Dynamic Prediction in Clinical Survival Analysis using Temporal Convolutional Networks

Daniel Jarrett, Jinsung Yoon, *Member, IEEE,* and Mihaela van der Schaar, *Fellow, IEEE*

***Abstract*—Accurate prediction of disease trajectories is critical for early identification and timely treatment of patients at risk. Conventional methods in survival analysis are often constrained by strong parametric assumptions and limited in their ability to learn from high-dimensional data. This paper develops a novel convolutional approach that addresses the drawbacks of both traditional statistical approaches as well as recent neural network models for survival. We present MATCH-Net: a Missingness-Aware Temporal Convolutional Hitting-time Network, designed to capture temporal dependencies and heterogeneous interactions in covariate trajectories and patterns of missingness. To the best of our knowledge, this is the first investigation of temporal convolutions in the context of dynamic prediction for personalized risk prognosis. Using real-world data from the Alzheimer's Disease Neuroimaging Initiative (ADNI), we demonstrate state-of-the-art performance without making any assumptions regarding underlying longitudinal or time-to-event processes—attesting to the model's potential utility in clinical decision support.**

***Index Terms*—Alzheimer's Disease Neuroimaging Initiative, dynamic prediction, survival analysis, temporal convolutions.**

I. INTRODUCTION

RECENT advances in data-driven healthcare have enhanced such various medical domains as disease identification, personalized treatment, epidemic prediction, and drug discovery. With the explosion of comprehensive and systematically collected electronic health record data, judicious application of machine learning methods have the potential to pave the way for more productive healthcare interactions and more intelligent screening, interventions, treatments, and clinical trial design, contributing to improvements in patient outcomes.

Clinical survival analysis is the study of time-to-event data, modeling the expected duration of time until clinical events occur—such as the onset of a disease, relapse of a condition, development of adverse reactions, and death. Accurate prediction of patient trajectories is critical for the early identification and timely treatment of individuals at risk. In Alzheimer's disease—the annual cost of which exceeds $800 billion globally [2]—the effectiveness of therapeutic treatments is often limited by the challenge of identifying patients at early enough stages of disease progression for treatments to be of potential use. As a result, accurate and personalized prognosis during earlier stages of cognitive decline is critical for effective intervention and subject selection in clinical trials.

Traditional statistical methods have often approached the survival problem by first choosing explicit functions to model the underlying stochastic processes, then using available data to estimate unknown parameters of the model [3]–[8]. However, this means that conventional models are often tightly coupled with their specific assumptions, such as linearity and proportionality in the case of the popular Cox model [9], [10]—constraints that may not be valid or verifiable in practice. Neural networks offer versatile alternatives by virtue of their capacity as general-purpose function approximation machines, capable of learning—without restrictive assumptions—the complex latent structure between an individual's prognostic factors and odds of survival. At the same time, the use of specialized architectures means that prior knowledge can still be flexibly incorporated into models to guide learning.

This investigation focuses on deep learning for survival prediction, using Alzheimer's disease as a case study for experimental validation. In particular, in light of recent evidence of the competitiveness of generic convolutional architectures for sequence processing, especially in comparison to recurrent models [11], we analyze and illustrate the effectiveness of *temporal convolutions* for survival prediction in the presence of time-dependent covariates. While much research has been devoted to studying patterns of risk in Alzheimer's disease, a conclusive understanding of disease progression remains elusive, owing to heterogeneous biological pathways [12], [13], complex temporal patterns [14], [15], and diverse interactions [16], [17]. Hence Alzheimer's data is a prime venue for leveraging the potential advantages of deep convolutional networks over conventional statistical techniques in modeling temporal patterns and issuing risk predictions—helping physicians estimate both the likelihood of dementia as well as the expected rate of progression for individual patients.

Contributions. Our goal is to establish a novel convolutional model for survival prediction in the longitudinal setting, using Alzheimer's disease as a case study for experimental validation. Primary contributions are threefold: First, we formulate a *generalized framework* for the task of longitudinal survival prediction, laying the foundation for effective cross-model performance comparison. Second, our proposed architecture is

This work was supported by Alzheimer's Research UK, the Office of Naval Research (ONR), and the NSF (Grant number: ECCS1462245, ECCS1533983, and ECCS1407712). Note that a preliminary version of this work was previously presented at the NeurIPS Workshop on Machine Learning for Health (2018), with a draft of results available at http://arxiv.org/abs/1811.10746 [1].

D. Jarrett is with the Department of Engineering Science, University of Oxford, Oxford OX1 3PJ UK (email: daniel.jarrett@eng.ox.ac.uk).

J. Yoon is with the Department of Electrical and Computer Engineering, UCLA, Los Angeles, CA 90095 USA (email: jsyoon0823@g.ucla.edu).

M. van der Schaar is with the Department of Engineering Science, University of Oxford, Oxford OX1 3PJ UK, the Alan Turing Institute, London NW1 2DB UK, and the Department of Electrical and Computer Engineering, UCLA, Los Angeles, CA 90095 USA (e-mail: mihaela.vanderschaar@eng.ox.ac.uk).

Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.usc.edu). As such, the ADNI investigators contributed to the design and implementation of ADNI and/or provided data, but did not participate in the analysis or writing of this report. A complete listing of ADNI investigators can be found here.

uniquely designed to capitalize on longitudinal data to issue *dynamically updated* survival predictions, as well as accommodating potentially informative patterns of *data missingness*. Third, we propose a method to visualize influential variables for insight into the model's predictions, along with a medical application. Finally, we demonstrate state-of-the-art results in comparison with a comprehensive suite of benchmarks.

In what follows, Section II analyzes recent work in deep learning for survival analysis, contextualizing the main contribution of our proposed method. Sections III and IV formalize the survival prediction task and provide a technical description of the MATCH-Net architecture. Sections V and VI put our proposal and novelties to the test via performance comparisons with a variety of statistical and deep learning benchmarks, demonstrating that our method achieves superior results on real-world Alzheimer's data. Finally, Sections VII and VIII conclude the study and illustrate a potential practical use case for MATCH-Net in the context of clinical decision support.

II. Related Work

The first study to investigate neural networks formally in the context of time-to-event analysis was done by [18]. By swapping out the linear functional in the standard Cox model for the topology of a hidden layer, their maximum-likelihood approach generalized the hazard function to accommodate nonlinear relationships with covariates, and was readily extendible to other models for censored data, such as [19] and [20]. Specifically, the Cox model assumes that the effect of covariates $\mathbf{x}$ is to increase or decrease the hazard λ by a time-invariant *proportionate* amount; thus the hazard is given by

$$\lambda(t, \mathbf{x}) = \lambda_0(t) e^{\phi^\top \mathbf{x}} \tag{1}$$

for some baseline hazard function $\lambda_0(t)$, where ϕ is a vector of coefficients and t denotes time. Instead of assuming a strictly *linear* relationship $\phi^\top \mathbf{x}$, the proposal in [18] was to use neural networks to allow for a more flexible parameterization.

In 2016, [21] were the first to apply modern techniques in deep learning to survival analysis, in particular without prior feature selection or domain expertise. While previous studies following [18]'s model generally produced mixed results in relation to conventional statistical methods [22], [23], [21] demonstrated comparable or superior performance through the use of multilayer perceptrons. Employing such modern techniques as stochastic gradient descent, weight decay, batch normalization, dropout, and gradient clipping, they illustrated the advantage of the nonlinear proportional hazards approach through real and synthetic datasets with both linear and nonlinear risks. Their model was successfully adapted to alternative forms of input, such as unstructured medical images [24] and high throughput transcriptomics data [25].

Instead of predicting the hazard function as an intermediate objective, [26] first proposed—and [27] further developed—an alternative approach to predict survival directly for grouped time intervals—that is, by formulating the problem in a manner more akin to multi-label classification. In 2017, [28] combined the use of the Cox partial likelihood with the goal of predicting probabilities for pre-specified time intervals. Inspired by the work of [29] on multi-task logistic regression models for survival, they generalized the idea to deep learning via multilayer perceptrons with a multi-task framework. However, all aforementioned models still maintained the limiting assumption that hazard ratios are time-invariant. In particular, [26] explicitly constrained the weights between layers to safeguard the proportional hazards assumption; likewise, [28] constrained the penultimate layer to consist of a single bottleneck neuron with linear activation for estimating the hazard—assumed to capture all task-relevant aspects of the input data.

TABLE I
SUMMARY OF PRIMARY IMPROVEMENTS BY RELATED WORK USING NEURAL NETWORKS FOR SURVIVAL ANALYSIS

Model	Non-Linearity	Deep Learning	Direct-to-Probability	Time-Variance	Dynamic Prediction
Faraggi et al. [18]	✓	✗	✗	✗	✗
Katzman et al. [21]	✓	✓	✗	✗	✗
Luck et al. [28]	✓	✓	✓	✗	✗
Lee et al. [30]	✓	✓	✓	✓	✗
MATCH-Net	✓	✓	✓	✓	✓

Note that statistical methods have also been developed that relax the conditions of linearity and proportionality and enable dynamic predictions. The sliding landmarking [31] and joint modeling [32] approaches are the most popularly used techniques developed with these objectives. In Section VI, we describe and include the performance of both approaches as benchmarks in our experiments.

Recently, [30] proposed learning the distribution of survival times directly, making no assumptions regarding the underlying stochastic processes—in particular with respect to the time-invariance of hazards. Specifically, they proposed dispensing entirely with relying on the relationship

$$\mathbb{P}(T_{\text{surv}} > t) = e^{-\int_0^t \lambda(u) du} \tag{2}$$

where T_{surv} denotes the survival time. Instead, they opted to directly estimate the failure function $F(t|\mathbf{x}) = 1 - \mathbb{P}(T_{\text{surv}} > t)$, thereby dropping the proportionality assumption altogether. The end-to-end neural network parameterization of the failure function allowed smoothly handling the presence of competing risks, and demonstrated significant improvements over existing statistical, machine learning, and neural network survival models on multiple real and synthetic datasets. At the same time, all models so far had issued survival predictions using only information from a single time point; in the presence of panel data with time-dependent covariates, this approach may potentially give up valuable temporal information.

Within the context of Alzheimer's disease, [33] studied the use of longitudinal medical image sequences for classifying stable and progressive patients. While their graph-based approach explicitly accounted for the temporal aspect of the input, their outputs were binary estimates with no temporal dimension. Specifically for modeling survival, [34] investigated the use of recurrent neural networks in forecasting Alzheimer's disease trajectories with longitudinal data. By generalizing the joint modeling framework [32] to deep learning, they demonstrated improvements over conventional methods. However, they explicitly relied on the exponential distribution (*i.e.* the hazard function $\lambda_0(t)$ is assumed not to vary with time) for modeling survival, falling back on potentially restrictive assumptions.

Building on these developments, one of the main contributions of this study is the use of temporal convolutions for dynamic survival prediction. Importantly, this is done *without* making any assumptions whatsoever, leveraging the longitudinal aspect of inputs while allowing nonlinear associations between covariates and risks to evolve over time (see Table I).

Most recently, [35] and [36] both illustrate the advantages of deep learning over the Cox framework within the static setting, conducting applied analyses with prostate cancer and cervical cancer respectively. In the temporal setting, [37] demonstrated the use of recurrent networks in predicting survival from irregularly sampled data for cystic fibrosis, paying specific attention to competing risks. For additional context, we refer the interested reader to a recent survey of statistical and machine learning methods applicable to survival analysis [38].

