Received: 7 November 2021 | Revised: 18 February 2022 | Accepted: 8 March 2022

DOI: 10.1002/sim.9392

RESEARCH ARTICLE

Statistics in Medicine WILEY

Deep learning for the dynamic prediction of multivariate longitudinal and survival data

Jeffrey Lin[1] | **Sheng Luo**[2]

[1]Department of Biostatistics and Data Science, The University of Texas Health Science Center at Houston, Houston, Texas,

[2]Department of Biostatistics and Bioinformatics, Duke University Medical Center, Durham, North Carolina,

Correspondence
Sheng Luo, Biostatistics and Bioinformatics, Duke University Medical Center, Durham, NC, USA.
Email: sheng.luo@duke.edu

Funding information
National Institute on Aging, Grant/Award Numbers: P30AG028716, P30AG072958, R01AG064803; ADNI (National Institutes of Health), Grant/Award Number: U01 AG024904; DOD ADNI (Department of Defense), Grant/Award Number: W81XWH-12-2-0012; NIA/NIH Grant, Grant/Award Number: U24 AG072122

Abstract

The joint model for longitudinal and survival data improves time-to-event predictions by including longitudinal outcome variables in addition to baseline covariates. However, in practice, joint models may be limited by parametric assumptions in both the longitudinal and survival submodels. In addition, computational difficulties arise when considering multiple longitudinal outcomes due to the large number of random effects to be integrated out in the full likelihood. In this article, we discuss several recent machine learning methods for incorporating multivariate longitudinal data for time-to-event prediction. The presented methods use functional data analysis or convolutional neural networks to model the longitudinal data, both of which scale well to multiple longitudinal outcomes. In addition, we propose a novel architecture based on the transformer neural network, named TransformerJM, which jointly models longitudinal and time-to-event data. The prognostic abilities of each model are assessed and compared through both simulation and real data analysis on Alzheimer's disease datasets. Specifically, the models were evaluated based on their ability to dynamically update predictions as new longitudinal data becomes available. We showed that TransformerJM improves upon the predictive performance of existing methods across different scenarios.

KEYWORDS

Alzheimer's disease, functional data analysis, joint model, personalized medicine, temporal convolutions, transformer neural network

1 | INTRODUCTION

Alzheimer's disease (AD) is a progressive, neurodegenerative disorder which begins with mild memory loss and advances to more severe inhibition of other mental functions.[1,2] Although no cure currently exists for AD, therapeutic treatments may alleviate symptoms and slow progression when administered early in the disease process.[1,3] In particular, mild cognitive impairment (MCI) is often considered a transitional state between normal cognitive aging and AD, with 32% of MCI patients progressing to AD within 5 years.[4,5] As such, there is a crucial need for accurate prediction of AD progression during the early stages of the disease for both efficacious treatment and aiding in subject selection for clinical trials.

AD patients are typically followed over the course of the disease, resulting in repeated longitudinal measurements for multiple outcome variables. An important consideration for improving predictions for time-to-AD onset is the inclusion of these multivariate longitudinal outcomes in the survival model. Many existing survival methods use only the last

available observation even if repeated measurements are available. This poses a serious limitation, as modeling based on accrued longitudinal information leads to significant improvement in the prediction of survival risk.[6,7] For example, the joint model (JM) for longitudinal and survival data[6] was used to assess the ability of various longitudinal biomarkers to predict the time to AD conversion.[8] This univariate analysis showed several clinical assessments such as the Alzheimer's Disease Assessment Scale-Cognitive 13 (ADAS-Cog13) and Rey Auditory Verbal Learning Tests (RAVLT) are important indicators of conversion from MCI to AD. While the JM approach has also been extended to include multiple longitudinal outcomes,[7] it remains computationally challenging due to the large number of random effects to be integrated out in the full likelihood.

Several alternatives to modeling time-to-event data using multiple longitudinal outcomes have been proposed. MFPCCox[9] is a two-staged modeling framework which considers the multivariate longitudinal outcomes as functional data, allowing it to have significantly faster computational times in comparison to multivariate JM. In the first step, multivariate functional principal component analysis (MFPCA)[10] is used to extract informative features from the multiple longitudinal outcomes in the form of multivariate functional principal component (MFPC) scores. The scores account for both the autocorrelation from within a longitudinal outcome and the correlation between multiple longitudinal outcomes. In the second step, the MFPC scores are included directly as time-independent covariates together with other baseline variables in a Cox proportional hazards model. Similarly, the functional survival forest[11] extended this concept to use MFPCA in conjunction with the random survival forest (RSF) to capture more complex interactions between variables in modeling the survival data without further subjective input. These findings show that MFPCA can be used to easily incorporate longitudinal information into other flexible and nonparametric survival models, such as neural networks for time-to-event data. Specifically, multilayer perceptrons were used to extend the Cox model to consider nonlinear covariate effects.[12] This model was later improved using modern deep learning techniques such as activation functions and dropout, resulting in an updated model named DeepSurv.[13]

Recently, end-to-end neural network models using longitudinal data for survival analysis have also been proposed. MATCH-net,[14] a missingness-aware temporal convolutional time-hitting network, was developed to adapt a convolutional neural network (CNN) to predict survival probabilities based on multiple longitudinal outcomes. MATCH-net addressed the challenges of unstructured longitudinal data by accounting for both irregular sampling (allowing for the interval between clinical visits to vary) and asynchronous missingness (some longitudinal measurements may be missing at some visits). Incorporating multiple longitudinal outcomes allowed MATCH-net to achieve high predictive performance in modeling AD onset.

