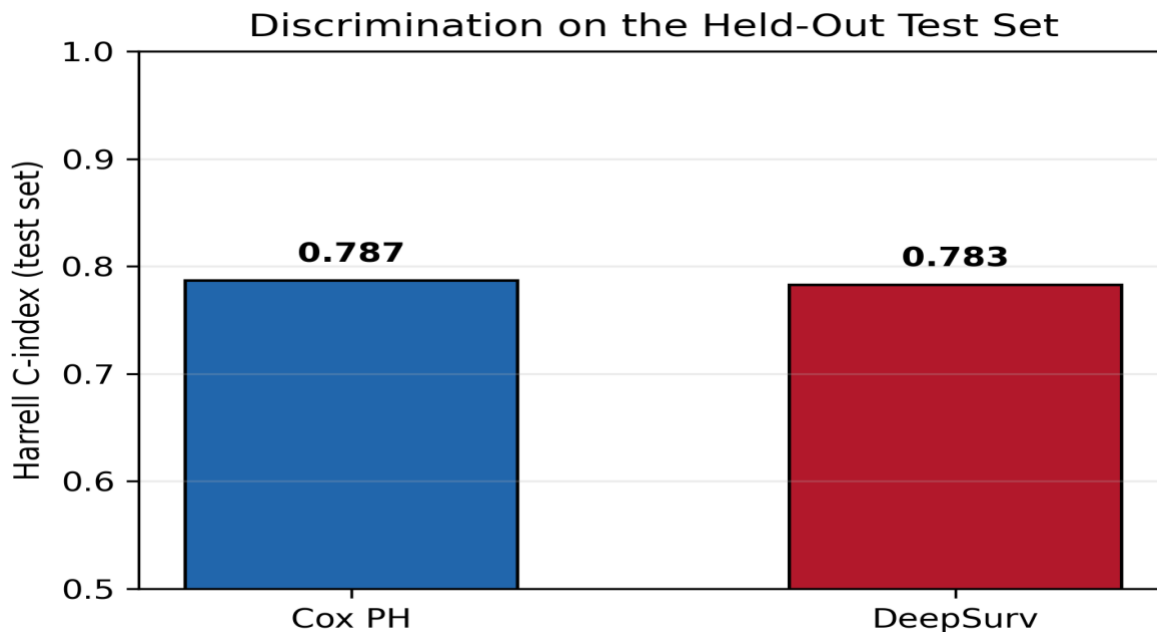
Predictor	β	HR	95% CI	p-value	PH test p
Alternative ADL indicator	0.108	1.114	0.960–1.293	0.154	0.628
Health condition	0.034	1.035	0.892–1.201	0.650	0.476
Worry-related limitation	-0.065	0.937	0.793–1.107	0.447	0.224
Region	0.080	1.083	1.019–1.150	0.010	0.902
Social networks	-0.007	0.993	0.947–1.041	0.766	0.855
Residential status	0.256	1.292	1.035–1.613	0.023	0.784
Gender	-0.083	0.921	0.810–1.046	0.203	0.003

Note. HR = hazard ratio; CI = confidence interval; ADL = activities of daily living. The model used ridge penalization = 0.05.

4.3 Cox Versus DeepSurv Performance

On the independent test set, Cox PH achieved a Harrell C-index of 0.7879, compared with 0.7833 for DeepSurv. Antolini C-indices were likewise 0.7879 for Cox PH and 0.7833 for DeepSurv. The integrated Brier score was 0.1290 for Cox PH and 0.1279 for DeepSurv, indicating a very small advantage for DeepSurv in overall prediction error. Across 500 paired bootstrap resamples, the mean Harrell C-index was 0.7878 for Cox PH and 0.7830 for DeepSurv. The mean paired difference (DeepSurv – Cox PH) was -0.0048 (95% CI $[-0.0112, 0.0021]$). Because the confidence interval included zero, the analysis did not indicate a meaningful discrimination advantage for DeepSurv. Table 5 summarizes the comparative model performance measures, while Figure 3 presents the held-out Harrell C-index comparison.

Figure 3. Held-out Harrell C-index with paired-bootstrap 95% confidence intervals for Cox PH and DeepSurv



4.4 Time-Dependent Discrimination and Calibration

Table 5 shows that Cox PH and DeepSurv exhibited closely aligned time-dependent discrimination across the evaluated prediction horizons. Cox PH had slightly higher AUCs at 1, 3, 5, and 10 years, though the absolute differences between models were small. We interpreted the longer-horizon estimates, particularly at 10 years, cautiously because of reduced follow-up support.