Finally, there has recently been an increase in research specifically based on ADNI data. In the imaging domain, [39] and [40] develop novel methods for analyzing brain images for classifying disease states, while [41] and [42] demonstrate the benefit of transfer learning techniques in improving generative and discriminative performance using limited data. Focusing on the temporal dimension, [43] predict cognitive test scores at multiple future time points using baseline MRI features, while [44] and [45] are more related to our work in using longitudinal biomarkers and cognitive test data to predict the probability of dementia within predefined periods. While these methods focus on the goal of binary classification accuracy at a fixed threshold, our interest lies in modeling the evolving survival curve itself by dynamically predicting the *probabilities* of failure as new information comes in at each time step.

III. Problem Formulation

Let there be N patients in a study, indexed $i \in \{1, ..., N\}$. Each patient is associated with a sequence of longitudinal observations, and time is treated as a discrete dimension. Each longitudinal data point therefore consists of the tuple $(t, \mathbf{x}_{i,t}, s_{i,t})$, where $\mathbf{x}_{i,t}$ is the vector of observed covariates recorded at time step t, and $s_{i,t}$ is the binary survival indicator corresponding to the event of interest, such as death or the diagnosis of a condition. Per convention in survival literature, we assume that the censoring of observations is not correlated with the eventual survival outcomes of patients [46]–[49].

For patient i, let random variable $T_{i,\text{surv}}$ denote the time of event occurrence and $T_{i,\text{cens}}$ denote the time of right-censoring; then the time of last measurement is the random variable $T_i = \min\{T_{i,\text{surv}}, T_{i,\text{cens}}\}$. In the context of survival, the event is observed for a maximum of only one time, after which no further observations are recorded; that is, by construction $s_{i,t} = 1$ only where $t = t_i$. In addition, due to the right-censoring of patient trajectories, final event occurrences may not always be observed; trajectories for such patients correspond to sequences $\langle (t, \mathbf{x}_{i,t}, s_{i,t}) \rangle_{t=1}^{t_i}$ for which all values $s_{i,t} = 0$.

Now, let the complete longitudinal survival dataset be given by $\{\langle (t, \mathbf{x}_{i,t}, s_{i,t}) \rangle_{t=1}^{t_i}\}_{i=1}^{N}$. Since our objective is to predict survival, we restrict our attention to input sequences in which the event of interest has not yet occurred (*i.e.* where $s_{i,t} = 0$ for all t in the input sequence). After all, it is not particularly useful, for instance, to predict the future probability of death for a patient who has already died. Then we can define

$$\mathbf{X}_{i,t,w} = \langle \mathbf{x}_{i,t'} \rangle_{t' \in T} \quad \text{where} \quad T = \{t' : t - w \leq t' \leq t\} \tag{3}$$

to be the set of observations for patient i extending from time t into a width-w window of the past, where hyperparameter w depends on the model under consideration. Again, the survival indicators are implicit as $s_{i,t} = 0$ for all t. Given longitudinal measurements in $\mathbf{X}_{i,t,w}$, our task is to issue risk predictions corresponding to length-τ horizons into the future. Formally, given a *backward-looking* historical window $(t - w, t]$, we are interested in the failure function for *forward-looking* prediction intervals $(t, t + \tau]$; that is, we want to estimate the probability

$$F_i(t + \tau | t, w) = \mathbb{P}(T_{i,\text{surv}} \leq t + \tau | T_{i,\text{surv}} > t, \mathbf{X}_{i,t,w}) \tag{4}$$

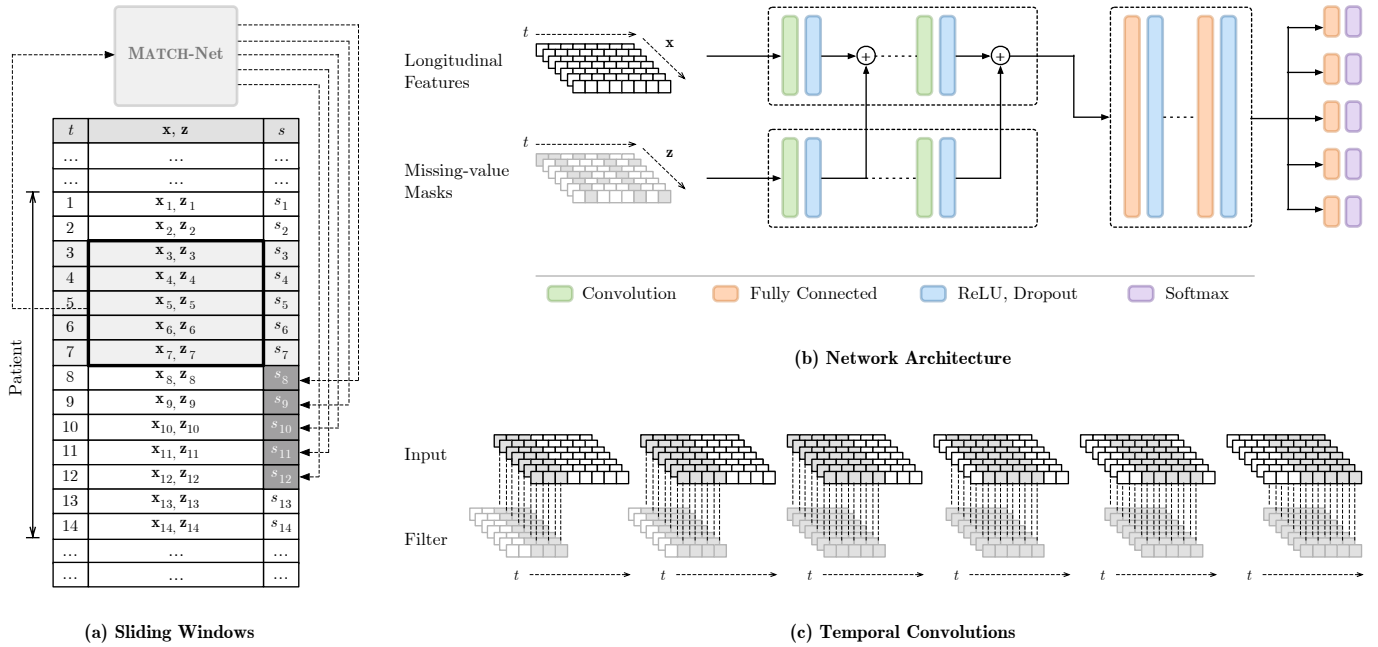
of event occurrence within each prediction interval. Naturally, the true distribution of survival times cannot be known on the basis of a finite dataset; our objective is therefore to obtain estimates of the true probability. In other words, we want to minimize $\mathcal{L}(\hat{F}_i(t + \tau | t, w), s_{i,t+\tau})$ over estimates $\hat{F}$, where $\mathcal{L}$ is some appropriate measure of loss in relation to the model's estimated failure function and the empirical distribution of survival times. Observe that parameterizing the width of the historical window results in a generalized framework. For instance, a Cox landmarking approach would typically utilize the most recent set of measurements; that is, $w = 1$. At the other end of the spectrum, recurrent network models may consume the entire history of measurements since the beginning; that is, $w = t$. As we shall see, the best performance is in fact obtained via a flexible intermediate approach—that is, by incorporating information from a sliding window of history, and allowing the optimal width of the window to be selected as a hyperparameter. See Section IV-A for details of architecture, and see Appendix E in the supplementary material for a list of hyperparameters.

IV. Match-Net

We propose MATCH-Net: a Missingness-Aware Temporal Convolutional Hitting-time Network for survival prediction, innovating on current approaches in two main respects:

- **Temporal Convolutions**: Existing deep learning models issue prognoses on the basis of information from a single time point [18]–[30]. With the increasing availability of longitudinal survival data, this approach discards potentially valuable information. We investigate the use of temporal convolutions in capturing explicit representations of covariate trajectories, in order to make full use of historical information in issuing dynamic predictions.
- **Informative Missingness**: Current survival methods rely on the common assumption that the timing and frequency of covariate measurements is uninformative [31], [50]. By contrast, our model explicitly accounts for informative missingness by learning correlations between patterns of data missingness and disease progression.

Among other design choices, each innovation is a source of gain in performance; see Section VI for a detailed account.



(a) Sliding Windows

(b) Network Architecture

(c) Temporal Convolutions

Fig. 1. (a) Longitudinal survival data and the sliding-window input mechanism. In this example, as of time $t = 7$, MATCH-Net takes as input a width-$w = 5$ window of historical information (encompassing times $t = 3$ through 7), and issues survival predictions for $t = 8$ through 12 (that is, for $\tau = 1$ through 5). As the input window slides forward through time, targets for survival predictions are similarly shifted forward. (b) The MATCH-Net architecture, shown here with $\tau_{\max} = 5$. Input covariates and binary masks are processed by parallel convolutional streams, with filter activations from the auxiliary branch concatenated with the main branch after each layer. This is followed by fully-connected block that captures more global relationships between locally extracted features. The output block consists of a single fully-connected layer and softmax function that produces the output for each prediction horizon. (c) Illustration of a single temporal convolutional filter acting over an input window. Individual longitudinal features are represented as parallel channels within the sequence of inputs.

A. Network Architecture

MATCH-Net accepts as input a *sliding-window* of observed covariates $\mathbf{X}_{i,t,w}$, as well as a corresponding binary mask of missing-value indicators $\mathbf{Z}_{i,t,w}$ defined analogously,

$$\mathbf{Z}_{i,t,w} = \langle \mathbf{z}_{i,t'} \rangle_{t' \in T} \quad \text{where} \quad T = \{t' : t - w \leq t' \leq t\} \tag{5}$$

where $z^d_{i,t} = 1$ if and only if $x^d_{i,t}$ is missing, for any covariate d. Figure 1(a) illustrates the longitudinal context within which MATCH-Net operates, as well as the network's prediction targets in association with the sliding window mechanism.

Figure 1(b) shows the core MATCH-Net architecture. Starting from the base of the network, the convolutional block first learns representations of longitudinal covariate trajectories by extracting local features from temporal patterns in the data. Indicator masks are processed in a parallel stream, and filter activations from the auxiliary branch are concatenated with those in the main branch after each layer. The fully-connected block then captures more global relationships by combining local information extracted from the convolutional block. ReLUs are used for nonlinear activation after each layer, followed by MC dropout [51]. Finally, employing the multi-task approach of [28], [30], each prediction task in the output block is associated with a single fully-connected layer followed by the softmax function, producing the array of failure estimates

$$\hat{\mathbf{y}}_{i,t} = [\hat{F}_i(t+1|t,w), ..., \hat{F}_i(t+\tau_{\max}|t,w)] \tag{6}$$

for pre-specified prediction intervals, where $\tau_{\max}$ is the maximal prediction horizon desired; the sequence $\hat{\mathbf{y}}_{i,t}$ traces out the survival curve for patient i conditioned on survival until t.