Lastly, the transformer[15] is a recent neural network architecture designed for natural language processing tasks. It has achieved state-of-the-art benchmarks in language translation, text generation, and image recognition in comparison to previous methods using recurrent neural networks (RNN) and CNN. In addition, the transformer has been successfully applied to EHR data, using longitudinal clinical data for tasks such as classification of patient outcomes and providing diagnoses across a range of diseases.[16,17] In this article, we propose a novel transformer architecture named TransformerJM to jointly model multivariate longitudinal and survival data. Analogous to how the original transformer takes in a sequence of words and translates it from one language to another, the proposed model takes a sequence of multivariate longitudinal data as input and predicts sequences of both the future longitudinal outcomes and survival probabilities. Similar to other joint models, TransformerJM allows for personalized predictions which can be dynamically updated as new longitudinal information is made available. Providing accurate prognosis of the longitudinal outcomes and the survival risk to clinicians may aid them in making guided decisions and selecting appropriate treatment plans for their patients.

This article is structured as follows. In Section 2, a review of current methods for longitudinal and survival analysis is provided. The details of the proposed TransformerJM architecture are given. In Section 3, a simulation study was conducted to assess the performance of the presented methods in different scenarios. In Section 4, the presented methods were applied to AD datasets. Lastly in Section 5, we discuss the results and make some recommendations regarding each model.

2 | METHODS

Let I be the total number of patients enrolled in a study. Each patient i ($i = 1, \ldots, I$) is followed for multiple visits, indexed by $j = 1, \ldots, J_i$. Each patient is observed until time $T_i^* = \min(T_i, C_i)$, where T_i is the true event time, and C_i is the independent censoring time. The event indicator $\delta_i = \mathbf{1}(T_i \leq C_i)$ denotes whether the event was observed or censored. For each patient, a set of P time-independent covariates and Q longitudinal outcomes were collected. Let the matrix $\boldsymbol{Z}_{I\times P}$ denote

the values of time-independent covariates. Let $Y_{iq}(t_{ij})$ refers to the observed values for the qth longitudinal outcome for subject i at visit j. When implementing the models, $Y_{iq}(t_{ij})$ is represented as the input matrix $\boldsymbol{Y}_{I\times J\times Q}$, where $J \geq J_i,\ \forall i$. The form of $\boldsymbol{Y}$ differs between methods and will be defined in each of the following sections. In addition, the notation $\boldsymbol{Y}^{(t)}$ is also used to denote the subset of longitudinal data up to time t.

2.1 | Functional principal component analysis-based methods

By treating longitudinal variables as sparse functional data, functional principal component analysis (FPCA)[18] can be used to extract informative features from the longitudinal process as a set of scores. The FPC scores are defined using the Karhunen–Loève theorem which states that for a single longitudinal variable q, the trajectory $X_{iq}(t)$ can be expanded as

$$X_{iq}(t) = \mu_q(t) + \sum_{m=1}^{\infty} \xi_{iqm}\phi_{qm}(t) \ .$$

Here, $\mu_q(t)$ is the unknown smoothed mean function with the covariance between time points t and t' denoted as the function $\Sigma_q(t,t')$. The spectral decomposition of the covariance function yields $\Sigma_q(t,t') = \sum_{m=1}^{\infty}\lambda_{qm}\phi_{qm}(t)\phi_{qm}(t')$, where λ_{qm} and $\phi_{qm}(t)$ are the nonincreasing eigenvalues and the eigenvectors, respectively. $X_{iq}(t)$ can be sufficiently approximated with the first M_q eigenfunctions, $X_{iq}(t) \approx \mu_q(t) + \sum_{m=1}^{M_q}\xi_{iqm}\phi_{qm}(t)$, where M_q can be chosen based on a desired percentage of variance explained (PVE).[18] In practice, the true longitudinal trajectories are not observed. Rather, the observed data are given by $Y_{iq}(t_{ij}) = X_{iq}(t_{ij}) + \epsilon_{iq}(t_{ij})$, where $\epsilon_{iq}(t_{ij})$ is measurement error assumed to be normally distributed. Therefore, FPCA uses the Principal Analysis by Conditional Estimation algorithm to first obtain estimates of the mean function $\hat{\mu}_q(t)$, error variance $\hat{\sigma}_q^2$, covariance function $\hat{\Sigma}_q(t,t')$, eigenvalues $\hat{\lambda}_{qm}$, and eigenfunctions $\hat{\phi}_{qm}(t)$. The FPC scores can then be calculated using these estimated values as

$$\hat{\xi}_{iqm} = \hat{\lambda}_{qm}(\hat{\boldsymbol{\phi}}_{iqm})^T\hat{\Sigma}_{Y_{iq}}^{-1}(\boldsymbol{Y}_{iq} - \hat{\boldsymbol{\mu}}_q) \ , \tag{1}$$

where $\boldsymbol{Y}_{iq} = \{Y_{iq}(t_{ij})\}_{j=1,\dots,J_i}$, $\hat{\boldsymbol{\mu}}_{iq} = \{\hat{\mu}_{iq}(t_{ij})\}_{j=1,\dots,J_i}$, $\hat{\boldsymbol{\phi}}_{iqm} = \{\hat{\phi}_{iqm}(t_{ij})\}_{j=1,\dots,J_i}$, and $\hat{\Sigma}_{Y_{iq}}$ is a $J_i \times J_i$ matrix with the (j,j') entry defined as $(\hat{\Sigma}_{Y_{iq}})_{jj'} = \hat{\Sigma}_q(t_{ij},t_{ij'}) + \hat{\sigma}_q^2\delta_{jj'}$, $\delta_{jj'} = 1$ for $j = j'$ and 0 otherwise.