Table 5. Comparative Predictive Performance of Cox PH and DeepSurv

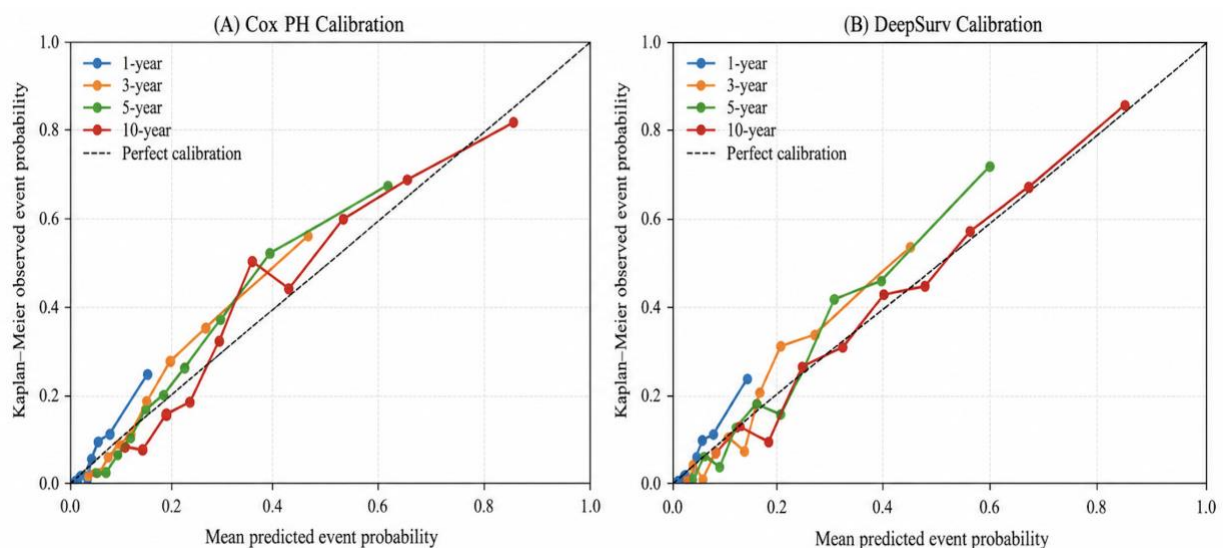
Performance metric	Cox PH	DeepSurv	Difference (DeepSurv – Cox PH)
Held-out Harrell C-index	0.7879	0.7833	–0.0046
Bootstrap mean C-index	0.7878	0.7830	–0.0048
Bootstrap 95% CI	0.7637 – 0.8089	0.7600 – 0.8043	—
1-year time-dependent AUC	0.8528	0.8473	–0.0055
3-year time-dependent AUC	0.8365	0.8336	–0.0029

Performance metric	Cox PH	DeepSurv	Difference (DeepSurv – Cox PH)
5-year time-dependent AUC	0.8440	0.8384	–0.0056
10-year time-dependent AUC	0.8488	0.8447	–0.0041
Antolini C-index	0.7879	0.7833	–0.0046
Integrated Brier score	0.1290	0.1279	–0.0011

Note 1. Lower IBS indicates lower prediction error. Paired bootstrap difference in Harrell C-index (DeepSurv – Cox PH) = –0.0048, 95% CI [–0.0112, 0.0021].

We assessed calibration for Cox PH and DeepSurv at 1-, 3-, 5-, and 10-year prediction horizons (Figure 4). At each horizon, we grouped participants by predicted event probability and compared mean predicted probabilities with Kaplan–Meier-estimated observed event probabilities. Figure 4A presents the Cox PH calibration results, and Figure 4B presents the DeepSurv results. Both models generally followed the 45-degree reference line, with departures in some predicted-risk groups across the evaluated horizons. The calibration patterns did not indicate a consistent advantage of either model across all prediction horizons.

Figure 4. Calibration of Cox PH and DeepSurv Models at 1-, 3-, 5-, and 10-Year Prediction Horizons. (A) Cox PH calibration. (B) DeepSurv calibration.



Note. Mean model-predicted dementia event probabilities are plotted against Kaplan–Meier-estimated observed event probabilities across predicted-risk groups. The dashed 45-degree reference line represents perfect agreement between predicted and observed probabilities. Calibration was evaluated on the held-out test set.

4.5 Cross-Model Predictor Importance and Agreement

Table 6 compares Cox PH associations with DeepSurv global SHAP importance. Age category ranked first in DeepSurv importance, followed by digital communication, self-rated memory, online daily activity, and mobility-aid use. For cross-model comparison, we ranked Cox PH predictors by the absolute magnitude of the fitted log-hazard coefficient ($|\beta|$), whereas we ranked DeepSurv predictors by mean absolute SHAP values. Three of the Top 5 predictors (60%) and seven of the Top 10 predictors (70%) were shared across the models, indicating substantial overlap among the most influential predictors. Some differences were also observed. ADL impairment ranked tenth in DeepSurv SHAP importance despite a nonsignificant Cox association (HR = 1.050, $p = 0.266$), indicating that it contributed to DeepSurv predictions without, by itself, establishing a nonlinear or interaction effect.

Table 6. Cross-Model Comparison of Cox PH Associations and DeepSurv SHAP Importance

Predictor	Cox HR	95% CI	p-value	Mean SHAP	SHAP rank
Age category	1.713	1.568–1.871	< 0.001	0.3694	1
Digital communication	0.772	0.708–0.841	< 0.001	0.1724	2
Self-rated memory	1.227	1.134–1.328	< 0.001	0.1329	3
Online daily activity	0.824	0.688–0.986	0.035	0.0807	4
Mobility-aid use	1.327	1.059–1.663	0.014	0.0722	5
Race	1.119	1.038–1.208	0.004	0.0581	6
Social participation	0.754	0.655–0.869	< 0.001	0.0542	7
Anxiety	1.083	1.001–1.172	0.048	0.0467	8
Hospitalization	1.076	0.939–1.233	0.294	0.0458	9
ADL impairment	1.050	0.964–1.143	0.266	0.0434	10

Note. SHAP rank is based on mean absolute SHAP value. SHAP importance describes model contribution and should not be interpreted causally.

To assess robustness to extensive missingness in predictors, we conducted a sensitivity analysis excluding variables with more than 60% missing observations. We excluded five predictors (online daily activity, online health engagement, multiple falls, worry-related limitation, and mobility-aid use), leaving 20 predictors. We then refitted Cox PH and DeepSurv with the reduced predictor set and evaluated them on the same independent test set using Harrell's C-index. Table 7 presents the results after removing five predictors, leaving 20 of the original 25. After exclusion, the Cox PH C-index remained essentially unchanged, increasing slightly from 0.7879 to 0.7888, whereas the DeepSurv C-index decreased from 0.7833 to 0.7594. Consequently, the difference between DeepSurv and Cox PH increased from -0.0046 in the primary analysis to -0.0294 in the reduced-predictor analysis. These findings support the primary conclusion that DeepSurv did not outperform Cox PH, while also showing that DeepSurv was more sensitive to removing predictors with extensive missingness.