This *convolutional dual-stream* architecture explicitly captures representations of temporal dependencies within each stream, as well as between covariate trajectories and missingness patterns in association with disease progression. This accounts for the potential informativeness of both *irregular* sampling (*i.e.* the intervals between consecutive clinical visits and measurements may vary) as well as *asynchronous* sampling (*i.e.* not all features are measured at the same time or at the same frequency) [52], [53]; for instance, a patient suspected of exhibiting progressive cognitive impairment might be more likely to be scheduled more frequent visits by the clinician for the purposes of repeated lab measurements and cognitive tests. In addition, the dual-stream architecture also encourages the network to distinguish between actual measurements and imputed values, thereby reducing its sensitivity to the specific imputation method chosen. Finally, since we expect the indicator tensors to contain less information than the covariate measurements themselves, the parallel-stream design—as opposed to simply encoding missingness via additional dummy feature channels—allows us to restrict the capacity of the auxiliary path by reducing the relative number of filters.

B. Loss Function

Let $\boldsymbol{\theta}$ denote the set of trainable parameters that characterize the survival network. Then the negative log-likelihood of a single empirical result $s_{i,t+\tau}$ and model estimate $\hat{F}_i(t+\tau|t,w)$ in association with some input window $\mathbf{X}_{i,t,w}$ is given by

$$\mathcal{L}_{i,t,\tau}(\boldsymbol{\theta}) = -[s_{i,t+\tau} \log \hat{F}_i(t+\tau|t,w) + (1 - s_{i,t+\tau}) \log(1 - \hat{F}_i(t+\tau|t,w))] \tag{7}$$

The total loss function is then computed to simultaneously take into account the quality of survival predictions for all desired prediction horizons τ, for all times t available along each patient's recorded longitudinal trajectory, and for all patients $i \in \{1, ..., N\}$ present in the survival dataset:

$$\mathcal{L}(\boldsymbol{\theta}) = \eta \cdot \sum_{i=1}^{N} \sum_{j=1}^{t_i} \sum_{k=1}^{\tau_i} \alpha(i,j,k) \cdot \mathcal{L}_{i,j,k} \quad \text{where} \quad \eta = \frac{1}{\sum_{i=1}^{N} \sum_{j=1}^{t_i} \tau_i} \tag{8}$$

where $\tau_i = \min\{t_i - t, \tau_{\text{max}}\}$ accounts for failure and right-censoring occurring prior to $t + \tau_{\text{max}}$. This is a natural generalization of the log-likelihood à la [26] to accommodate longitudinal survival. The weight function $\alpha(i,t,\tau)$ allows trading off the relative importance of different patients, time steps, and prediction horizons. First, this allows us to standardize patient contributions by setting $\alpha(i,t,\tau) \propto 1/t_i$, thereby counteracting the automatic bias against patients with shorter durations to failure or censoring. Second and more importantly in the context of heavily imbalanced classes, this allows up-weighting positive training instances—that is, input windows that correspond to eventual failure. Finally, any convex combination of losses across prediction horizons will be valid; here we simply take the unweighted sum across intervals.

C. Training Procedure

Training begins with the tuple of input data $\{\langle \mathbf{X}_{i,t,w} \rangle_{t=1}^{t_i}\}_{i=1}^{N}$ and $\{\langle \mathbf{Z}_{i,t,w} \rangle_{t=1}^{t_i}\}_{i=1}^{N}$, and terminates with a set of calibrated network weights $\boldsymbol{\theta}$. The network is trained until convergence, up to a maximum of 50 epochs. As described in Section VI, performance will be evaluated on the basis of the area under the receiver operating characteristic curve (AUROC), as well as the area under the precision-recall curve (AUPRC), and both metrics are computed as functions of the prediction horizon τ. Analogous to our definition for the total loss, the convergence metric is defined as the following weighted sum of performance scores across all prediction tasks (coefficients $\beta(\tau)$, $\gamma(\tau)$ optionally allow trading off the relative importance between the two measures and different horizons):

$$\mathcal{C} = \sum_{k=1}^{\tau_{\text{max}}} (\beta(k) \cdot \text{AUROC}_k + \gamma(k) \cdot \text{AUPRC}_k) \tag{9}$$

In this investigation, we simply take the unweighted sum across both dimensions, although any convex combination would be valid. Empirically, results are not meaningfully improved by favoring one metric over another. Validation performance is computed every 10 iterations. For early stopping, validation scores serve as proxies for the generalization error. See Appendix A in the supplementary material for detailed pseudocode of the MATCH-Net training algorithm.

In addition, current approaches such as [21], [30] take the default option of using weight decay regularization. In the presence of high-dimensional feature spaces, we employ elastic net regularization [54] to additionally leverage the feature selection effect of sparse coefficients. Formally, we compute the overall penalty as a convex combination of the L_1 and L_2 penalties from the lasso and ridge regularizers; that is,

$$\mathcal{R} = \rho \cdot \mathcal{R}_{\text{lasso}} + (1 - \rho) \cdot \mathcal{R}_{\text{ridge}} \tag{10}$$

where $\rho \in [0, 1]$ is treated as a model hyperparameter.

V. DATASET

The Alzheimer's Disease Neuroimaging Initiative (ADNI) study data is a longitudinal survival dataset of per-visit measurements for 1,737 patients [2]. The data tracks disease progression through clinical measurements at ½-year intervals, including quantitative biomarkers, cognitive tests, risk factors, as well as clinician diagnoses of patients' disease status. Our objective is to predict the *first stable diagnosis* of Alzheimer's disease for each patient. Further information on the study data can be found at https://tadpole.grand-challenge.org/Data/.

A. Details on Dataset

We are interested in the disease status for each patient at any given time. A clinical diagnosis is recorded at each patient's visit, and consists of two attributes. First, each diagnosis may be either stable or transitive. The former consists of stable diagnoses of normal brain functioning (NL), mild cognitive impairment (MCI), or Alzheimer's disease (AD), and the latter consists of diagnoses indicating transitions between these categories, which may take the form of either conversions or reversions. Conversions indicate probable forward progression in the disease trajectory, and reversions indicate probable regression back towards an earlier stage of the disease.

Patients are observed to remain in stable or transition states for various durations. The average patient who receives a transition diagnosis is observed to persist in that state for one year, while some patients do not exit this state until almost 5 years have elapsed. Patients who receive a transition diagnosis may not actually receive a subsequent stable diagnosis; in fact, less than half of the transition diagnoses for dementia were confirmed by a stable diagnosis at the next time step, and almost one quarter are never followed by a stable diagnosis at any point until eventual right-censoring. In addition, patients often actually undergo reversion transitions back towards earlier stages of the disease; in fact, over 5% of the study population receive reversion diagnoses at some point in time.

Event labels are defined as positive upon the first occurrence of stable diagnosis of Alzheimer's disease. Given the preliminary nature of transition diagnoses observed, we adopt the more conservative approach of relying on stable diagnoses to alleviate the problem of noisy labels. Note that this generates a far smaller number of positive training instances, resulting in a more difficult analysis. At the same time, it translates into a more clinically relevant prediction task that more faithfully models the underlying disease process, instead of simply learning any patterns of misdiagnosis present. The overall per-patient failure rate is 14% (243 patients out of the total 1,737 are eventually receive a stable diagnosis of Alzheimer's disease). However, given the online nature of the sliding window mechanism in training and testing, the effective fraction of observations with positive event labels for

TABLE II
SUMMARY AND DESCRIPTION OF VARIABLES USED IN ADNI DATASET.

		Type	Min	Max	Mean	S.D.	Missing
Event	Diagnosis of AD	Categorical	-	-	-	-	30.1%
Static	Age	Numeric	5.44E+01	9.14E+01	7.38E+01	7.20E+00	0.0%
	APOE4 (Risk)	Numeric	0.00E+00	2.00E+00	5.37E-01	6.56E-01	0.1%
	Education Level	Numeric	4.00E+00	2.00E+01	1.59E+01	2.86E+00	0.0%
	Ethnicity	Categorical	-	-	-	-	0.0%
	Gender	Categorical	-	-	-	-	0.0%
	Marital Status	Categorical	-	-	-	-	0.0%
	Race	Categorical	-	-	-	-	0.0%
Biomarker	Entorhinal	Numeric	1.04E+03	6.71E+03	3.44E+03	8.12E+02	49.2%
	Fusiform	Numeric	7.74E+03	3.00E+04	1.71E+04	2.82E+03	49.2%
	Hippocampus	Numeric	2.22E+03	1.12E+04	6.68E+03	1.24E+03	46.6%
	Intracranial	Numeric	2.92E+02	2.11E+06	1.53E+06	1.66E+05	37.6%
	Mid Temp	Numeric	8.04E+03	3.22E+04	1.92E+04	3.14E+03	49.2%
	Ventricles	Numeric	5.65E+03	1.63E+05	4.21E+04	2.33E+04	41.6%
	Whole Brain	Numeric	6.49E+05	1.49E+06	1.01E+06	1.12E+05	39.7%
Cognitive	ADAS (11-item)	Numeric	0.00E+00	7.00E+01	1.14E+01	8.63E+00	30.1%
	ADAS (13-item)	Numeric	0.00E+00	8.50E+01	1.75E+01	1.17E+01	30.7%
	CRD Sum of Boxes	Numeric	0.00E+00	1.80E+01	2.17E+00	2.81E+00	29.7%
	Mini Mental State	Numeric	0.00E+00	3.00E+01	2.66E+01	3.95E+00	29.9%
	RAVLT Forgetting	Numeric	-1.20E+01	1.50E+01	4.23E+00	2.53E+00	30.9%
	RAVLT Immediate	Numeric	0.00E+00	7.50E+01	3.45E+01	1.36E+01	30.7%
	RAVLT Learning	Numeric	-5.00E+00	1.40E+01	4.03E+00	2.81E+00	30.7%
	RAVLT Percent	Numeric	-5.00E+02	1.00E+02	5.97E+01	3.84E+01	31.4%

any prediction horizon is around 2%—by construction, only the final instances out of each patient's collection of observations is positive. Finally, note that left-truncated patients (20%) are omitted from the analysis, since survival cannot be defined for patients who are already registered positive at baseline.

B. Data Preparation

Since the ADNI dataset is an amalgamation of data from multiple related studies, most features are sparsely populated. Features with less than half of the entries missing are retained, leaving 18 numeric and 4 categorical features (see Table II); the latter are represented by one-hot encoding, resulting in 16 binary features. Consistent with existing Alzheimer's studies, patients are aligned according to time elapsed since baseline measurements [55]–[57]. Timestamps are discretized by mapping onto an axis with a fixed resolution of ½-year intervals; where multiple measurements qualify for the same destination, the most recent measurement per feature takes precedence. Since original measurements were already made at roughly ½-year intervals, we observe that the average absolute deviation between original and final timestamps amounts to an insignificant 4 days (*i.e.* less than 2% of each interval).

Where measurements are missing, values are reconstructed using zero-order hold interpolation. In addition, due to the fixed-width nature of the sliding window, the input tensor $\mathbf{X}_{i,t,w}$ for initial prediction times $t < w$ correspond to left-truncated information $t - w < 0$; feature values are therefore similarly extrapolated backwards with constant values for all intervals of the form $[-w, 0]$. Note that regardless of the imputation mechanism, information on original patterns of missingness—due to truncation, irregular sampling, and asynchronous sampling alike [52], [53]—is preserved in the missing-value mask $\mathbf{Z}_{i,t,w}$ provided in parallel to the network. Finally, to improve numerical conditioning, features for all patients and time steps are normalized with their empirical means and standard deviations from the training set data.