The correlation between the multiple longitudinal outcomes may lead to nonnegligible correlation between the sets of FPC scores for each outcome in Equation (1). Multivariate FPCA (MFPCA)[10] indirectly accounts for the correlation between longitudinal variables by modeling the correlation between each set of univariate FPC scores. Let $M_+ = \sum_{q=1}^{Q}M_q$ denote the total number of univariate FPC scores, where M_q is the number of FPC scores for the qth outcome. Let $\boldsymbol{\Theta}_{I\times M_+}$ be a matrix with each row containing the concatenated FPC scores of the longitudinal variables. Decompose the matrix $\boldsymbol{H}_{M_+\times M_+} = (n-1)^{-1}\boldsymbol{\Theta}^T\boldsymbol{\Theta}$ to obtain the orthonormal eigenvectors $\{\hat{\boldsymbol{c}}_k\}_{k=1,\dots,M_+}$ and respective eigenvalues $\{\hat{v}_k\}_{k=1,\dots,M_+}$. Then the estimated multivariate eigenfunctions is given by $\hat{\psi}_{qk}(t) = \sum_{m=1}^{M_q}[\hat{\boldsymbol{c}}_k]_m^{(q)}\hat{\phi}_{qm}(t)$, where $[\hat{\boldsymbol{c}}_k]^{(q)}$ denotes the qth block of $\hat{\boldsymbol{c}}_k$. The estimated MFPC scores can be calculated for each subject using the univariate FPC scores and $[\hat{\boldsymbol{c}}_k]^{(q)}$ with the following expression:

$$\hat{\rho}_{ik} = \sum_{q=1}^{Q}\sum_{m=1}^{M_q}[\hat{\boldsymbol{c}}_k]_m^{(q)}\hat{\xi}_{iqm} \ .$$

Lastly, the qth longitudinal outcome, $X_{iq}(t)$, can be sufficiently approximated by selecting the first $M^* < M_+$ scores and eigenfunctions based on PVE or other information criterion.

$$\hat{X}_{iq}(t) \approx \hat{\mu}_q(t) + \sum_{k=1}^{M^*}\hat{\rho}_{ik}\hat{\psi}_{qk}(t) \ .$$

MFPCA is implemented in the R package `MFPCA`.[19] The current implementation requires the longitudinal data to lie on a grid with fixed time intervals (ie, every 6 months), however, missing data is allowed. The observed data $Y_{iq}(t_{ij})$ is first rounded to the nearest time corresponding to the fixed grid. Missing values are marked with `NA`'s. The input matrix is given by $\boldsymbol{Y}_{I\times J\times Q}$, where J is the grid length needed to accommodate the largest observation time.

2.1.1 | MFPCA-Cox

Previous works have used the scores extracted by MFPCA directly as covariates in subsequent survival models such as the Cox model and RSF.[9,11] If a Cox model is adopted in modeling the survival data, the framework is referred to as MFPCA-Cox (also known as MFPCCox in the original paper[9]). MFPCA-Cox defines the hazard function for the ith subject as $\lambda_i(t) = \lambda_0(t)\exp\left(\boldsymbol{Z}_i^T\boldsymbol{\gamma} + \hat{\boldsymbol{\rho}}_i^T\boldsymbol{\beta}\right)$, where $\boldsymbol{Z}_i$ is time-independent covariate vector and $\hat{\boldsymbol{\rho}}_i$ is the estimated MFPC scores, and $\boldsymbol{\gamma}$ and $\boldsymbol{\beta}$ are their corresponding regression coefficient vectors. In particular, during model building, $\hat{\boldsymbol{\rho}}_i$ is the set of MFPC scores derived from all longitudinal observations that occur prior to the event time. To make a prediction for a new subject who has not yet experienced the event at time t, their MFPC scores are calculated from all available longitudinal observations up to t and passed to the fitted Cox model to obtain subject-specific survival probabilities. Details on fitting the MFPCA-Cox model and using it for predictions can be found in the original paper.[9]

2.1.2 | MFPCA-DeepSurv

The same methodology was applied to incorporate longitudinal information into the DeepSurv model. DeepSurv[13] is a deep feed-forward neural network based on the Cox model. In the traditional Cox model, a patient's log-risk of experiencing the event is modeled as a linear combination of the patient's covariates, $\hat{h}(x) = \boldsymbol{Z}^T\gamma$. Nonlinear effects such as interactions between covariates are added through domain expertise or variable selection procedures. DeepSurv, however, learns complex and nonlinear relationships between patients' covariates and their event risk without explicitly modeling them. The patient's covariates are passed into the network as inputs and propagated through a series of hidden layers of a feed-forward neural network. When MFPCA is used to extract informative features from the longitudinal data, the MFPC scores are passed together with the baseline covariates as the network inputs. The output of the model is a single node which estimates the patient's log-risk, $\hat{h}(x)$, which accounts for nonlinear patterns by using multiple hidden layers and activation functions. The hazard function of the ith subject for DeepSurv takes the form $\lambda_i(t) = \lambda_0(t)\exp\left(\hat{h}(x)\right)$.

DeepSurv is trained on a loss function based on the Cox partial likelihood function. More specifically, the model minimizes the average negative log likelihood given by

$$l(\theta) = -\frac{1}{I_{\delta=1}}\sum_{i:\delta_i=1}\left(\hat{h}_\theta(x_i) - \log\sum_{j\in R(T_i^*)} e^{\hat{h}_\theta(x_j)}\right).$$

Here, $I_{\delta=1}$ is the total number of subjects who had the event, and $R(t)$ is the set of patients still at risk at time t. Additional details of DeepSurv are illustrated in Section S1.

Similar to the Cox model, DeepSurv represents the hazard of each patient with a single log-hazard value and assumes proportional hazards: that covariates are a multiplicative constant of the baseline hazard function. When the proportional hazards assumption is met, the MFPCA-DeepSurv model provides a straightforward way to model survival data using longitudinal variables. In the presence of nonproportional hazards, the Cox and DeepSurv models may not be appropriate.