Table 7. Sensitivity Analysis of Model Discrimination After Excluding Predictors with More Than 60% Missingness

Analysis	Cox PH C-index	DeepSurv C-index	DeepSurv - Cox
Primary, 25 predictors	0.7879	0.7833	-0.0046
Excluding > 60% missing, 20 predictors	0.7888	0.7594	-0.0294

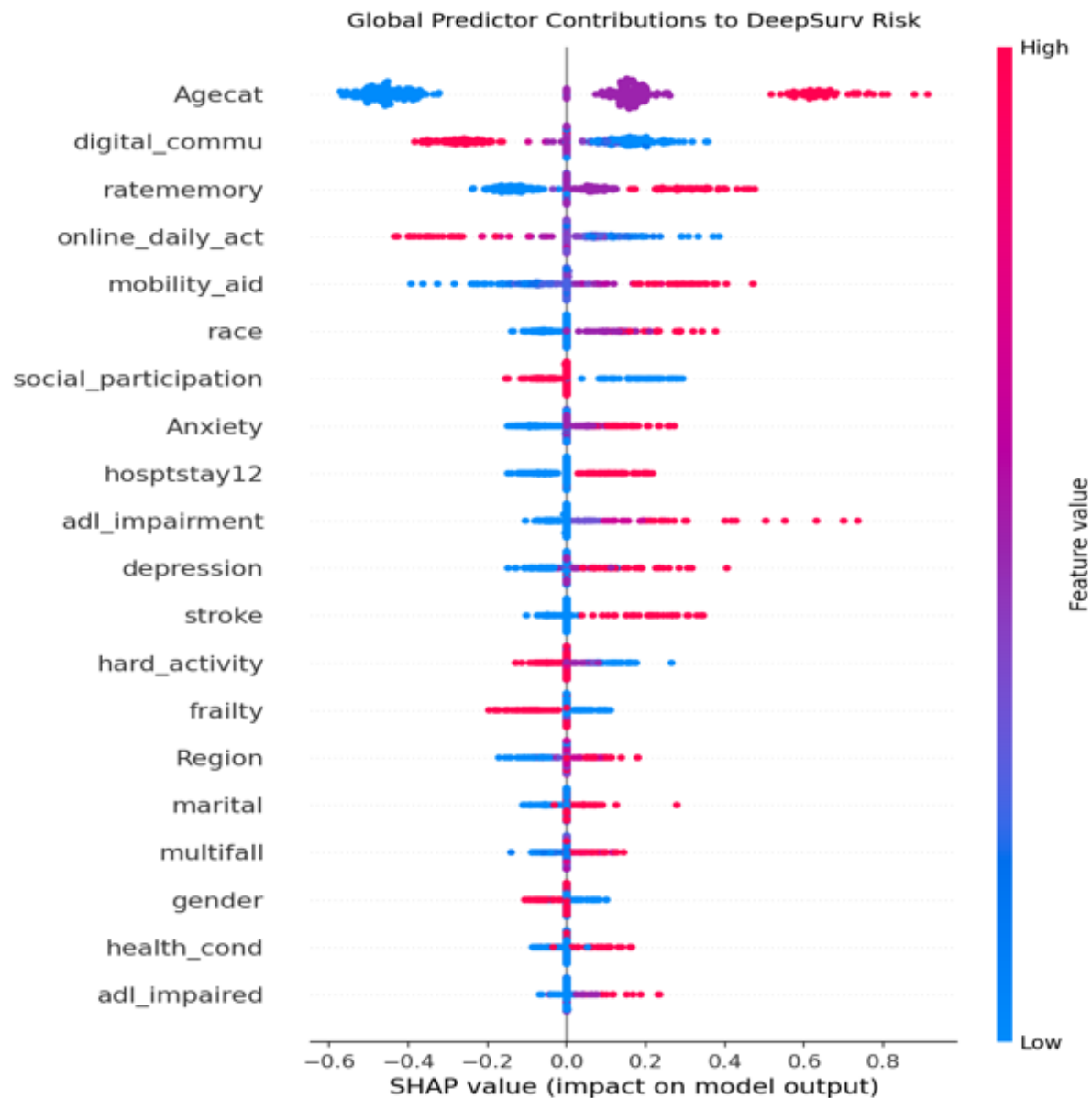
4.6 SHAP and Cox Convergence

Figure 5 presents the global SHAP summary for the DeepSurv model. Age category showed the largest overall contribution to model predictions, followed by digital communication, self-rated memory, online daily activity, mobility-aid use, race, and social participation. The SHAP value distribution also illustrates the direction and heterogeneity of predictor contributions across participants. Higher age-category values were predominantly associated with positive SHAP values, indicating higher model-predicted risk, whereas higher digital communication and online daily activity values were generally associated with negative SHAP values. Higher self-rated memory and mobility-aid use tended to contribute positively to predicted risk, while higher social participation generally contributed negatively. Several clinical and psychological variables, including anxiety, hospitalization, ADL impairment, depression, and stroke, also contributed to individual predictions but had lower global importance. Overall, the SHAP results indicate that DeepSurv incorporated information from demographic, cognitive, technology-related, functional,

and social domains when generating individualized risk predictions. These SHAP values describe the fitted model's behavior and should not be interpreted as causal effects.

From a decision-support perspective, it offers a transparent view of how DeepSurv uses information across demographic, cognitive, functional, social, and technology-related domains, complementing the population-level hazard-ratio interpretation from Cox PH.