C. Class Imbalance

Recent work on deep learning for survival have largely performed experiments on relatively balanced data [21], [28], [30]. In stark contrast, as noted in Section V-A the ADNI data is characterized by an imbalance of 2%, posing substantive practical challenges for training and optimization [58]–[60]. In this study, we employ two techniques to counteract class imbalance. First, oversampling has been shown—especially in the context of convolutional architectures—to be more consistently effective than alternative methods, especially with binary class labels [61]. At the same time, it is known to potentially result in overfitting [62]. In this investigation, oversampling is applied to achieve a target ratio of positive to negative observations, with the ratio is treated as a model hyperparameter. Second, we employ label-forwarding to passively increase the frequency of positive event labels seen during training; that is, positive labels for any horizon $(t, t + \tau]$ are propagated forwards to all subsuming intervals $(t, t + \tau + k]$, for any positive integer k.

VI. EXPERIMENTS

To form a comprehensive basis for performance evaluation, we compare MATCH-Net against both traditional longitudinal methods in survival analysis, as well as an assortment of deep learning approaches—including those employed in the most recent studies. The former includes the Cox landmarking and joint modeling approaches, and the latter includes multilayer perceptrons and recurrent neural network models. As we shall see in Section VI-D, this allows us to examine the incremental gains in performance due to various design choices.

TABLE III
CROSS VALIDATION PERFORMANCE FOR MATCH-NET AND BENCHMARKS (BOLD VALUES INDICATE BEST PERFORMANCE) FOR $\tau_{\text{MAX}} = 5$

	τ	MATCH-Net	SW-TCN	SW-MLP	WaveNet†	FCN	D-Atlas	RNN	MLP	JM	LM
AUPRC	1	**0.594**	0.580	0.500	0.551	0.536	0.517	0.464*	0.469*	0.473*	0.469*
	2	**0.513**	0.505	0.447	0.476	0.453	0.423	0.410*	0.435	0.415*	0.412*
	3	**0.373**	0.367	0.354	0.363	0.357	0.364	0.340	0.340	0.319	0.325
	4	**0.390**	0.380	0.364	0.379	0.375	0.352	0.355	0.359	0.362	0.367
	5	**0.384**	0.381	0.371	0.374	0.365	0.360	0.365	0.356	0.366	0.363
AUROC	1	**0.962**	0.961	0.959	0.959	0.954	0.959	0.949*	0.948*	0.913*	0.909*
	2	**0.942**	0.941	0.932	0.939	0.930	0.929	0.930	0.930	0.917*	0.914*
	3	**0.902**	**0.902**	0.897	**0.902**	0.895	0.892	0.891	0.890	0.881	0.878
	4	**0.909**	0.908	0.904	0.908	0.903	0.896	0.901	0.895	0.894	0.890
	5	0.886	0.884	0.881	**0.888**	0.883	0.884	0.883	0.874	0.883	0.878

Left to right: Our proposed model (MATCH-Net) as well as its sliding window precursors, including sliding-window temporal convolutional networks (SW-TCN) and sliding-window multilayer perceptrons (SW-MLP). Performance benchmarks include generic adaptations of sequence models WaveNet (WaveNet†) and fully-convolutional networks (FCN) for the task of survival prediction, Disease Atlas (D-Atlas), recurrent networks (RNN), static multilayer perceptrons (MLP), as well as traditional survival methods including joint models (JM) and landmarking (LM). Under the RNN category, we experiment with vanilla RNNs, as well as models using either GRUs or LSTMs; due to the similarity of their underlying architectures, we only report results for the best-performing RNN specimen (which in this case is the vanilla RNN). The two-sample *t*-test for difference of means is conducted on the cross-validations results. Asterisks next to benchmark results indicate a statistically significant difference (p-value < 0.05) relative to the MATCH-Net result. A detailed breakdown of gains is found in Table V.

TABLE IV
PROPORTION OF RUNS (OUT OF TOTAL 25) WHERE MATCH-NET GIVES SUPERIOR RESULTS RELATIVE TO ALTERNATIVE ARCHITECTURES

	τ	SW-TCN	SW-MLP	WaveNet†	FCN	D-Atlas
AUPRC	1	68%	100%	80%	88%	100%
	2	76%	100%	92%	100%	100%
	3	84%	92%	88%	52%	100%
	4	80%	84%	64%	40%	100%
	5	52%	76%	76%	68%	100%
AUROC	1	84%	100%	84%	96%	100%
	2	68%	100%	68%	84%	100%
	3	68%	84%	44%	40%	92%
	4	64%	84%	76%	44%	100%
	5	56%	72%	20%	52%	80%

Performance is evaluated on the basis of AUROC and AUPRC, both computed with respect to each prediction horizon τ.

All benchmarks are structured to perform the same prediction tasks as MATCH-Net. For strict comparability, all deep learning models are optimized on the basis of the same loss function and regularization techniques, with training guided by identical convergence mechanisms and pipeline decisions such as oversampling and label forwarding. This allows us to subsequently isolate the gains from novelties in MATCH-Net.

A. Benchmarks

Cox Landmarking. To account for time-dependent covariates, we use the entry-exit implementation of the Cox model [63] and create separate records for each interval between measurements. In addition, we use the sliding landmarking approach to compute dynamic predictions of survival—that is, by basing predictions after each time step on information of all patients still alive just prior to that time [31]. Optimal groupings for the sequence of regression models are determined by exhaustive search in ½-year increments. Preliminary feature-selection is performed by stepwise regression using [64]. Consistent with literature, the time dimension is defined in terms of years since initial follow-up [34], [55]–[57].

Joint Modeling. Joint models have been shown to offer potential advantages in precision, especially by accounting for measurement error [65]. We adopt the common two-stage method described in [66]. First, linear mixed effects models are fit for longitudinal response variables (cubic B-splines are also considered); second, Cox models are fit as above, but using the mean estimates from the longitudinal sub-models. Given the predictive value of cognitive scores for the transition to Alzheimer's disease [67], candidates for response variables are chosen among the various test scores. As before, significant variables are first identified with [64]; the final set of responses is obtained by searching for the optimal combination.

Multilayer Perceptrons. First, we consider the performance of static multilayer perceptrons in the manner of [30]—that is, using only a single set of covariate values from the most recently available measurements. Second, we also consider the performance of dynamic models—that is, using a sliding window of history as input to the network, with a flattening operation prior to the first fully-connected layer.

Recurrent Architectures. First, we evaluate the performance of vanilla recurrent networks with sequence-to-vector architectures, where final states feed into a softmax output layers for survival predictions. In addition, the space of architectures searched over includes GRU and LSTM models. Furthermore, we compare against the exponentially parameterized joint modeling recurrent architecture Disease Atlas in [34].

Convolutional Sequence Models. Finally, as additional points for comparison, we consider generic sequence models with convolutional architectures. While not originally designed for use in survival analysis, we implement adaptations of fully-convolutional networks [68] and WaveNet [69] to enable survival prediction. For the latter, this importantly involves the manual addition of a fully-connected block on top of the causal convolutional residual-block layers in order to provide adequate longitudinal context for issuing risk predictions.

B. Experimental Setup

For all neural network models, hyperparameter optimization is carried out via 100 iterations of random search (see Appendix E in the supplementary material for a full list of hyperparameters and their selection ranges). Training loss is only computed

TABLE V
PRIMARY SOURCES OF GAIN (CROSS VALIDATION MEAN ± STANDARD DEVIATION; BOLD VALUES INDICATE BEST PERFORMANCE)

	τ	MLP		RNN	
AUPRC	1	**0.469**	(±**0.064**)	0.464	(±0.079)
	2	**0.435**	(±**0.056**)	0.410	(±0.060)
	3	**0.340**	(±**0.067**)	**0.340**	(±**0.067**)
	4	**0.359**	(±**0.065**)	0.355	(±0.068)
	5	0.356	(±0.085)	**0.365**	(±**0.087**)
AUROC	1	0.948	(±0.010)	**0.949**	(±**0.009**)
	2	**0.930**	(±**0.011**)	**0.930**	(±**0.012**)
	3	0.890	(±0.027)	**0.891**	(±**0.026**)
	4	0.895	(±0.029)	**0.901**	(±**0.025**)
	5	0.874	(±0.039)	**0.883**	(±**0.031**)

(a) Gain from covariate history. Recurrent neural networks (RNN) for time series prediction—including GRU and LSTM architectures—do not exhibit convincing performance improvements over static multilayer perceptrons (MLP).

	τ	RNN		SW-MLP	
AUPRC	1	0.464	(±0.079)	**0.500**	(±**0.066**)
	2	0.410	(±0.060)	**0.447**	(±**0.056**)
	3	0.340	(±0.067)	**0.354**	(±**0.061**)
	4	0.355	(±0.068)	**0.364**	(±**0.054**)
	5	0.365	(±0.087)	**0.371**	(±**0.084**)
AUROC	1	0.949	(±0.009)	**0.950**	(±**0.009**)
	2	0.930	(±0.012)	**0.932**	(±**0.012**)
	3	0.891	(±0.026)	**0.897**	(±**0.025**)
	4	0.901	(±0.025)	**0.904**	(±**0.026**)
	5	**0.883**	(±**0.031**)	0.881	(±0.035)

(b) Gain from limited window. The sliding-window multilayer perceptron (SW-MLP) allows selecting the optimal width of a sliding window of historical data. Improvements are more promising, boosting the average AUPRC by 4%.

	τ	SW-MLP		SW-TCN	
AUPRC	1	0.500	(±0.066)	**0.580**	(±**0.066**)
	2	0.447	(±0.056)	**0.505**	(±**0.065**)
	3	0.354	(±0.061)	**0.367**	(±**0.063**)
	4	0.364	(±0.054)	**0.380**	(±**0.052**)
	5	0.371	(±0.084)	**0.381**	(±**0.085**)
AUROC	1	0.950	(±0.009)	**0.961**	(±**0.005**)
	2	0.932	(±0.012)	**0.941**	(±**0.007**)
	3	0.897	(±0.025)	**0.902**	(±**0.025**)
	4	0.904	(±0.026)	**0.908**	(±**0.026**)
	5	0.881	(±0.035)	**0.884**	(±**0.032**)

(c) Gain from temporal convolutions. Sliding-window temporal convolutional networks (SW-TCN) include the use of convolutions over time, better capturing the temporal nature of historical information; AUPRC further improves by 9%.

	τ	SW-TCN		MATCH-Net	
AUPRC	1	0.580	(±0.066)	**0.594**	(±**0.058**)
	2	0.505	(±0.065)	**0.513**	(±**0.059**)
	3	0.367	(±0.063)	**0.373**	(±**0.065**)
	4	0.380	(±0.052)	**0.390**	(±**0.059**)
	5	0.381	(±0.085)	**0.384**	(±**0.081**)
AUROC	1	0.961	(±0.005)	**0.962**	(±**0.004**)
	2	0.941	(±0.007)	**0.942**	(±**0.007**)
	3	**0.902**	(±**0.025**)	**0.902**	(±**0.024**)
	4	0.908	(±0.026)	**0.909**	(±**0.027**)
	5	0.884	(±0.032)	**0.886**	(±**0.033**)

(d) Gain from missingness-awareness. By incorporating a dual-stream architecture to incorporate potentially informative patterns of missingness, MATCH-Net displays added AUPRC gains of 2% on average over the single stream model.

for event labels corresponding to actual recorded clinical visits (*i.e.* timestamps with recorded covariate values) in relation to actual recorded patient states (*i.e.* neither imputed nor forward-filled labels are included). While forward-propagated labels are used during training (see Section V-C), they are excluded from calculations for the purposes of validation and testing. Model selection is performed on the basis of final composite scores—as defined in Equation 9—for each candidate model. We employ stratified five-fold cross validation to evaluate model performance, with the set of patients randomly selected into datasets for training (60%), validation (20%), and testing (20%).