2.2 | MATCH-Net

CNNs have been adopted for modeling sequential data such as time series and longitudinal data.[20-22] In the case of univariate longitudinal data, the convolutional operation is often illustrated as sliding a one-dimensional filter over the observed values. More specifically, a univariate longitudinal outcome $Y_{iq}(t_{ij})$ of length J_i is convolved with a filter of length ℓ by calculating the dot product as the filter is passed along the lengths of the sequence. This results in a new sequence of length $J_i - \ell + 1$. Alternatively, the length of the original sequence can be preserved by padding the ends of the sequence with zeros or repeating the first and last values. When $Y_{iq}(t_{ij})$ refers to the case of multiple longitudinal outcomes, the Q outcomes are stacked along a new dimension called the depth. They are convolved with a set of Q filters of the same length: one filter for each longitudinal outcome. Applying this set of filters yields Q sequences which are then summed together, resulting in a single sequence of length $J_i - \ell + 1$. This sequence produced from a convolution is often referred to as a feature map. For clarity, these steps are illustrated in Figure S2. Typically multiple sets of filters are applied resulting in multiple sequences or feature maps which are stacked along the depth dimension and processed by subsequent

convolutional layers. The use of multiple feature maps allows the model to learn different discriminative features, aiding in predicting the task of interest.

MATCH-net[14] is a CNN for incorporating multiple longitudinal variables for survival prediction. Several modifications are made to accommodate the characteristics of clinical longitudinal data. First, the CNN requires observations to lie on a grid of fixed time steps (ie, every 6 months). Medical data from clinical visits are typically sparse as patients may not be frequently observed or patients may miss appointments. Therefore, the observed data $Y_{iq}(t_{ij})$ is first rounded to the nearest time corresponding to an established grid. Missing values are imputed from the last observed measurement. This results in the input matrix $\boldsymbol{Y}_{I\times J\times Q}$, where J is the grid length needed to accommodate the largest observation time. Furthermore, to account for patterns of missingness that may be informative of the survival event, an indicator mask with the same dimensions as $\boldsymbol{Y}$ is used to denote the positions of the imputed values. The mask is propagated through auxiliary convolutional layers in parallel to the main longitudinal branch. After each convolutional layer, the outputs of the auxiliary convolutions are concatenated with the outputs of the main branch to indicate a notion of missingness.

The last convolutional layer is followed by a global average pooling operation which averages over the values of a feature map. The resulting vector with length equal to the number of feature maps is concatenated with other baseline covariates and propagated through a feed-forward neural network. MATCH-net directly predicts the failure probability or the probability of the event occurring, $F_i(t_{J+\tau}|t_J) = P(t_{J+\tau-1} \le T_i^* \le t_{J+\tau}|T_i^* > t_J, \boldsymbol{Y}_i^{(t_J)}, \boldsymbol{Z}_i)$ for some prespecified prediction window $(t_{J+\tau-1}, t_{J+\tau}]$, where $\tau \ge 1$ is the number of prediction windows. The final output of MATCH-net is a vector with the same length as the number of contiguous prediction windows of interest, $[\hat{F}_i(t_{J+1}|t_J),\ \hat{F}_i(t_{J+2}|t_J),\ \dots\ ]$. A softmax function is employed on the final output vector to constrain the failure probabilities across all predictions windows to sum to one. Applying softmax on a vector x with length τ is defined as $\text{Softmax}(x) = \exp(x)/\sum_{\tau'\in\tau}\exp(x_{\tau'})$. Often, we are interested in the survival probability, $P(T_i^* > t_{J+\tau}|T_i^* > t_J, \boldsymbol{Y}_i^{(t_J)}, \boldsymbol{Z}_i)$. This can be expressed in terms of the failure probabilities as $1 - \left(\hat{F}_i(t_{J+1}|t_J) + \cdots + \hat{F}_i(t_{J+\tau}|t_J)\right)$. A visualization of the MATCH-net model is given in Figure S3.

The loss function for MATCH-net is calculated by

$$\mathcal{L} = -\frac{1}{\eta}\sum_{i=1}^{I}\sum_{j=1}^{J_i}\sum_{k=1}^{\tau_i} r_{ij}\cdot\log[F_i(t_{j+k}|t_j)] + (1-r_{ij})\cdot\log[1-F_i(t_{j+k}|t_j)]\ ,$$

where $\eta = \sum_{i=1}^{I}\sum_{j=1}^{J_i}\tau_i$ and $\tau_i = \min(J_i - j, \tau)$ which accounts for when the event or censoring occurs prior to the last prediction window. Lastly, $r_{ij} = 1$ if the event is observed at t_{ij}, otherwise $r_{ij} = 0$.

2.3 | TransformerJM

The transformer[15] is a neural network originally developed for natural language processing tasks such as translation. It is a sequence-to-sequence model that takes a sequence of elements (ie, words in a sentence) and transforms it into another sequence (a translated sentence). Here, we propose a transformer-based architecture named TransformerJM which takes a sequence of longitudinal measurements and predicts sequences of the future longitudinal and survival trajectories. Modeling longitudinal outcome data are analogous to natural language processing. Specifically, each visit is a vector containing Q longitudinal outcomes which corresponds to a *word*. The sequence of J_i visits from a subject can be viewed as a *sentence*. The data is composed of I subjects or a number of *sentences*.

A key tenet of the transformer is the attention mechanism. Attention is a function that takes a sequence and compares it to a second sequence, identifying areas with matching or relevant context. When attention is used to compare a sequence to itself, it is referred to as self-attention. This allows each element in the sequence to be represented based on other parts of the same sequence. This is useful in natural language processing as the meaning of a word can change or become more specific when placed in the context of the entire sentence. Likewise, although the same set of measurements could be observed in two patients at a given visit, their disease progression may not be the same. Prior visits should be taken into consideration to see if the overall condition of each patient is improving or declining. Through self-attention, the model is able to learn the dependencies between visits and build up a more complete profile of the patients' conditions. Several modifications made to the attention mechanism are used in the transformer. In particular, an important distinction when applying self-attention to longitudinal data is the need to derive context from only the past and not from future visits which have not yet occurred. Removing inappropriate connections from the attention calculation is called masked self-attention. In addition, attention can be calculated multiple times in parallel to give several different representations of the data. This