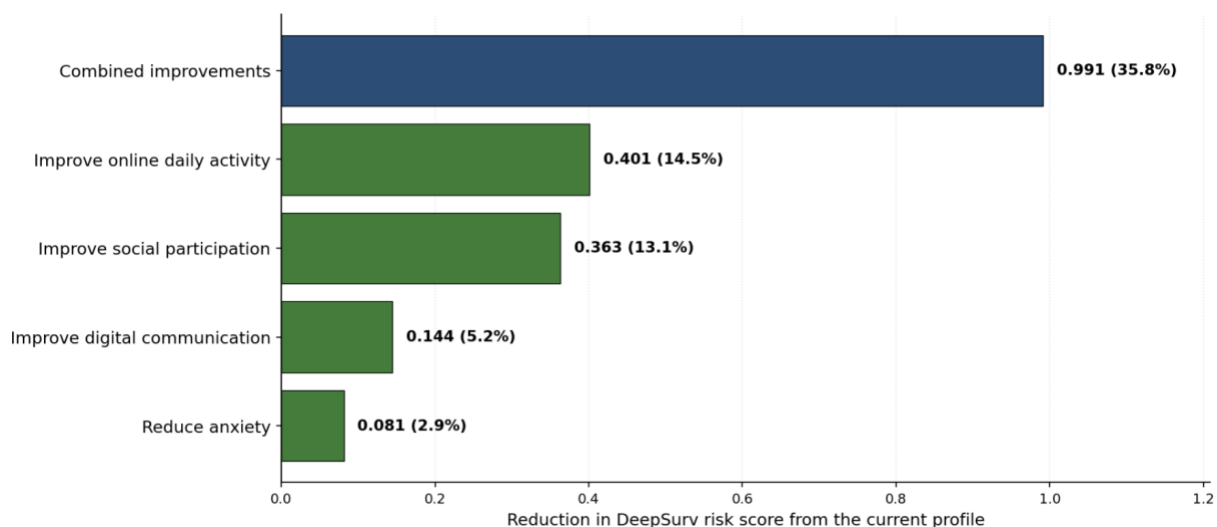
Figure 5. Global SHAP summary for DeepSurv



4.7 Model-Based Scenarios

Figure 6 presents the model-based scenario analysis for the deliberately selected highest-risk test participant, whose baseline DeepSurv risk score was 2.7682. Reducing anxiety from 1 to 0 lowered the score to 2.6868 (2.9%); improving social participation from 1 to 2 lowered it to 2.4048 (13.1%); improving digital communication from 0 to 1 lowered it to 2.6240 (5.2%); and improving online daily activity from 1 to 2 lowered it to 2.3672 (14.5%). The combined favorable profile lowered the score to 1.7774, a 35.8% reduction in the model-derived score. Because the reference participant was selected specifically for the highest predicted risk, these values represent an illustrative extreme-risk case and should not be generalized to typical participants or the overall cohort. They are model-based scenario outputs, not causal treatment effects. Analyses using median-risk and other representative profiles would help assess the stability and practical relevance of these patterns.

Figure 6. Illustrative model-based scenario analysis for the highest-risk test participant



To assess whether the scenario patterns extended beyond the deliberately selected highest-risk case, we repeated the analysis for a representative participant whose baseline DeepSurv risk score was closest to the median predicted risk in the independent test set. As shown in Table 8, scenario responses varied across baseline-risk profiles.

Table 8. Comparison of Model-Based Scenario Changes for Highest-Risk and Median-Risk Test Participants

Predictor	Highest-risk case	Median-risk case
Reduce anxiety	0.081 (2.9%)	0.000

Improve digital communication	0.144 (5.2%)	0.000
Improve social participation	0.363 (13.1%)	0.224
Improve online daily activity	0.401 (14.5%)	0.163
Combined improvements	0.991 (35.8%)	0.386

Note. Values represent reductions in the fitted DeepSurv risk score relative to each participant's baseline profile. Scenario results are participant-specific model outputs and should not be interpreted as causal effects.

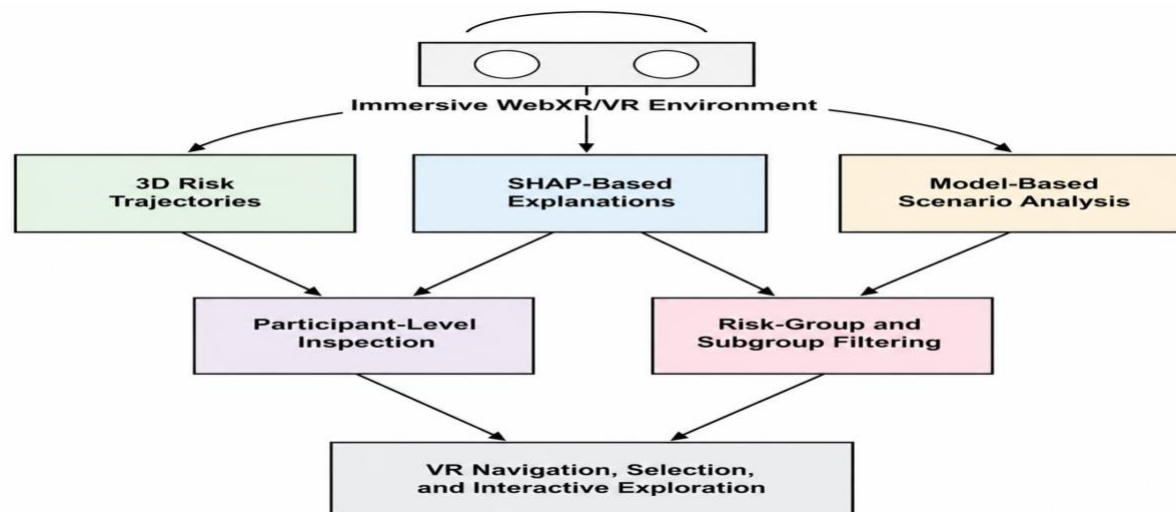
For the median-risk participant, reducing anxiety and improving digital communication produced no measurable change in the fitted DeepSurv risk score, whereas improvements in social participation and online daily activity reduced the score by 0.224 and 0.163, respectively. The combined favorable scenario produced the largest reduction (0.386). These profile-dependent differences illustrate that individualized scenario responses vary by participant characteristics and baseline risk. As with the highest-risk case, these changes reflect fitted-model behavior and should not be interpreted as causal intervention effects.