C. Results

Average Performance. Performance metrics are reported in Table III for prediction horizons up to $\tau_{\max} = 5$ (with time steps of ½ years). MATCH-Net results are shown in relation to its sliding-window precursors, including sliding window temporal convolutional networks (SW-TCN) and sliding-window multilayer perceptrons (SW-MLP). Also shown are conventional statistical benchmarks for survival, including Cox landmarking (LM) and joint modeling (JM). Deep learning survival benchmarks include static multilayer perceptrons (MLP), recurrent neural networks (RNN) including GRUs and LSTMs, Disease Atlas (D-Atlas), as well as generic sequence-to-sequence models adapted for survival prediction, including fully-convolutional networks (FCN) and a modification of WaveNet with an additional fully-connected block to provide global context (WaveNet†). Bold values indicate best performance, and asterisks on benchmark results indicate statistically significant difference with MATCH-Net at the 5% level (using the two-sample t-test for difference of means).

Proportion of Outperformance. While Table III provides the values of each performance metric *averaged* across all train-test splits, Table IV additionally provides the *proportion* of runs for which the MATCH-Net architecture performs better than the closest alternatives. Specifically, for each of the 5 random train-test splits, we run the model a total of 5 times, producing an overall total of 25 runs, out of which we compute the proportion for which MATCH-Net achieves a higher performance score than each alternative. We expect that this proportion be generally higher than 50% if the proposed model is more suitable than other methods. In fact, we observe that the proportion of runs for which MATCH-Net outperforms is fairly consistently in the majority, in particular for the more sensitive metric of AUPRC scores.[1] This provides an alternative argument as to how the proposed method is more performant: Should we choose to model the data with MATCH-Net, we know the results are better than if we had chosen an alternative method—much more often than not.

D. Sources of Gain

MATCH-Net produces state-of-the-art results, consistently outperforming both conventional statistical and neural network benchmarks. Gains are especially apparent in AUPRC

[1]Note that the AUROC score is much less sensitive than the AUPRC score in the context of highly imbalanced classes such as the ADNI dataset [70].

TABLE VI
CROSS VALIDATION PERFORMANCE FOR DIFFERENT WAYS OF HANDLING MISSING VALUES (BOLD VALUES INDICATE BEST PERFORMANCE)

	τ	MATCH-Net	Concatenated	Zeroed Out	Not Included
AUPRC	1	**0.594**	0.587	0.363*	0.580
	2	**0.513**	0.503	0.232*	0.505
	3	**0.373**	0.365	0.276*	0.367
	4	**0.390**	0.378	0.216*	0.380
	5	**0.384**	0.376	0.265*	0.381
AUROC	1	0.962	**0.963**	0.851*	0.961
	2	**0.942**	0.941	0.763*	0.941
	3	**0.902**	0.901	0.795*	**0.902**
	4	**0.909**	0.900	0.732*	0.908
	5	**0.886**	0.878	0.797*	0.884

(a) Average performance metrics across train-test splits for each method.

	τ	Concatenated	Zeroed Out	Not Included
AUPRC	1	88%	100%	68%
	2	56%	100%	76%
	3	52%	100%	84%
	4	80%	100%	80%
	5	68%	100%	52%
AUROC	1	56%	100%	84%
	2	72%	100%	68%
	3	44%	100%	68%
	4	68%	100%	64%
	5	60%	100%	56%

(b) Proportion of runs (out of total 25) where MATCH-Net outperforms.

Left to right: Our proposed dual-stream architecture (MATCH-Net); an alternative that concatenates missingness masks as additional channels of input (Concatenated); an alternative that leaves missing values zeroed out (Zeroed Out); a baseline that simply ignores missingness information (Not Included).

TABLE VII
MISCELLANEOUS SOURCES OF GAIN (CROSS VALIDATION MEAN $\pm$ STANDARD DEVIATION; BOLD VALUES INDICATE BEST PERFORMANCE)

Over-Sampling	Label Forwarding	Elastic Net Regularizer	AUPRC					AUROC				
			$\tau=1$	$\tau=2$	$\tau=3$	$\tau=4$	$\tau=5$	$\tau=1$	$\tau=2$	$\tau=3$	$\tau=4$	$\tau=5$
✗	✗	✗	0.493	0.418	0.348	0.351	0.355	0.946	0.923	0.895	0.902	**0.888**
✗	✗	✓	0.534	0.477	0.368	0.381	**0.384**	0.956	0.938	0.901	0.906	0.886
✗	✓	✗	0.435	0.413	0.334	0.351	0.357	0.944	0.920	0.888	0.896	0.886
✗	✓	✓	0.575	0.502	0.370	0.384	0.377	0.961	0.941	**0.902**	0.908	0.887
✓	✗	✗	0.499	0.425	0.352	0.365	0.365	0.946	0.923	0.894	0.902	0.887
✓	✗	✓	0.533	0.473	0.362	0.379	0.385	0.956	0.938	0.901	0.907	0.887
✓	✓	✗	0.448	0.404	0.332	0.348	0.361	0.942	0.918	0.885	0.894	0.881
✓	✓	✓	**0.594**	**0.513**	**0.373**	**0.390**	**0.384**	**0.962**	**0.942**	**0.902**	**0.909**	0.886

Gain from oversampling, label forwarding, and elastic net regularization (techniques applied across all models). Results are shown for MATCH-Net, and each row gives the results of the model with some, none, or all of the techniques applied. The left-hand columns indicate the specific combinations of techniques.

scores—improving on the MLP by an average of 15% and on joint models by 16% across all horizons, and by 27% and 26% for one-step-ahead predictions. While the advantage of using multilayer perceptrons over traditional statistical survival models has been studied (see, *e.g.*, [30]), we now account for the additional sources of gain from each design choice.

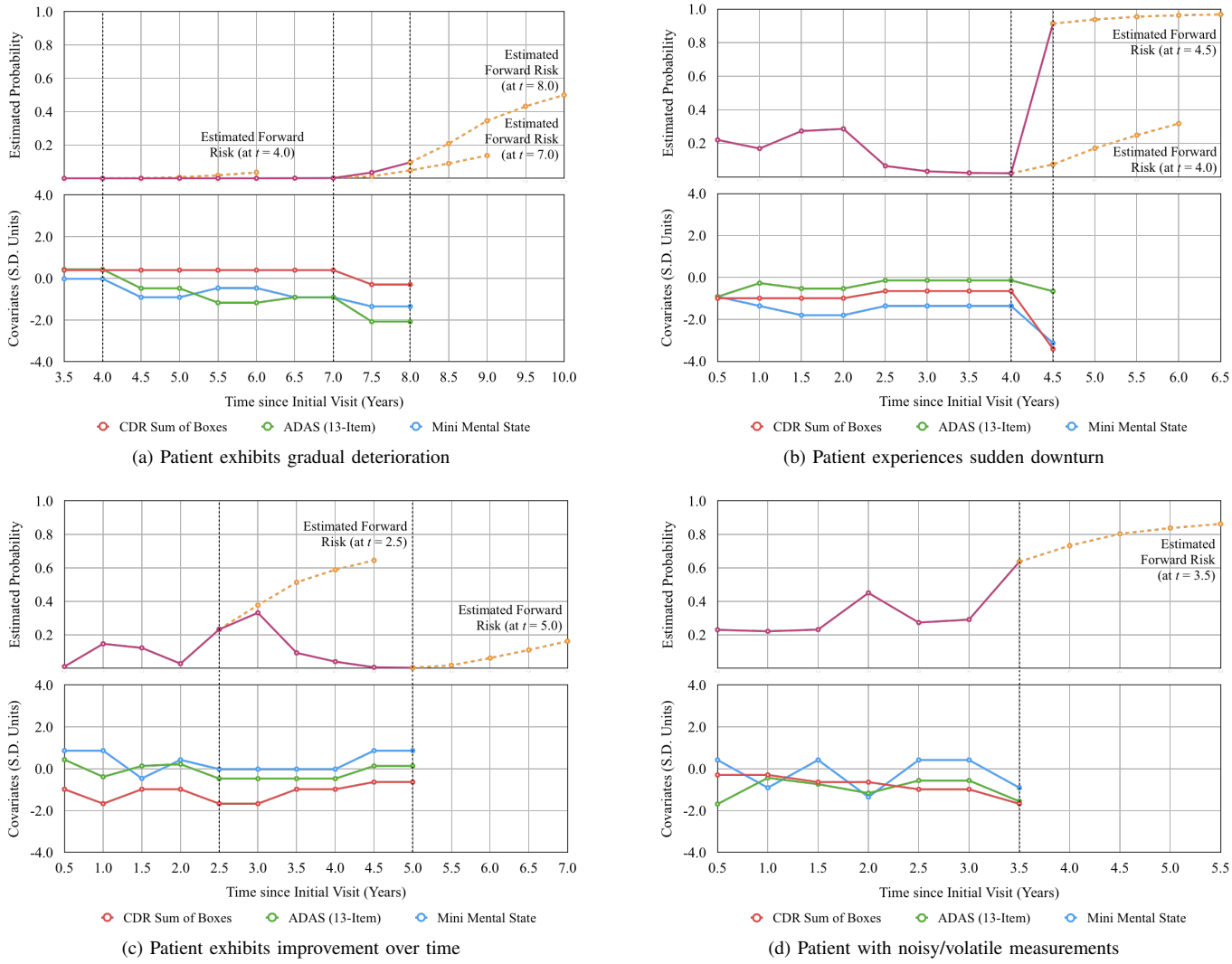
Covariate History. First, what is the value of past measurements? Table V(a) shows the initial benefit from incorporating longitudinal histories of covariate measurements in the most straightforward way—through recurrent neural networks. While this is a reasonable starting point, performance improvements—where positive—appear marginal at best.

Limited Window. In accordance with our hypothesis that *not all* historical information may be beneficial (especially in the presence of noise), we then allow the optimal width of a limited sliding window of history to be selected as a model hyperparameter. Table V(b) shows the benefit from this mechanism; improvements are more promising, boosting average AUPRC by 4% over both RNN and MLP models.

Temporal Convolutions. Now, what is the best way of incorporating this window of information? We answer this question by demonstrating the incremental gain from incorporating temporal convolutions (SW-TCN) over simply flattening the time dimension of the input features for feeding into a fully-connected network (SW-MLP). Table V(c) gives the results, showing added AUPRC gains of 9% with this method.

Missingness Information. Finally, we consider the possible improvement from incorporating informative missingness. As shown in Table V(d), this yields incremental AUPRC gains of 2%. Overall, compared with the joint modeling baseline commonly employed in the context of time-dependent covariates, MATCH-Net achieves average AUPRC improvements of 15% and one-step-ahead improvements of 26%.