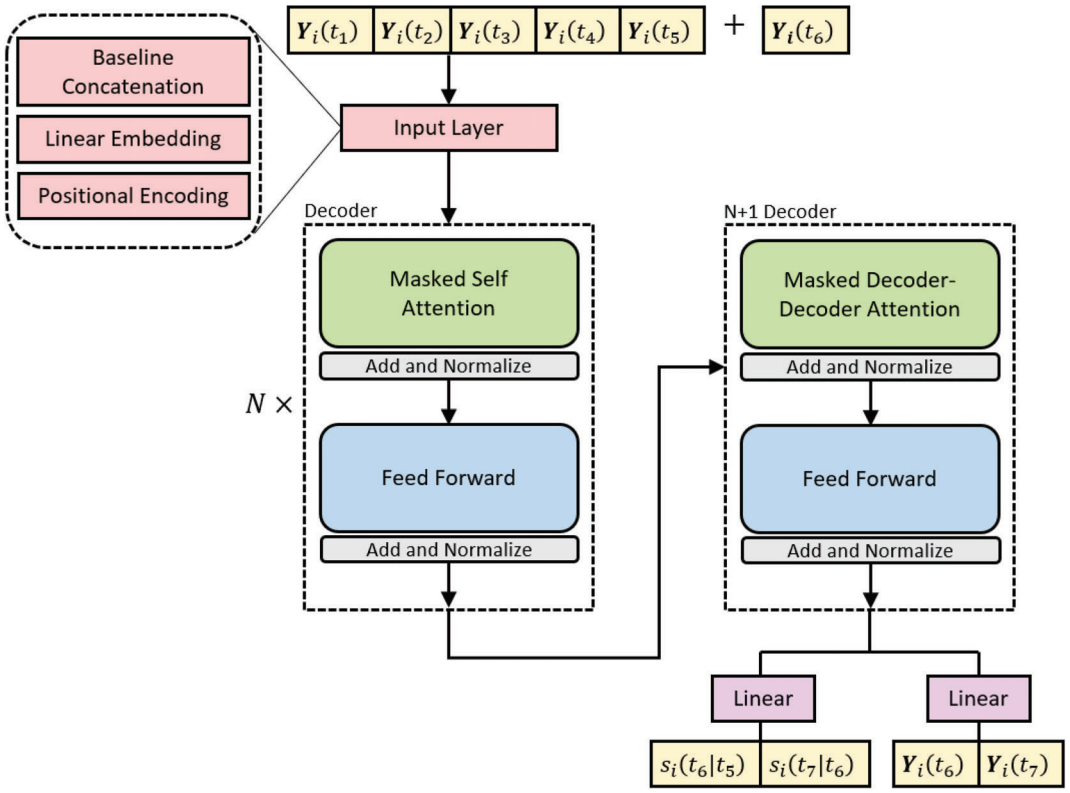

FIGURE 1 Architecture of the TransformerJM model. Given longitudinal outcomes observed for five visits $[\boldsymbol{Y}_i(t_1), \ldots, \boldsymbol{Y}_i(t_5)]$, TransformerJM predicts the longitudinal outcomes and survival probabilities for visit 6 $[\boldsymbol{Y}_i(t_6), s_i(t_6|t_5)]$. The predictions can be fed autoregressively back into the model to obtain predictions for visit 7 $[\boldsymbol{Y}_i(t_7), s_i(t_7|t_6)]$

is referred to as multihead attention. While a single attention head finds relationships between visits, multihead attention allows each attention head to focus on a more nuanced type of relationship. For example in the repeated measures from an individual patient, a prior visit may be relevant to a current visit because the longitudinal outcomes are close in value, or they may be related because they occurred close in time to each other. Providing the model with different notions of how visits are related allows for more refined representations of the patient's disease status. A final remark on the transformer is its use of positional encodings. The attention mechanism learns dependencies between all allowable time points; however, it does not account for the actual time at which the visit takes place. One solution to give the transformer a sense of order is to add a piece of information to each visit denoting the time it took place. While there are several options for positional encodings, a common choice for transformers is to represent the time using sine and cosine functions of different frequencies. The technical details of positional encoding and the different forms of attention are reviewed in Section S3.

In Figure 1, we outline the architecture of TransformerJM and illustrate how an example input sequence of $J_i = 5$ visits is passed through the network. Let $\{\boldsymbol{Y}_i(t_j)\}_{j=1,\ldots,5}$ refer to the sequence of vectors representing the disease status of patient i at a given visit j. Note that the form of $\boldsymbol{Y}_i(t_j)$ will change as it is passed through the network. $\boldsymbol{Y}_i(t_j)$ initially refers to the vector of Q longitudinal outcomes observed at visit j. The input sequence is processed through several steps. First, the baseline variables are concatenated to each $\boldsymbol{Y}_i(t_j)$. Each vector is then passed through a feed-forward neural network to learn interactions between the elements of $\boldsymbol{Y}_i(t_j)$ (linear embedding). Lastly, positional encoding is used to tie the observation times to each $\boldsymbol{Y}_i(t_j)$. At this point, each $\boldsymbol{Y}_i(t_j)$ is derived from only the data observed during visit j. For instance, $\boldsymbol{Y}_i(t_3)$, is a vector derived from the baseline covariates and the longitudinal outcomes observed at the third visit.

TransformerJM is composed of a stack of $N+1$ "decoder" blocks, where N is a hyperparameter to be selected. The first N decoder blocks consists of two sublayers: a masked multihead self-attention mechanism, followed by a feed-forward neural network. Self-attention allows each $\boldsymbol{Y}_i(t_j)$ to give a fuller representation of the patient's condition by incorporating context from prior visits, while the feed-forward layers further integrates this information. This means that the patient's disease status at the third visit, $\boldsymbol{Y}_i(t_3)$, now takes into account the information from the preceding two visits. The output