4.8 Longitudinal and Exploratory Visualization Outputs

These visualizations are supplementary exploratory outputs, not validated decision-support tools. The WebXR/VR component is retained only as a conceptual future-interface example and is not treated as an empirical contribution of the study. We did not deploy or evaluate it with healthcare professionals (Figure 7).

We developed a conceptual interactive visualization framework to illustrate potential presentations of longitudinal risk trajectories and explainability outputs. We also developed an exploratory WebXR/VR visualization concept to show how such an interface could look in the future, including representations of longitudinal risk trajectories and explainability results. Immersive, three-dimensional visualization approaches have been explored for interactive analytics and medical applications (Kraus et al., 2022; Yeung et al., 2021). Because this interface has not been implemented or tested with users, it was not included in the model's empirical evaluation.

Figure 7. Conceptual WebXR/VR-Enhanced Framework for Future Longitudinal Dementia-Risk Visualization



Note. The immersive interface was neither implemented nor user-tested and is presented only as a conceptual illustration of a potential future decision-support interface; it is not an empirically validated contribution of the present study.

5. Discussion

In the direct comparison, we found that the DeepSurv model did not discriminate statistically better than the Cox PH model. Although the neural network model had greater flexibility and provided case-specific interpretations, its added complexity did not confer a clear predictive advantage. This underscores the need for clear benchmarking when analyzing health care data.

The similar discrimination achieved by both models is consistent with previous research showing that deep-survival models do not necessarily outperform conventional survival models when well specified on structured data (Wiegerebe et al., 2024). The results therefore support using transparent statistical models as benchmarks when evaluating more complex AI-based survival approaches.

Predictor analyses showed strong alignment between the two methods. Age group emerged as the most robust predictor, and self-assessed memory function proved relevant in both Cox regression and SHAP analyses. These results align with prior literature on age-related factors and cognitive markers linked to dementia susceptibility (Livingston et al., 2020; Drabo et al., 2025). Digital communication, online activities, and social participation also emerged as influential predictors. This pattern is broadly consistent with previous research showing that social engagement, digital connectedness, and participation in cognitive and social activities are associated with cognitive outcomes among older adults (Floud et al., 2021; Piolatto et al., 2022; Yu et al., 2026). These results show statistical associations, not causal evidence that increased tech adoption or social engagement alone prevents dementia.

Some differences between the models remained informative. For example, ADL impairment was not statistically significant in the Cox model but contributed to DeepSurv predictions according to SHAP. This divergence should not, by itself, be interpreted as evidence of nonlinearity or interaction; rather, it shows that the predictor played a different role in the fitted neural survival model than in the global Cox coefficient. Cox PH therefore offers directly interpretable hazard-ratio estimates, whereas DeepSurv provides a complementary framework for individualized post hoc explanations and flexible risk modeling (Katzman et al., 2018).

The scenario analysis further illustrated DeepSurv's potential value for exploratory decision support. For the highest-risk test participant, favorable model-based shifts in online daily activity, social participation, digital communication, and anxiety were associated with lower predicted risk, and the combined scenario reduced the DeepSurv risk score by 35.8%. Because the reference case was deliberately selected as the highest-risk participant, this result is an illustrative extreme-risk scenario and should not be generalized to the full cohort. These changes are model-derived predictions, not causal intervention effects.

From a healthcare management perspective, the results support this complexity–value alignment approach. Organizations should not automatically select a more sophisticated AI model if a simpler, more transparent one delivers comparable predictive performance. DeepSurv may still offer complementary value, including nonlinear modeling, individualized explanations, scenario analysis, and longitudinal visualization. Ultimately, healthcare organizations should weigh predictive performance, interpretability, implementation, and governance requirements when selecting AI-based decision-support models.

6. Managerial and Practical Implications

The main implication for healthcare managers is that model choice should reflect the incremental value the organization can achieve, not solely the algorithm's complexity. When a traditional model such as Cox Proportional Hazards delivers predictive performance comparable to a complex deep learning model, transparency, interpretability, ease of implementation, validation processes, and governance should drive the decision. The additional value that DeepSurv brings through personalization, flexibility, and what-if analysis should justify the resources required for implementation and project management.

6.1 Model Selection

When Cox PH and DeepSurv deliver comparable predictive performance, the transparent model offers practical advantages in interpretation, validation, implementation, and governance. DeepSurv should therefore be considered when its additional capabilities provide demonstrated incremental value that justifies its greater complexity and variability. Such value may include individualized explanations, nonlinear risk characterization, or scenario-based analyses that

improve decision-making beyond what Cox PH provides. The present study did not directly evaluate these potential organizational benefits, which require prospective validation.

6.2 Implementation and Governance

Implementing such technologies should also address how to incorporate predictive models ethically into healthcare processes. Sophisticated models are likely to require more computation, validation, explainability, monitoring, and governance, especially when predictions inform risk stratification or resource allocation. Thus, explainability outputs should support, not override, professional judgment. The discrepancy between the Cox Proportional Hazards and DeepSurv models might also warrant additional review rather than being interpreted as a sign of model malfunction. Furthermore, predictors related to technology and social engagement might help organizations identify potential obstacles to digital care, but these relationships cannot be taken as evidence of cause and effect.

6.3 Performance Monitoring

Model evaluations should extend beyond implementation. Healthcare organizations must monitor not only discrimination but also calibration, subgroup performance, process times, referrals, utilization, end-user understanding, overrides, and any unintentionally introduced disparities. This is required to determine whether the predictive accuracy has practical value within an organization. Because this study focuses on predictive and analytical outcomes rather than managerial outcomes, these suggestions should be regarded as recommendations for future research and application.