Sensitivities on Handling Missingness. Our proposed model uses a restricted-capacity auxiliary convolutional path to handle missingness masks. Table VI shows the performance of different ways of handling missingness. First, we consider a straightforward alternative that simply concatenates the missingness masks as additional channels to the input sequences (Concatenated). This means that missingness masks are handled in exactly the same way as actual features—and are therefore processed simultaneously using the same number of convolutional filters as the feature vectors in the convolutional block. This is in contrast to the reduced-capacity parallel stream proposed for MATCH-Net. Next, we consider a parsimonious method that leaves the missing values zeroed out in the input sequences (Zeroed Out); that is, there is no prior imputation procedure for missing data. Finally, for comparison we also show the baseline method that simply does not incorporate missingness information (Not Included); by definition this is identical to the SW-TCN model. In Table VI(a), bold values indicate best performance, and asterisks on benchmark results indicate statistically significant difference with MATCH-Net at the 5% level (using the two-sample t-test for difference of means). Table VI(b) shows the proportion of runs where the proposed model outperforms (analogous to Table IV). We observe that directly zeroing out the missing values leads to very poor performance, possibly since it is difficult for the



(a) Patient exhibits gradual deterioration

(b) Patient experiences sudden downturn

(c) Patient exhibits improvement over time

(d) Patient with noisy/volatile measurements

Fig. 2. Example application of MATCH-Net for personalized risk scoring, shown for several typical ADNI patients. In each top panel, the historical sequence of one-step-ahead risk (*i.e.* $\tau = 1$) is displayed in solid purple, and estimated forward risks (*i.e.* $\tau > 1$) at selected time points are displayed in dotted yellow, with the specific time points indicated by the adjacent annotations. Each bottom panel displays the trajectories of covariate measurements for the top three longitudinal features most predictive of one-step-ahead risk of Alzheimer's disease diagnosis. Covariate values are normalized and expressed in standard deviation units, where zero corresponds to the mean value across all examples within the training set. Positive (negative) numbers correspond to healthy (unhealthy) covariate values. See Appendix D in the supplementary material for an explanation and visualization of how the variable influence is determined.

network to distinguish between missingness and actual feature values of zero. We also observe that restricting the capacity of learning on the missingness mask leads to slightly more favorable performance, which accords with our hypothesis that there is less information there than in the feature values (hence require fewer parameters so as to prevent overfitting).

Sensitivities on Miscellaneous Techniques. Finally, using MATCH-Net as an example, Table VII indicates individual and cumulative benefits attributable to miscellaneous design choices that we apply to all models in experiments, including oversampling (see Section V-C), label forwarding (see Section VI-B), and the use of elastic net regularization (see, *e.g.*, [54]).

VII. USE CASE: PERSONALIZED SCREENING

While a variety of medical settings may benefit from MATCH-Net as a matter of clinical decision support, we give an example application in the context of personalized screening. In the following, we describe illustrative scenarios involving the disease trajectories of several typical ADNI patients. Each top panel in Figure 2 shows the historical risk trajectory (in terms of one-step-ahead risk $\tau = 1$, in solid purple), as well as forward risk estimates ($\tau > 1$, in dotted yellow) at selected time points for each patient. For additional context, each example is accompanied by a bottom panel indicating the corresponding evolution of the three covariates most predictive of the one-step-ahead risk (see Appendix D in the supplementary material for an explanation of how variable influence is determined; these are consistent with the fact that cognitive scores are known to be indicative of disease state [43], [71], [72]).

Figure 2(a) traces the path of a patient who exhibits gradual deterioration over time. During the first four years of bi-annual clinical visits, the patient exhibits healthy and unremarkable measurements. As of $t = 4.0$, the estimated forward risk—computed by MATCH-Net on the basis of these and other regularly measured biomarkers and tests—is less than 4%. Among other covariate movements, the patient exhibits

gradual declines in ADAS (13-Item) and Mini Mental State exams over the course of the next three years, while CDR Sum of Boxes remains above average in the cohort. As of $t = 7.0$, while the one-step-ahead risk remains low, the predicted 30-month forward risk increases to 14% to reflect these developments. Two time steps later at $t = 8.0$, the projected 30-month forward risk rises to over 50%. Upon inspection, we understand that the recent and simultaneous deterioration of all three cognitive test scores may have contributed to the heightened risk prognosis. Depending on clinical protocols, the physician may be alerted to such sudden increases in dementia risk at various risk thresholds of choice—and may then decide to advise more frequent checkups, or to administer a wider range of tests and biomarker measurements in the immediate term to better assess overall risks. As it turns out in this case, the patient is indeed eventually diagnosed with Alzheimer's disease at $t = 10.5$ years, shedding light on MATCH-Net's potential as an early warning and subject selection system.

Figure 2(b) shows an example where the patient deteriorates much more abruptly over the course of six months. As of $t = 4.0$ years, the predicted one-step-ahead risk is only 2%, and the 30-month forward risk is just over 30%. However, at $t = 4.5$ all risk estimates are above 90% as a result of—among other factors—the sudden and coincidental deterioration in all three cognitive test scores. In fact, the patient is very soon after diagnosed with Alzheimer's disease—at $t = 5.5$, lending credence to the magnitude of MATCH-Net's risk prognoses.

As a contrary example, Figure 2(c) illustrates a patient whose covariate trajectories actually exhibit improvements over time. At $t = 2.5$ years, the 30-month forward risk for the patient is estimated at almost 70%, consistent with the generally negative covariate measurements obtained. However, consistent improvements are recorded over the course of the next two-and-a-half years, such that by $t = 5.0$ the patient's estimated forward risk is less than 20%. In fact, as it turns out in this case, the patient indeed remains *without* any diagnosis of Alzheimer's until the time of right-censoring at $t = 8.0$.

Finally, Figure 2(d) gives an example involving noisy or volatile measurements—due to a variety of possible reasons including fatigue, order of measurements, or inconsistent environments. In this example, the patient's test score trajectories appear rather noisy, and movements seem seldom consistent across tests. For instance, simply analyzing the Mini Mental State exam results would paint a somewhat puzzling picture of a fluctuating patient. However, MATCH-Net issues risk predictions on the basis of aggregate information from *all* available longitudinal trajectories, thereby giving a more holistic assessment of risk invariant to individual noise. Meaningful changes in risk are only observed when important covariate movements are synchronized—for instance, at $t = 3.5$.

VIII. CONCLUSION

In this work we proposed a novel deep learning model for clinical survival analysis. Formulating a generalized conceptual framework for the task of dynamic survival prediction, we presented and assessed MATCH-Net—uniquely designed to leverage longitudinal data for issuing dynamically updated survival predictions. Via performance comparisons with a suite of statistical and deep learning benchmarks, we demonstrated state-of-the-art results on real-world Alzheimer's data, and accounted for incremental sources of gains from various design choices. Future work will benefit from more thorough experimentation within alternative medical settings and datasets, potentially in the context of diseases with different time scales, quantities of features, interactions, as well as competing risks.

REFERENCES

[1] D. Jarrett, J. Yoon, and M. van der Schaar, "Match-net: Dynamic prediction in survival analysis using convolutional neural networks," *NeurIPS Workshop on Machine Learning for Healthcare*, 2018.

[2] R. V. Marinescu, N. P. Oxtoby, A. L. Young *et al.*, "Tadpole challenge: Prediction of longitudinal evolution in alzheimer's disease," *arXiv preprint arXiv:1805.03909*, 2018.

[3] K. A. Doksum and A. Hbyland, "Models for variable-stress accelerated life testing experiments based on wiener processes and the inverse gaussian distribution," *Technometrics*, vol. 34, no. 1, pp. 74–82, 1992.

[4] D. Kleinbaum and M. Klein, "Survival analysis statistics for biology and health," *Survival*, vol. 510, p. 91665, 2005.

[5] G. Rodríguez, "Parametric survival models," *Lectures Notes, Princeton University*, 2005.

[6] M.-L. T. Lee and G. Whitmore, "Proportional hazards and threshold regression: their theoretical and practical connections," *Lifetime data analysis*, vol. 16, no. 2, pp. 196–214, 2010.

[7] T. Fernández, N. Rivera, and Y. W. Teh, "Gaussian processes for survival analysis," in *Advances in Neural Information Processing Systems*, 2016, pp. 5021–5029.

[8] A. M. Alaa and M. van der Schaar, "Deep multi-task gaussian processes for survival analysis with competing risks," in *Proceedings of the 30th Conference on Neural Information Processing Systems*, 2017.

[9] R. Singh and K. Mukhopadhyay, "Survival analysis in clinical trials: Basics and must know areas," *Perspectives in clinical research*, vol. 2, no. 4, p. 145, 2011.

[10] D. R. Cox, "Regression models & life-tables," in *Breakthroughs in statistics*. Springer, 1992.

[11] S. Bai, J. Z. Kolter, and V. Koltun, "An empirical evaluation of generic convolutional and recurrent networks for sequence modeling," *arXiv preprint arXiv:1803.01271*, 2018.

[12] L. A. Beckett, M. C. Donohue, C. Wang *et al.*, "The alzheimer's disease neuroimaging initiative phase 2: Increasing the length, breadth, and depth of our understanding," *Alzheimer's & Dementia*, vol. 11, no. 7, pp. 823–831, 2015.

[13] J. Hardy and D. J. Selkoe, "The amyloid hypothesis of alzheimer's disease: progress and problems on the road to therapeutics," *science*, vol. 297, no. 5580, pp. 353–356, 2002.

[14] B. M. Jedynak, A. Lang, B. Liu *et al.*, "A computational neurodegenerative disease progression score: method and results with the alzheimer's disease neuroimaging initiative cohort," *Neuroimage*, vol. 63, no. 3, pp. 1478–1486, 2012.

[15] M. C. Donohue, H. Jacqmin-Gadda, M. Le Goff *et al.*, "Estimating long-term multivariate progression from short-term data," *Alzheimer's & Dementia*, vol. 10, no. 5, pp. S400–S410, 2014.

[16] L. Frölich, O. Peters, P. Lewczuk *et al.*, "Incremental value of biomarker combinations to predict progression of mci to alzheimer's dementia," *Alzheimer's research & therapy*, vol. 9, no. 1, p. 84, 2017.

[17] T. A. Pascoal, S. Mathotaarachchi, M. Shin *et al.*, "Synergistic interaction between amyloid and tau predicts the progression to dementia," *Alzheimer's & Dementia*, vol. 13, no. 6, pp. 644–653, 2017.

[18] D. Faraggi and R. Simon, "A neural network model for survival data," *Statistics in medicine*, vol. 14, no. 1, pp. 73–82, 1995.

[19] R. L. Prentice and J. D. Kalbfleisch, "Hazard rate models with covariates," *Biometrics*, pp. 25–39, 1979.

[20] J. Buckley and I. James, "Linear regression with censored data," *Biometrika*, vol. 66, no. 3, pp. 429–436, 1979.

[21] J. L. Katzman, U. Shaham, A. Cloninger *et al.*, "Deep survival: A deep cox proportional hazards network," *stat*, vol. 1050, p. 2, 2016.

[22] L. Mariani, D. Coradini, E. Biganzoli *et al.*, "Prognostic factors for metachronous contralateral breast cancer: a comparison of the linear cox regression model and its artificial neural network extension," *Breast cancer research and treatment*, vol. 44, no. 2, pp. 167–178, 1997.

[23] A. Xiang, P. Lapuerta, A. Ryutov, J. Buckley, and S. Azen, "Comparison of the performance of neural network methods and cox regression for censored survival data," *Computational statistics & data analysis*, vol. 34, no. 2, pp. 243–257, 2000.

[24] X. Zhu, J. Yao, and J. Huang, "Deep convolutional neural network for survival analysis with pathological images," in *Bioinformatics and Biomedicine (BIBM), 2016 IEEE International Conference on*. IEEE, 2016, pp. 544–547.