of the final decoder block is still a sequence of vectors $\boldsymbol{Y}^* = \{\boldsymbol{Y}_i(t_j)\}_{j=1,\dots,5}$. However, now each $\boldsymbol{Y}_i(t_j)$ has incorporated contextual information from prior visits. Note that $\boldsymbol{Y}^*$ represents the patient's condition for the observed visits, but in the example we are interested in predicting the patient's state at the next visit, $j = 6$. To make a prediction for the next visit, the positional encoding of the desired prediction time, t_6, is added to the vector of the last visit, $\boldsymbol{Y}_i(t_5)$, and passed to a final "N+1" decoder block. This decoder block differs in that it calculates the attention between the prediction time encoded vector and $\boldsymbol{Y}^*$. The result is a single vector predicting the patient's state at time t_6. This output is passed to two branches of linear layers. The first branch predicts $\boldsymbol{Y}_i(t_6)$, the Q longitudinal outcomes at time t_6. The second branch predicts $s_i(t_6|t_5)$, the conditional survival probability at time t_6. During evaluation, the predicted longitudinal measures are fed back into the beginning of the model as inputs to auto-regressively predict additional future time points. In the example, the predicted values $\boldsymbol{Y}_i(t_6)$ may be added back to the input sequence to generate the predictions, $\boldsymbol{Y}_i(t_7)$ and $s_i(t_7|t_6)$.

To predict both the longitudinal outcomes and the survival probability, the model is trained to jointly minimize two loss functions. The longitudinal loss is given by the mean squared error, while the survival loss is given by the negative log likelihood function.

$$\mathcal{L}_{\text{long}} = \frac{1}{\eta_\ell} \sum_{i=1}^{I} \sum_{j=1}^{J_i} \sum_{q=1}^{Q} [\hat{Y}_{iq}(t_{ij}) - Y_{iq}(t_{ij})]^2$$

$$\mathcal{L}_{\text{surv}} = \frac{1}{\eta_s} \sum_{i=1}^{I} \sum_{j=1}^{J_i} (1 - r_{ij}) \cdot \log(\hat{s}_{ij}) + r_{ij} \cdot \log(1 - \hat{s}_{ij}) \ ,$$

where $\eta_\ell = \sum_{i=1}^{I} \sum_{q=1}^{Q} J_i$ and $\eta_s = \sum_{i=1}^{I} J_i$. In the survival loss, $\hat{s}_{ij}$ is shorthand for $\hat{s}_i(t_{ij}|t_{i,j-1})$, the estimated conditional survival probability at visit j. If the event is observed at t_{ij} then $r_{ij} = 1$, otherwise $r_{ij} = 0$. The total loss is given by the sum of the two loss functions.

2.4 | Model evaluation

The performance of MFPCA-Cox, MFPCA-DS, MATCH-net, and TransformerJM were measured in the context of dynamic prediction which considers how predictions are updated when new longitudinal information becomes available. To validate the models' performance, subjects were split into training and testing sets. The models were first trained using all available longitudinal data. During testing, several landmark times of clinical interest, t, were evaluated. Importantly, at each landmark time, predictions were made using only longitudinal data up to time t, denoted $\boldsymbol{Y}^{(t)}$, for the subjects who were still event-free. Let Δt be a fixed prediction window, and $t' = t + \Delta t$ be a future time point of interest. Then the conditional survival probability for a subject i is defined as $\hat{s}_i(t'|t) = p(T_i^* \geq t' | T_i^* \geq t, \boldsymbol{Z}_i, \boldsymbol{Y}_i^{(t)})$. That is, the predicted risk of an event occurring in the interval $(t, t']$ is the conditional probability of the subject being event-free at time t' given that they were event-free at time t and taking into account baseline covariates and longitudinal observations up to time t.

The models were evaluated on the basis of both discrimination and calibration. Discrimination indicates how well a model is able to differentiate between subjects who experienced the event and subjects who do not. Discrimination was measured using the the time-dependent area under the ROC curve (AUC) for right censored data.[23,24] Calibration measures the agreement between predicted risks and true risks. The dynamic expected Brier score[25] was used as a second metric and accounts for both discrimination and calibration. The Kaplan–Meier estimator was used to obtain the inverse probability of censoring weights to adjust for right censoring. Strong predictive performance is indicated by a high AUC and low Brier score. Because AUC and Brier score metrics were calculated for each landmark time, we reported an integrated AUC (iAUC) and integrated Brier score (iBS) by integrating over the landmark times to further summarize the results. The integrated AUC is given by $\text{iAUC} = \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} \text{AUC} \ dt$, where $t_{\min}$ and $t_{\max}$ are the minimum and maximum landmark times. Likewise, the integrated Brier score is given by $\text{iBS} = \frac{1}{t_{\max} - t_{\min}} \int_{t_{\min}}^{t_{\max}} \text{BS} \ dt$.

In addition to the survival predictions, predictions were made for the longitudinal outcomes using MFPCA and the TransformerJM model. Note that MATCH-net only predicts survival probabilities and was not included in this comparison. For each observed time occurring after baseline, the values for the longitudinal outcomes were predicted using all the longitudinal data up to the prior visit. That is, $\hat{Y}_{iq}(t_{ij}) = f\left(Y_{iq}(t_{i0}) \ \dots \ Y_{iq}(t_{i,j-1})\right)$ for $j \geq 1$, where f represents the longitudinal prediction of MFPCA or TransformerJM. The longitudinal predictions were assessed using the root mean squared error (RMSE) where the mean squared error was defined as $\text{MSE} = \frac{1}{I} \sum_{i=1}^{I} \frac{1}{Q} \sum_{q=1}^{Q} \frac{1}{J_i} \sum_{j=1}^{J_i} [\hat{Y}_{iq}(t_{ij}) - Y_{iq}(t_{ij})]^2$.

3 | SIMULATION STUDY

Model performance was evaluated through several simulation settings. We created 100 simulated datasets, each with a sample size of $I = 1000$ subjects. Each subject had $P = 2$ baseline covariates and $Q = 3$ longitudinal outcomes with a maximum of $J_i = 11$ annual visits from observation times $t_{ij} = [0, 10]$. We refer to this initial simulation setting as Scenario 1. To demonstrate the effects of model misspecification, an interaction term was added to the baseline covariates when creating the simulated data. When training the models, the interaction term was not specified as a predictor. This simulation setting is referred to as Scenario 2 and evaluates the models' ability to capture the effect of the interaction term without needing explicit input. Lastly, in Scenario 3, a time-dependent covariate was used to create a simulation setting with nonproportional hazards. The details of the data simulation process for each scenario are given in Section S4.1. The code for the simulation study is provided in Section S4.2.