6.4 Practical Application Considerations

Prediction and analytical outcomes dominate the results: these are not yet direct measures of management outcomes. The managerial implications should therefore be interpreted as potential applications whose validity requires future evaluation. Nearly equal performance by Cox PH and DeepSurv supports healthcare organizations setting benchmarks based on transparent statistical models when appraising complex AI systems. Healthcare organizations should consider deploying complex models only if they perform no worse than statistically transparent ones in predicting outcomes, and they should look for benefits such as personalized explanations, improved workflow efficiency, or other advantages to justify their use. Using communication-related engagement measures as predictors (online activity, social participation, digital communication) could motivate additional attention to digital access and social isolation in service planning for older adults. We encourage future implementation studies to measure calibration, subgroup error, time on workflow, service utilization, number of referrals completed, and unintended disparities before concluding the organizational value of complex models.

Healthcare organizations should set guidelines based on transparent statistical models as benchmarks for complex AI. With performance nearly tied, organizations should rely on measures beyond standard metrics when deploying more complex algorithms. At a minimum, we

recommend evaluating calibration, subgroups, operational metrics such as workflow time, referral success, and service use, unintended disparities, and end-user understanding. A complex model that does not significantly outperform a statistical benchmark needs proof of incremental value, such as enabling personalized explanations, improving efficiency in the care pathway, and supporting a greater focus on proactive risk detection to inform better, targeted care. Cox and DeepSurv predictions provide risk information while also identifying areas where model outputs warrant cautious interpretation, thereby supporting a quality-control function.

Digital engagement supports healthcare decision-making. As telehealth services expand, and as patient portals and remote care grow, organizations need to integrate engagement measures (digital communication, social participation, online activity) into service design to provide adequate care for older adults who may experience social isolation or lack digital access. Instead of relying solely on a high C-index, organizations must monitor calibration, subgroup error, overrides, referral completion rate, time to response, service-use levels, and unintended disparities.

7. Theoretical and Research Implications

The findings align with a socio-technical perspective in which the value of predictive analytics depends on the relationship among technical performance, interpretability, organizational capabilities, workflow design, and governance. In this study, greater model complexity did not advance discrimination, raising the question of whether incremental analytical capabilities justify additional implementation, validation, explanation, and governance requirements. We therefore propose complexity value alignment as a conceptual proposition motivated by the empirical model comparison, rather than as a validated management theory. From this viewpoint, organizational adoption should be guided by demonstrated incremental utility rather than technical novelty alone. Future research will directly test this proposition in organizational settings by examining implementation costs and how model complexity and explanation formats affect managerial trust, decision consistency, workflow time, resource allocation, and accountability.

8. Limitations and Future Research

Identical five-fold cross-validation and comparative calibration strengthened the head-to-head model evaluation, but several limitations remain. DeepSurv exhibited greater between-fold variability than Cox PH, suggesting less stable discrimination across development subsets. In addition, sensitivity analysis excluding predictors with more than 60% missingness preserved Cox PH performance but reduced the DeepSurv C-index from 0.7833 to 0.7594, indicating that the neural model is more sensitive to removing these variables. Schoenfeld residual diagnostics identified nominal proportional-hazards violations for gender and social participation; therefore, the corresponding Cox coefficients should be interpreted as average associations over follow-up rather than strictly time-invariant effects. We assessed cross-model agreement in predictor importance using Top-5 and Top-10 overlap. These comparisons capture overlap among the most influential predictors but do not imply equivalence between Cox coefficients and SHAP values.

External validation is needed to assess the generalizability of these results. For example, future research should test Cox PH and DeepSurv using independent longitudinal samples from healthcare institutions that differ in demographics, geography, and care pathways. This should include evaluating discrimination, calibration, and subgroup analyses to determine whether the comparable results observed in the NHATS sample continue to hold under different conditions. Prospective implementation studies should further assess whether explainable survival models improve clinical utility, workflow efficiency, resource allocation, and organizational decision-making.

9. Conclusion

Penalized Cox PH and DeepSurv showed comparable discrimination in predicting time to first observed dementia. Cox PH also demonstrated greater stability across cross-validation folds and greater robustness when predictors with extensive missingness were excluded. Hence, DeepSurv's added complexity did not confer a clear predictive advantage in this analysis. For healthcare organizations, model selection should weigh predictive performance alongside interpretability, implementation requirements, validation, and monitoring. When a transparent model performs comparably to a more complex approach, additional complexity should be justified by demonstrated decision-support benefits. DeepSurv may offer complementary value through individualized explanations, flexible risk modeling, and scenario analysis, but prospective evaluation is needed to confirm these potential benefits. Future research should externally validate these findings and assess calibration, subgroup performance, and decision-support utility in independent healthcare settings.

Acknowledgment and Funding

The authors would like to acknowledge the support provided in part through Sub Award No. NSF00123-08 (NSF Award 2118285) and the Data Visualization and Extended Reality (DVXR) Lab at the College of Information, University of North Texas.

Ethics statement: This study involved secondary analysis of publicly available, de-identified data from the National Health and Aging Trends Study (NHATS) and did not involve direct participant recruitment, contact, or author intervention. We collected no new human-participant data for this analysis.

Conflict of interest: The authors declare no conflicts of interest relevant to this study.

Data availability: NHATS public-use data are available through the National Health and Aging Trends Study subject to its applicable registration and data-use requirements.

Author contributions: Ashrafe Alam: Conceptualization, Methodology, Formal Analysis, Investigation, Data Curation, Visualization, Writing – Original Draft. Md Golam Rabbani: Data Curation, Validation. Gahangir Hossain: Validation, Supervision, Writing – Review & Editing.