[25] T. Ching, X. Zhu, and L. X. Garmire, "Cox-nnet: An artificial neural network method for prognosis prediction of high-throughput omics data," *PLoS computational biology*, vol. 14, no. 4, p. e1006076, 2018.

[26] K. Liestbl, P. K. Andersen, and U. Andersen, "Survival analysis and neural nets," *Statistics in medicine*, vol. 13, no. 12, pp. 1189–1200, 1994.

[27] E. M. Biganzoli, F. Ambrogi, and P. Boracchi, "Partial logistic artificial neural networks (plann) for flexible modeling of censored survival data," in *Neural Networks, 2009. IJCNN 2009. International Joint Conference on*. IEEE, 2009, pp. 340–346.

[28] M. Luck, T. Sylvain, H. Cardinal, A. Lodi, and Y. Bengio, "Deep learning for patient-specific kidney graft survival analysis," *arXiv preprint arXiv:1705.10245*, 2017.

[29] C.-N. Yu, R. Greiner, H.-C. Lin, and V. Baracos, "Learning patient-specific cancer survival distributions as a sequence of dependent regressors," in *Advances in Neural Information Processing Systems*, 2011, pp. 1845–1853.

[30] C. Lee, W. R. Zame, J. Yoon, and M. van der Schaar, "Deephit: A deep learning approach to survival analysis with competing risks." AAAI, 2018.

[31] H. van Houwelingen and H. Putter, *Dynamic prediction in clinical survival analysis*. CRC, 2011.

[32] G. L. Hickey, P. Philipson, A. Jorgensen, and R. Kolamunnage-Dona, "Joint modelling of time-to-event and multivariate longitudinal outcomes: recent developments and issues," *BMC medical research methodology*, vol. 16, no. 1, p. 117, 2016.

[33] S. Parisot, S. I. Ktena, E. Ferrante *et al.*, "Disease prediction using graph convolutional networks: Application to autism spectrum disorder and alzheimer's disease," *Medical image analysis*, 2018.

[34] B. Lim and M. van der Schaar, "Forecasting disease trajectories in alzheimer's disease using deep learning," *KDD Workshop on Machine Learning for Medicine and Healthcare*, 2018.

[35] M. Z. Nezhad, N. Sadati, K. Yang, and D. Zhu, "A deep active survival analysis approach for precision treatment recommendations: Application of prostate cancer," *Expert Systems with App*, vol. 115, pp. 16–26, 2019.

[36] K. Matsuo, S. Purushotham, B. Jiang, R. S. Mandelbaum, T. Takiuchi, Y. Liu, and L. D. Roman, "Survival outcome prediction in cervical cancer: Cox models vs deep-learning model," *American journal of obstetrics and gynecology*, vol. 220, no. 4, pp. 381–e1, 2019.

[37] C. Lee, J. Yoon, and M. Van Der Schaar, "Dynamic-deephit: A deep learning approach for dynamic survival analysis with competing risks based on longitudinal data," *IEEE Transactions on Biomedical Engineering*, 2019.

[38] P. Wang, Y. Li, and C. K. Reddy, "Machine learning for survival analysis: A survey," *ACM Computing Surveys (CSUR)*, vol. 51, no. 6, p. 110, 2019.

[39] R. Cui and M. Liu, "Hippocampus analysis by combination of 3d densenet and shapes for alzheimer's disease diagnosis," *IEEE journal of biomedical and health informatics*, 2018.

[40] E. Alickovic, A. Subasi *et al.*, "Automatic detection of alzheimer disease based on histogram and random forest," in *International Conference on Medical and Biological Engineering*. Springer, 2019, pp. 91–96.

[41] W. Li, Y. Zhao, X. Chen, Y. Xiao, and Y. Qin, "Detecting alzheimer's disease on small dataset: A knowledge transfer perspective," *IEEE journal of biomedical and health informatics*, 2018.

[42] R. V. Marinescu, M. Lorenzi, S. Blumberg, A. L. Young, P. P. Morell, N. P. Oxtoby, A. Eshaghi, K. X. Yong, S. J. Crutch, and D. C. Alexander, "Disease knowledge transfer across neurodegenerative diseases," *arXiv preprint arXiv:1901.03517*, 2019.

[43] P. Jiang, X. Wang, Q. Li *et al.*, "Correlation-aware sparse and low-rank constrained multi-task learning for longitudinal analysis of alzheimer's disease," *IEEE journal of biomedical and health informatics*, 2018.

[44] G. Lee, K. Nho, B. Kang, K.-A. Sohn, and D. Kim, "Predicting alzheimer's disease progression using multi-modal deep learning approach," *Scientific reports*, vol. 9, no. 1, p. 1952, 2019.

[45] M. Grassi, N. Rouleaux, D. Caldirola *et al.*, "A novel ensemble-based ml algorithm to predict the conversion from mild cognitive impairment to alzheimer's disease using socio-demographic characteristics, clinical information and neuropsychological measures," *bioRxiv*, p. 564716, 2019.

[46] A. A. Tsiatis and M. Davidian, "Joint modeling of longitudinal and time-to-event data: an overview," *Statistica Sinica*, pp. 809–834, 2004.

[47] Y. Zheng and P. J. Heagerty, "Partly conditional survival models for longitudinal data," *Biometrics*, vol. 61, no. 2, pp. 379–391, 2005.

[48] H. C. Van Houwelingen, "Dynamic prediction by landmarking in event history analysis," *Scandinavian Journal of Statistics*, vol. 34, no. 1, pp. 70–85, 2007.

[49] J. Barrett and L. Su, "Dynamic predictions using flexible joint models of longitudinal and time-to-event data," *Statistics in medicine*, vol. 36, no. 9, pp. 1447–1460, 2017.

[50] D. Rizopoulos, *Joint models for longitudinal and time-to-event data: With applications in R*. Chapman and Hall/CRC, 2012.

[51] Y. Gal and Z. Ghahramani, "Dropout as a bayesian approximation: Representing model uncertainty in deep learning," in *international conference on machine learning*, 2016, pp. 1050–1059.

[52] D. B. Rubin, "Inference and missing data," *Biometrika*, vol. 63, no. 3, pp. 581–592, 1976.

[53] S.-J. Bang, Y. Wang, and Y. Yang, "Phased-lstm based predictive model for longitudinal ehr data with missing values."

[54] H. Zou and T. Hastie, "Regularization and variable selection via the elastic net," *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, vol. 67, no. 2, pp. 301–320, 2005.

[55] K. Li and S. Luo, "Functional joint model for longitudinal and time-to-event data: an application to alzheimer's disease," *Statistics in medicine*, vol. 36, no. 22, pp. 3560–3572, 2017.

[56] K. Liu, K. Chen *et al.*, "Prediction of mci conversion using a combination of independent component analysis and the cox model," *Frontiers in human neuroscience*, vol. 11, p. 33, 2017.

[57] S. J. Vos, F. Verhey, L. Frölich *et al.*, "Prevalence and prognosis of alzheimer's disease at the mild cognitive impairment stage," *Brain*, vol. 138, no. 5, pp. 1327–1338, 2015.

[58] A. Avati, "Classifier evaluation metrics," in *Machine Learning*. Stanford University, 2017.

[59] J. W. Grzymala-Busse, L. K. Goodwin, W. J. Grzymala-Busse, and X. Zheng, "An approach to imbalanced data sets based on changing rule strength," in *Rough-Neural Computing*. Springer, 2004, pp. 543–553.

[60] B. Mac Namee, P. Cunningham, S. Byrne *et al.*, "The problem of bias in training data in regression problems in medical decision support," *Artificial intelligence in medicine*, vol. 24, no. 1, pp. 51–70, 2002.

[61] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018.

[62] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, "Smote: synthetic minority over-sampling technique," *Journal of artificial intelligence research*, vol. 16, pp. 321–357, 2002.

[63] T. M. Therneau, "A package for survival analysis in s," *R package version 2.42*, 2018.

[64] F. Hu, "Stepwise variable selection procedures for regression analysis," *R package version 0.1.0*, 2018.

[65] Y. Tripodis and K. Davis-Plourde, "A comparison between mixed effect and joint models for survival and longitudinal model estimates," *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, vol. 13, no. 7, p. P509, 2017.

[66] L. Wu, *Mixed effects models for complex data*. Chapman and Hall/CRC, 2009.

[67] W. Liu, B. Zhang, Z. Zhang, and X.-H. Zhou, "Joint modeling of transitional patterns of alzheimer's disease," *PloS one*, vol. 8, no. 9, p. e75487, 2013.

[68] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 3431–3440.

[69] A. Van Den Oord, S. Dieleman, H. Zen, K. Simonyan, O. Vinyals, A. Graves, N. Kalchbrenner, A. W. Senior, and K. Kavukcuoglu, "Wavenet: A generative model for raw audio." in *SSW*, 2016, p. 125.

[70] T. Saito and M. Rehmsmeier, "The precision-recall plot is more informative than the roc plot when evaluating binary classifiers on imbalanced datasets," *PloS one*, vol. 10, no. 3, p. e0118432, 2015.

[71] J. R. Petrella, R. E. Coleman, and P. M. Doraiswamy, "Neuroimaging and early diagnosis of alzheimer disease: a look to the future," *Radiology*, vol. 226, no. 2, pp. 315–336, 2003.

[72] S. E. O'Bryant, S. C. Waring, C. M. Cullum *et al.*, "Staging dementia using clinical dementia rating scale sum of boxes scores: a texas alzheimer's research consortium study," *Archives of neurology*, vol. 65, no. 8, pp. 1091–1095, 2008.

[73] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001.

[74] M. D. Zeiler and R. Fergus, "Visualizing and understanding convolutional networks," in *European conference on computer vision*. Springer, 2014, pp. 818–833.

Supplementary Material: Dynamic Prediction in Clinical Survival Analysis using Temporal Convolutional Networks

ACKNOWLEDGEMENTS

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). ADNI is funded by the National Institute on Aging, the National Institute of Biomedical Imaging and Bioengineering, and through generous contributions from the following: AbbVie, Alzheimer's Association; Alzheimer's Drug Discovery Foundation; Araclon Biotech; BioClinica, Inc.; Biogen; Bristol-Myers Squibb Company; CereSpir, Inc.; Cogstate; Eisai Inc.; Elan Pharmaceuticals, Inc.; Eli Lilly and Company; EuroImmun; F. Hoffmann-La Roche Ltd and its affiliated company Genentech, Inc.; Fujirebio; GE Healthcare; IXICO Ltd.; Janssen Alzheimer Immunotherapy Research & Development, LLC.; Johnson & Johnson Pharmaceutical Research & Development LLC.; Lumosity; Lundbeck; Merck & Co., Inc.; Meso Scale Diagnostics, LLC.; NeuroRx Research; Neurotrack Technologies; Novartis Pharmaceuticals Corporation; Pfizer Inc.; Piramal Imaging; Servier; Takeda Pharmaceutical Company; and Transition Therapeutics. The Canadian Institutes of Health Research is providing funds to support ADNI clinical sites in Canada. Private sector contributions are facilitated by the Foundation for the National Institutes of Health. The grantee organization is the Northern California Institute for Research and Education, and the study is coordinated by the Alzheimer's Therapeutic Research Institute at the University of Southern California. ADNI data are disseminated by the Laboratory for Neuroimaging at the University of Southern California.