For each simulated dataset, 70% of the data was randomly selected for training and the remaining 30% was set aside for testing using dynamic prediction. In our analysis, we made predictions at landmark times $t = [1, 2, 3, 4, 5]$. At each landmark time, predictions were made $\Delta t = 1$ into the future.

Table 1 presents the true iAUC and iBS values and compares them to the iAUC and iBS of each model. In Scenario 1, MFPCA-Cox performs extremely well as the survival data was simulated from a Cox model. However, the other methods also have strong performance with high iAUC and low iBS close to the true values. In Scenario 2, adding an unspecified interaction term when simulating the survival data contributes to a significant drop in iAUC and a markedly higher iBS for MFPCA-Cox, which is unable to implicitly consider interaction terms. In contrast, the neural network layers of MFPCA-DS, MATCH-net, and TransformerJM are able to sufficiently account for interactions that occur between covariates. Therefore, these methods maintain their high iAUC and low iBS values. Lastly in Scenario 3, when the proportional hazards assumption is not met, MFPCA-Cox and MFPCA-DS have iAUC and iBS values markedly different from the truth. In contrast, TransformerJM excels in this scenario with iAUC and iBS close to the true values. Although MATCH-net also accounts for nonproportional hazards, TransformerJM performs better in Scenario 3. This may be attributed to the autoregressive approach of TransformerJM which is more suitable for predicting the sequence of survival probabilities. In comparison, MATCH-net outputs the sequence of survival probabilities using a softmax function, which is typically used for classification of nominal outcomes. In Table S5, additional simulation results are presented for predictions made at

TABLE 1 Integrated AUC (iAUC), integrated Brier score (iBS), and root mean square error (RMSE) calculated for three scenarios: baseline (Scenario 1), unspecified interactions (Scenario 2), and non proportional hazards (Scenario 3). The iAUC and iBS evaluates the survival predictions made $\Delta t = 1$ into the future at landmark times $t = [1, 2, 3, 4, 5]$. The RMSE evaluates the predictions of the observed longitudinal outcomes, using all available longitudinal data leading up to each observation time

(a) Integrated AUC					
	True iAUC	**MFPCA-Cox**	**MFPCA-DS**	**MATCH-net**	**TransformerJM**
Scenario 1	0.915	0.912	0.909	0.907	0.909
Scenario 2	0.917	0.862	0.910	0.908	0.910
Scenario 3	0.940	0.847	0.871	0.917	0.930
(b) Integrated Brier Score					
	True iBS	**MFPCA-Cox**	**MFPCA-DS**	**MATCH-net**	**TransformerJM**
Scenario 1	0.060	0.065	0.065	0.069	0.063
Scenario 2	0.053	0.079	0.059	0.061	0.057
Scenario 3	0.059	0.105	0.092	0.092	0.066
(c) RMSE					
		MFPCA		**TransformerJM**	
Scenario 1		0.120		0.144	
Scenario 2		0.121		0.144	
Scenario 3		0.125		0.149	

$\Delta t = 2$ into the future. These results affirm the conclusions drawn for the predictions at $\Delta t = 1$. They demonstrate that the predictions made by the models at later prediction windows maintain good discrimination and calibration as measured by the iAUC and iBS.

Predictions were also made for the three longitudinal outcomes using MFPCA and TransformerJM. The predictions for any given visit were made using all available longitudinal data leading up to the visit. The RMSE was calculated for each of the 100 simulated datasets and averaged together. Table 1(c) shows that both methods have low RMSE with MFPCA having slightly lower RMSE compared to TransformerJM in the three scenarios.

One additional advantage of using MFPCA to model the longitudinal outcomes is the computational time, with the two-stage MFPCA models finishing the simulation significantly faster than the end-to-end neural network methods. Taking the first simulation scenario as an example, the MFPCA-DS model took 30 min to finish 100 replications of the simulation on a desktop computer (8GB RAM, 3.5 GHz CPU). In comparison, it took around 4.5–5 h for MATCH-net and TransformerJM to finish the same task. This may be of consideration if computation power is a limiting factor.

4 | APPLICATION TO AD

The described methods were applied to the Alzheimer's Disease Neuroimaging Initiative (ADNI) and the National Alzheimer's Coordinating Center (NACC) datasets. ADNI is a longitudinal multicenter study that collects clinical, imaging, genetic, and biological biomarkers to better track the progression of AD. ADNI is currently on its third phase, with previous phases being ADNI 1, Go, and 2. In our analysis, we focused specifically on MCI patients as they are more susceptible to AD conversion. We included baseline MCI patients with no missing baseline covariates from all phases of ADNI, including those in ADNI 3 (data extracted September 8, 2020). In addition, visits were only included when all longitudinal variables were observed during that visit. This resulted in a total of 1068 MCI patients with 331 progressing to AD. The mean time-to-event or censoring time was 3.33 years (SD: 2.57). Visits were scheduled every 6 months for the first 2 years with annual visits thereafter. Patients had on average 4.84 visits (SD: 2.51, range: 1-14).

We included age, gender, years of education, and the number of apolipoprotein E4 (APOE4) alleles as baseline demographic variables. In addition, we incorporated seven longitudinal clinical assessments shown to be informative predictors of AD progression in previous research.[8] These are the ADAS-Cog13, the Mini Mental State Examination (MMSE), the Functional Assessment Questionnaire (FAQ), the Clinical Dementia Rating Sum of Boxes (CDRSB), and the RAVLT (immediate, learning, and forgetting scores). Most of these variables are neuropsychological assessments which aims to measure the rate of cognitive decline. The exception is FAQ, a functional assessment which evaluates the patient's capabilities in performing normal everyday tasks. All outcomes were normalized using min-max scaling to have values between -1 and 1.