Sharad Sharma: Supervision, Writing – Review and Editing. All authors reviewed and approved the final manuscript.

References

- Akiba, T., Sano, S., Yanase, T., Ohta, T., & Koyama, M. (2019). Optuna: A next-generation hyperparameter optimization framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining* (pp. 2623–2631). <https://doi.org/10.1145/3292500.3330701>.
- Alam, A., Rabbani, M. G., & Prybutok, V. R. (2026). AI-enabled dementia risk prediction for smart and sustainable healthcare: An interpretable machine learning study using NHATS. *Applied Sciences*, 16(5), 2180. <https://doi.org/10.3390/app16052180>
- Ankuda, C. K., Freedman, V. A., Covinsky, K. E., & Kelley, A. S. (2021). Population-based screening for functional disability in older adults. *Innovation in Aging*, 5(1), igaa065. <https://doi.org/10.1093/geroni/igaa065>
- Antolini, L., Boracchi, P., & Biganzoli, E. (2005). A time-dependent discrimination index for survival data. *Statistics in Medicine*, 24(24), 3927–3944. <https://doi.org/10.1002/sim.2427>
- Bhadauria, U., Kumari, A., & Bhadauriya, S. (2026). Artificial intelligence and financial decision making in Indian banks: Adoption, effectiveness, and challenges. *International Journal of Business and Management*, 5(1), 969–995. <https://doi.org/10.56879/ijbm.v5i1.48>
- Brain, J., Kafadar, A. H., Errington, L., Kirkley, R., Tang, E. Y. H., Akyea, R. K., Bains, M., Brayne, C., Figueredo, G., Greene, L., Louise, J., Morgan, C., Pakpahan, E., Reeves, D., Robinson, L., Salter, A., Siervo, M., Tully, P. J., Turnbull, D., . . . Stephan, B. C. M. (2024). What's new in dementia risk prediction modelling? An updated systematic review. *Dementia and Geriatric Cognitive Disorders Extra*, 14(1), 49–74. <https://doi.org/10.1159/000539744>.
- Camacho, M., Atehortúa, A., Wilkinson, T., Gkontra, P., & Lekadir, K. (2025). Low-cost predictive models of dementia risk using machine learning and exposome predictors. *Health and Technology*, 15(2), 355–365. <https://doi.org/10.1007/s12553-024-00937-5>.
- Collins, G. S., Moons, K. G. M., Dhiman, P., Riley, R. D., Beam, A. L., Van Calster, B., Ghassemi, M., Liu, X., Reitsma, J. B., van Smeden, M., Boulesteix, A.-L., Camaradou, J. C., Celi, L. A., Denaxas, S., Denniston, A. K., Glocker, B., Golub, R. M., Harvey, H., Heinze, G., . . . Logullo, P. (2024). TRIPOD+AI statement: Updated guidance for reporting clinical prediction models that use regression or machine learning methods. *BMJ*, 385, e078378. <https://doi.org/10.1136/bmj-2023-078378>.



- Cox, D. R. (1972). Regression models and life-tables. *Journal of the Royal Statistical Society: Series B*, 34(2), 187-202. <https://doi.org/10.1111/j.2517-6161.1972.tb00899.x>.
- Deardorff, W. J., Barnes, D. E., Jeon, S. Y., Boscardin, W. J., Langa, K. M., Covinsky, K. E., Mitchell, S. L., Whitlock, E. L., & Smith, A. K. (2022). Development and external validation of a mortality prediction model for community-dwelling older adults with dementia. *JAMA Internal Medicine*, 182(11), 1161–1170. <https://doi.org/10.1001/jamainternmed.2022.4326>.
- Drabo, E. F., Wolff, J. L., Chyr, L. C., Zissimopoulos, J., & Lau, B. (2025). Subjective cognitive impairment (SCI) and future dementia risk in the National Health and Aging Trends Study (NHATS) during 2012–2019. *Journal of Aging and Health*, 37(3–4 Suppl), 76S–90S. <https://doi.org/10.1177/08982643241308450>.
- Efron, B. (1979). Bootstrap methods: Another look at the jackknife. *The Annals of Statistics*, 7(1), 1–26. <https://doi.org/10.1214/aos/1176344552>.
- Floud, S., Balkwill, A., Sweetland, S., Brown, A., Reus, E. M., Hofman, A., Blacker, D., Kivimäki, M., Green, J., Peto, R., Reeves, G. K., & Beral, V. (2021). Cognitive and social activities and long-term dementia risk: The prospective UK Million Women Study. *The Lancet Public Health*, 6(2), e116–e123. [https://doi.org/10.1016/S2468-2667\(20\)30284-X](https://doi.org/10.1016/S2468-2667(20)30284-X).
- Freedman, V. A., & Kasper, J. D. (2019). Cohort profile: The National Health and Aging Trends Study (NHATS). *International Journal of Epidemiology*, 48(4), 1044–1045g. <https://doi.org/10.1093/ije/dyz109>.
- Graf, E., Schmoor, C., Sauerbrei, W., & Schumacher, M. (1999). Assessment and comparison of prognostic classification schemes for survival data. *Statistics in Medicine*, 18(17–18), 2529–2545.
- Harrell, F. E., Jr. (2015). *Regression modeling strategies: With applications to linear models, logistic and ordinal regression, and survival analysis* (2nd ed.). Springer. <https://doi.org/10.1007/978-3-319-19425-7>.
- Jenne, T. (2026). More than a tool: How organizationally embedded AI competence enables sustainable competitive advantages in media companies. *International Journal of Business and Management*, 5(1), 186–202. <https://doi.org/10.56879/ijbm.v5i1.11>
- Katz, S., Ford, A. B., Moskowitz, R. W., Jackson, B. A., & Jaffe, M. W. (1963). Studies of illness in the aged: The index of ADL: A standardized measure of biological and psychosocial function. *JAMA*, 185(12), 914–919. <https://doi.org/10.1001/jama.1963.03060120024016>.
- Katzman, J. L., Shaham, U., Cloninger, A., Bates, J., Jiang, T., & Kluger, Y. (2018). DeepSurv: Personalized treatment recommender system using a Cox proportional hazards deep neural