APPENDIX A
PSEUDOCODE OF TRAINING ALGORITHM

Algorithm 1 MATCH-Net Training Procedure

Input: $\{\langle \mathbf{X}_{i,t,w} \rangle_{t=1}^{t_i}\}_{i \in \text{train}}$, $\{\langle \mathbf{Z}_{i,t,w} \rangle_{t=1}^{t_i}\}_{i \in \text{train}}$
Output: Calibrated network weights $\boldsymbol{\theta}$

$\boldsymbol{\theta}_{\text{best}} \leftarrow$ None; $\mathcal{C}_{\text{best}} \leftarrow 0$
for count $= 1$ to maximum iterations I **do**
  Sample minibatch $\mathcal{M} \in \{\langle (\mathbf{X}_{i,t,w}, \mathbf{Z}_{i,t,w}) \rangle_{t=1}^{t_i}\}_{i \in \text{train}}$
  Compute sample loss $l(\boldsymbol{\theta})$ on $\mathcal{M}$
  Update $\boldsymbol{\theta} \leftarrow \text{Adam}(l, \mathcal{M})$
  if count$|10$ **then**
    for $i \in$ validation **do**
      for $t < t_i$ **do**
        Predict failures $\langle \hat{F}_i(t+k|t,w) \rangle_{k=1}^{\tau_i}$
      end for
    end for
    Compute $\mathcal{C}_{\text{validation}}$
    if $\mathcal{C}_{\text{validation}} > \mathcal{C}_{\text{best}}$ **then**
      $\boldsymbol{\theta}_{\text{best}} \leftarrow \boldsymbol{\theta}_{\text{validation}}$
      $\mathcal{C}_{\text{best}} \leftarrow \mathcal{C}_{\text{validation}}$
    end if
  end if
  if converged **then**
    break
  end if
end for
return $\boldsymbol{\theta}_{\text{best}}$

APPENDIX B
EXAMPLES OF CENSORING

In general, censoring occurs when the data captures some information about an individual's survival, but the exact survival time remains unknown. Figure 3 illustrates examples of censored patient trajectories. In the longitudinal setting, survival status and covariate values are measured through time. Solid circles indicate failure, and empty circles indicate otherwise. Patients A and B are uncensored. Right-censoring occurs when the event of interest happens *after* some cutoff time, subsequent to which the event is no longer observable. Patient C is right-censored in the middle of the study period, or instance due to withdrawal or being lost to follow-up. Patient D is subject to administrative censoring at the end of the study period.

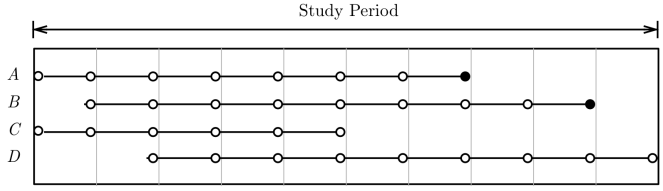


Fig. 3. Examples of censoring.

APPENDIX C
STATE SPACE OF DIAGNOSES

As described in Section V-A, the clinical diagnosis of a patient at each visit may be either stable or transitive; the former can be either normal brain functioning (NL), mild cognitive impairment (MCI), or Alzheimer's disease (AD), and the latter transitions between these categories. Figure 4 illustrates the space of all possible diagnosis states. Our objective is to predict the first stable diagnosis of Alzheimer's disease for each patient (*i.e.* the right-most state in Figure 4). See Section V-A in the main manuscript for a more detailed discussion of the dataset.

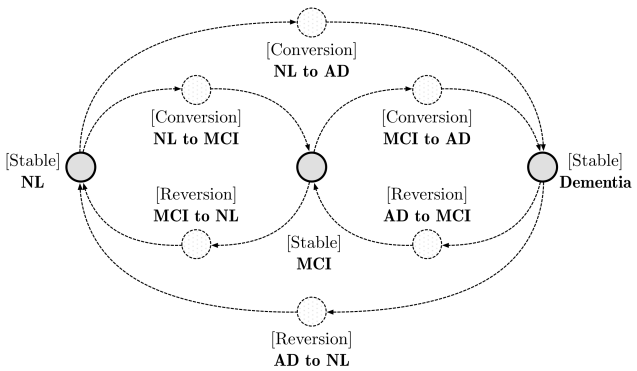


Fig. 4. State space of clinical diagnoses. Self-loops are omitted for clarity.

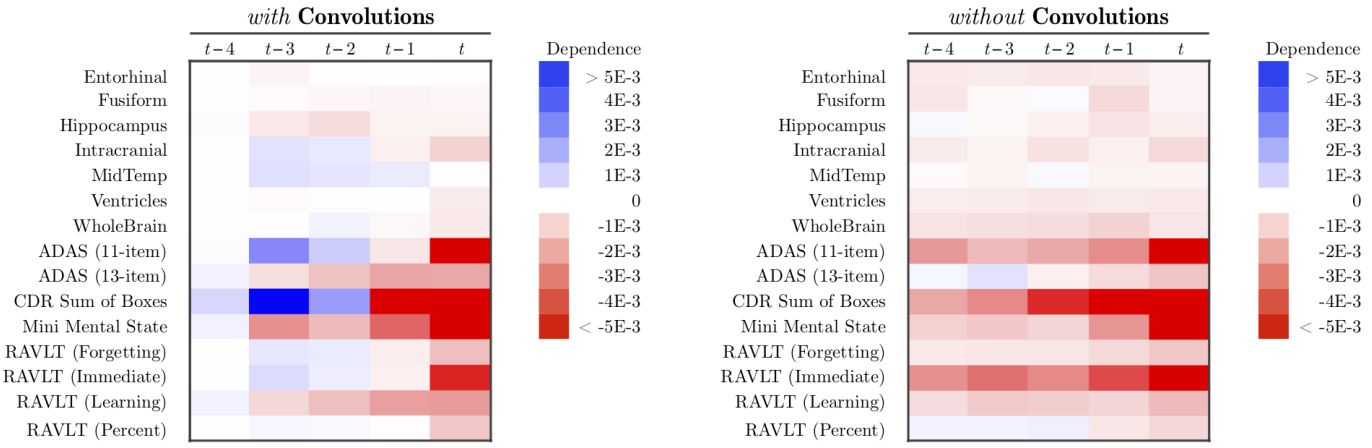


Fig. 5. Average saliency maps indicating feature and temporal influence within the sliding window, computed using slopes of partial dependence function π on numerical features, where window width $w = 5$. *Ceteris paribus*, without convolutions, worse covariate values are almost invariably associated with increased risk of failure. Incorporating convolutions, worse covariate values at earlier time steps—thereby inducing subsequent improvements *ceteris paribus*—sometimes results in decreased risk of failure. This accords with our motivation of employing temporal convolutions to better capture the importance of relative movements.

Appendix D
Visualizing Variable Influence

From the preceding, we observe the largest gains by introducing convolutions, compared to simply flattening all historical time steps into a one-dimensional input vector. This is consistent with our motivating hypothesis that convolutions are better able to capture explicit temporal patterns. Here we adopt the partial dependence approach in [73] to understand the input-output relationship in more detail, and give a possible explanation for the performance improvement from introducing convolutions. First, for each observed covariate d, we want to approximate how the estimated failure function varies based on the value of $\mathbf{x}^d_{t,w}$. We define partial dependence

$$\begin{aligned}\Pi(t+\tau, \mathbf{x}^d_{t,w}) &= \mathbb{E}_{\mathbf{X}^{(-d)}_{t,w}}[\hat{F}(t+\tau|\mathbf{X}_{t,w})] \\ &\approx \frac{1}{\sum_{i=1}^{N} t_i} \sum_{i=1}^{N} \sum_{j=1}^{t_i} \hat{F}(j+\tau|\mathbf{x}^d_{j,w}, \mathbf{X}^{(-d)}_{j,w})\end{aligned} \tag{11}$$

where $\mathbf{x}^d_{t,w} \cup \mathbf{X}^{(-d)}_{t,w} = \mathbf{X}_{t,w}$ decomposes the covariate input matrix into the feature of interest and the remaining features. In other words, the partial dependence is simply the expected value of the the failure probability estimate as a function of specified values for $\mathbf{x}^d_{t,w}$, with the expectation taken over the empirical distribution of $\mathbf{X}^{(-d)}_{t,w}$. In addition, to account for the temporal dimension of longitudinal covariate histories, we note that $\mathbf{x}^d_{t,w} = \langle x^d_{t-w+1}, ..., x^d_{t-1}, x^d_t \rangle$, which allows us to similarly define the time-dependent partial dependence

$$\pi(t+\tau, x^d_{t-k}) = \mathbb{E}_{\mathbf{X}_{t,w} \setminus x^d_{t-k}}[\hat{F}(t+\tau|\mathbf{X}_{t,w})] \tag{12}$$

for some choice of $k \in [0, w)$. While Equation 11 allows us to examine the expected failure by varying all historical values $\mathbf{x}^d_{t,w}$ for feature d simultaneously, Equation 12 allows us to examine the expected failure by varying individual values x^d_{t-k} for specific points within the longitudinal window.

Saliency Maps. By evaluating Equation 12 on the range of values $\mathbf{x}^d_{t,w}$ present in the data, the influence of each covariate and historical time step can be measured by estimating its slope. To obtain a global picture of what impact each feature and time step has on the model's predictions, we can compute the influence for all features and time steps to produce a *saliency map* [74]. Figure 5 shows such a map of the sliding input window, indicating the influence of each numerical feature and historical time step on the one-step-ahead failure estimates produced by the proposed architecture—both with and without temporal convolutions. Absent convolutions we observe, *ceteris paribus*, that having worse covariate values any time step almost invariably has an upward impact on risk (*i.e.* negatively correlated with failure). On the other hand, with temporal convolutions we observe, *ceteris paribus*, that having worse covariate values at earlier time steps—thereby producing a subsequent improvement—may sometimes actually result in a downward impact on risk (*i.e.* positively correlated with failure). This suggests that convolutions may better facilitate modeling relative movements in covariate trajectories (*e.g.* improvements or deteriorations) than simply paying attention to levels. This provides a possible explanation for the superior performance of temporal convolutional networks for survival prediction.

Appendix E
Hyperparameter Selection Ranges

Hyperparameter	Selection Range
Connected Layers	1, 2, 3, 4, 5
Convolutional Layers	1, 2, 3, 4, 5
Dropout Rate	0.1, 0.2, 0.3, 0.4, 0.5
Epochs for Convergence	1, 2, 3, 4, 5, 6, 7, 8, 9, 10
Learning Rate	1e-4, 3e-4, 1e-3, 3e-3, 1e-2, 3e-2
L1-Regularlisation	0, 1e-4, 3e-4, 1e-3, 3e-3, 1e-2, 3e-2
L2-Regularization	0, 1e-4, 3e-4, 1e-3, 3e-3, 1e-2, 3e-2
Minibatch Size	32, 64, 128, 256, 512
Number of Filters (Covariates)	32, 64, 128, 256, 512
Number of Filters (Masks)	8, 16, 32, 64, 128
Oversample Ratio	None, 1, 2, 3, 5, 10
Recurrent Unit State Size	1×, 2×, 3×, 4×, 5×
Width of Connected Layers	32, 64, 128, 256, 512
Width of Convolutional Filters	3, 4, 5, 6, 7, 8, 9, 10
Width of Sliding Window	3, 4, 5, 6, 7, 8, 9, 10