- network. BMC Medical Research Methodology, 18(1), 24. <https://doi.org/10.1186/s12874-018-0482-1>.
- Kraus, M., Weiler, N., Oelke, D., Kehrer, J., Keim, D. A., & Fuchs, J. (2022). Immersive analytics with abstract 3D visualizations: A survey. Computer Graphics Forum, 41(1), 201–229. <https://doi.org/10.1111/cgf.14430>.
- Kroenke, K., Spitzer, R. L., & Williams, J. B. W. (2003). The Patient Health Questionnaire-2: Validity of a two-item depression screener. Medical Care, 41(11), 1284–1292. <https://doi.org/10.1097/01.MLR.0000093487.78664.3C>
- Kvamme, H., Borgan, Ø., & Scheel, I. (2019). Time-to-event prediction with neural networks and Cox regression. Journal of Machine Learning Research, 20(129), 1–30.
- Lawton, M. P., & Brody, E. M. (1969). Assessment of older people: Self-maintaining and instrumental activities of daily living. The Gerontologist, 9(3 Part 1), 179–186. https://doi.org/10.1093/geront/9.3_Part_1.179
- Lee, C., Zame, W. R., Yoon, J., & van der Schaar, M. (2018). DeepHit: A deep learning approach to survival analysis with competing risks. Proceedings of the AAAI Conference on Artificial Intelligence, 32(1), 2314–2321. <https://doi.org/10.1609/aaai.v32i1.11842>
- Livingston, G., Huntley, J., Sommerlad, A., Ames, D., Ballard, C., Banerjee, S., Brayne, C., Burns, A., Cohen-Mansfield, J., Cooper, C., Costafreda, S. G., Dias, A., Fox, N., Gitlin, L. N., Howard, R., Kales, H. C., Kivimäki, M., Larson, E. B., Ogunniyi, A., . . . Mukadam, N. (2020). Dementia prevention, intervention, and care: 2020 report of the Lancet Commission. The Lancet, 396(10248), 413–446. [https://doi.org/10.1016/S0140-6736\(20\)30367-6](https://doi.org/10.1016/S0140-6736(20)30367-6)
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 4765–4774.
- National Health and Aging Trends Study (NHATS). (2024). National Health and Aging Trends Study: Rounds 1–14 public use files. Johns Hopkins University Bloomberg School of Public Health.
- Nichols, E., Steinmetz, J. D., Vollset, S. E., Fukutaki, K., Chalek, J., Abd-Allah, F., Abdoli, A., Abualhasan, A., Abu-Gharbieh, E., Akram, T. T., Al Hamad, H., Alahdab, F., Alanezi, F. M., Alipour, V., Almustanyir, S., Amu, H., Ansari, I., Arabloo, J., Ashraf, T., . . . Vos, T. (2022). Estimation of the global prevalence of dementia in 2019 and forecasted prevalence in 2050: An analysis for the Global Burden of Disease Study 2019. The Lancet Public Health, 7(2), e105–e125. [https://doi.org/10.1016/S2468-2667\(21\)00249-8](https://doi.org/10.1016/S2468-2667(21)00249-8).

- Obermeyer, Z., & Emanuel, E. J. (2016). Predicting the future—Big data, machine learning, and clinical medicine. *New England Journal of Medicine*, 375(13), 1216–1219. <https://doi.org/10.1056/NEJMp1606181>.
- Peng, W., Luo, Y., & Liu, M. (2024). Longitudinal association of digital health technology use with dementia in older adults. *Innovation in Aging*, 8(Supplement 1), 331. <https://doi.org/10.1093/geroni/igae098.1081>.
- Piolatto, M., Bianchi, F., Rota, M., Marengoni, A., Akbaritabar, A., & Squazzoni, F. (2022). The effect of social relationships on cognitive decline in older adults: An updated systematic review and meta-analysis of longitudinal cohort studies. *BMC Public Health*, 22, 278. <https://doi.org/10.1186/s12889-022-12567-5>
- Rockwood, K., & Mitnitski, A. (2007). Frailty in relation to the accumulation of deficits. *The Journals of Gerontology: Series A*, 62(7), 722–727. <https://doi.org/10.1093/gerona/62.7.722>.
- Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63(3), 581–592. <https://doi.org/10.1093/biomet/63.3.581>.
- Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5), 206–215. <https://doi.org/10.1038/s42256-019-0048-x>.
- Stensrud, M. J., & Hernán, M. A. (2020). Why test for proportional hazards? *JAMA*, 323(14), 1401–1402. <https://doi.org/10.1001/jama.2020.1267>.
- Wiegerebe, S., Kopper, P., Sonabend, R., Bischl, B., & Bender, A. (2024). Deep learning for survival analysis: A review. *Artificial Intelligence Review*, 57(3), 65. <https://doi.org/10.1007/s10462-023-10681-3>.
- Yeung, A. W. K., Tosevska, A., Klager, E., Eibensteiner, F., Laxar, D., Stoyanov, J., ... & Willschke, H. (2021). Virtual and augmented reality applications in medicine: analysis of the scientific literature. *Journal of medical internet research*, 23(2), e25499. <https://doi.org/10.2196/25499>.
- Yu, J., Huang, W., & Thorpe, R. J., Jr. (2026). Digital technology use and dementia among older adults: The moderating role of activity limitations. *The Journals of Gerontology: Series B*, 81(1), gbaf231. <https://doi.org/10.1093/geronb/gbaf231>.