We considered landmark times at $t = [1, 2, 3, 4]$ years and made survival predictions one year after each landmark time. Predictions of the patients' survival risk were made through 10-fold cross-validation with the predicted risk from each fold concatenated together prior to calculating the AUC and Brier score metrics. The iAUC and iBS measures are displayed in Table 2(a). The results show good performance across all models with high iAUC and low iBS.

We examined a second AD dataset provided by the NACC, which was established by the National Institute of Aging (NIA). The NACC is a collection of NIA-funded Alzheimer's Disease Research Centers (ADRC) across the United States and maintains a cumulative database to which each ADRC contributes. Each ADRC enrolls patients through physician referral, self-referral, or active recruitment and diagnoses patients according to their own protocol. In comparison, ADNI follows a specific recruitment and diagnostic protocol for all patients resulting in a more homogeneous study group.

The NACC dataset was preprocessed in the same manner as the ADNI dataset (data extracted October 16, 2019). MCI Patients were included if they had no missing baseline covariates. Only visits with no missing longitudinal variables during that visit were included. In total 5261 MCI patients satisfied these requirements, out of which 1516 progressed to AD. The mean time-to-event or censoring time was 2.76 (SD: 1.90). Visits were typically scheduled annually with patients having an average of 3.35 visits prior to receiving the event (SD: 1.65, range: 1-9). Age, gender, years of education, and the number of apolipoprotein E4 (APOE4) alleles were selected as baseline demographic variables. However, because the NACC does not contain all the longitudinal outcomes present in ADNI, we selected longitudinal clinical assessments based on the significance of an univariate joint model analysis, similar to the procedure described in Li et al.[8] The 10

TABLE 2 Integrated AUC and Brier score for the Alzheimer's disease neuroimaging initiative (ADNI) and National Alzheimer's Coordinating Center (NACC) datasets. One year predictions were made at landmark times 1, 2, 3, and 4 years

(a) ADNI				
	MFPCA-Cox	**MFPCA-DS**	**MATCH-net**	**TransformerJM**
iAUC	0.881	0.888	0.895	0.890
iBS	0.077	0.083	0.070	0.062
RMSE	0.211		—	0.214
(b) NACC				
	MFPCA-Cox	**MFPCA-DS**	**MATCH-net**	**TransformerJM**
iAUC	0.794	0.790	0.825	0.812
iBS	0.102	0.107	0.099	0.080
RMSE	0.225		—	0.243

most significant outcomes were selected for model building. The significant outcomes from the NACC dataset are the MMSE, the Clinical Dementia Rating Sum of Boxes (CDRSB), Logical Memory (immediate and delayed, respectively), the Wechsler Adult Intelligence Scale - Digit Symbol (WAIS-R), the Semantic Verbal Fluency Test (animal and vegetable categories, respectively), the Trail Making Test parts A and B, and the Boston Naming Test (30-items). All outcomes were normalized using min-max scaling to have values between −1 and 1.

Each model was trained with the NACC dataset using 10-fold cross-validation. The iAUC and iBS are presented in Table 2(b). MATCH-net and TransformerJM have better discrimination, with higher iAUC than MFPCA-Cox and MFPCA-DS. While MATCH-net has higher iAUC than TransformerJM, TransformerJM has markedly lower iBS, indicating better overall performance. Figure S4 plots the 95% confidence intervals for the difference in AUC and Brier score across different landmark times for TransformerJM vs the other models. In the NACC dataset, the difference in the AUC or BS for TransformerJM in comparison to MFPCA-Cox and MFPCA-DS is significantly different from zero at landmark times 1, 2, and 3. In the ADNI dataset, the differences were not significantly different from zero. This may suggest that while MFPCA-Cox and MFPCA-DS perform well on the ADNI dataset, MATCH-net and TransformerJM are better suited for the NACC dataset which contains more heterogeneous data. While the results of the two datasets cannot be directly compared due to different variables and sample sizes, the greater heterogeneity of the NACC dataset may also explain why the results for the NACC dataset produced lower iAUC and higher iBS and RMSE when compared to the results of the ADNI dataset. Although all ADRCs collect a common set of clinical observations, each have their own research focus and recruitment protocol. In addition, NACC patients are drawn from a broader population. For example, in the ADNI dataset, 94% of patients were white and less than 3% were African American. In contrast, in the NACC dataset, 82% of patients were white and 12% were African American. This is particularly important as prior studies have shown that AD has a higher prevalence among the African American community.[1,26,27] Although the greater heterogeneity of the NACC dataset increases the difficulty of prognostication, the results are more applicable to the general population and provides another valuable perspective on AD.

Predictions were also made for the longitudinal outcomes using MFPCA and TransformerJM. The RMSE was calculated from normalized values of the outcomes to allow for direct comparison. The results of TransformerJM are comparable to the benchmark set by MFPCA. While the RMSE of MFPCA appears to be slightly lower, TransformerJM remains very competitive, providing effective predictions of the longitudinal outcomes in addition to the survival probabilities. This is further illustrated in Figure S5, which compares the RMSE of MFPCA and TransformerJM for individual longitudinal outcomes.

To demonstrate the ability of TransformerJM to dynamically update predictions as new data are collected, individualized predictions were made for two ADNI patients in Figure 2. The mean trajectories of selected longitudinal outcomes and the survival probabilities are plotted with their corresponding pointwise 95% confidence intervals, which are generated by bootstrapping the data 100 times. For the longitudinal outcomes, 1-year predictions ($\Delta t = 1$) are displayed for each visit using all the accumulated data up to that point. The survival probabilities, $P(T > t + \Delta t | T > t)$ are presented for two landmark times t and $\Delta t = 1, 2, 3, 4$. The longitudinal predictions for Patient A matches up closely to the true observed values, with the 95% confidence intervals nearly always covering the observed values. In particular, the model is

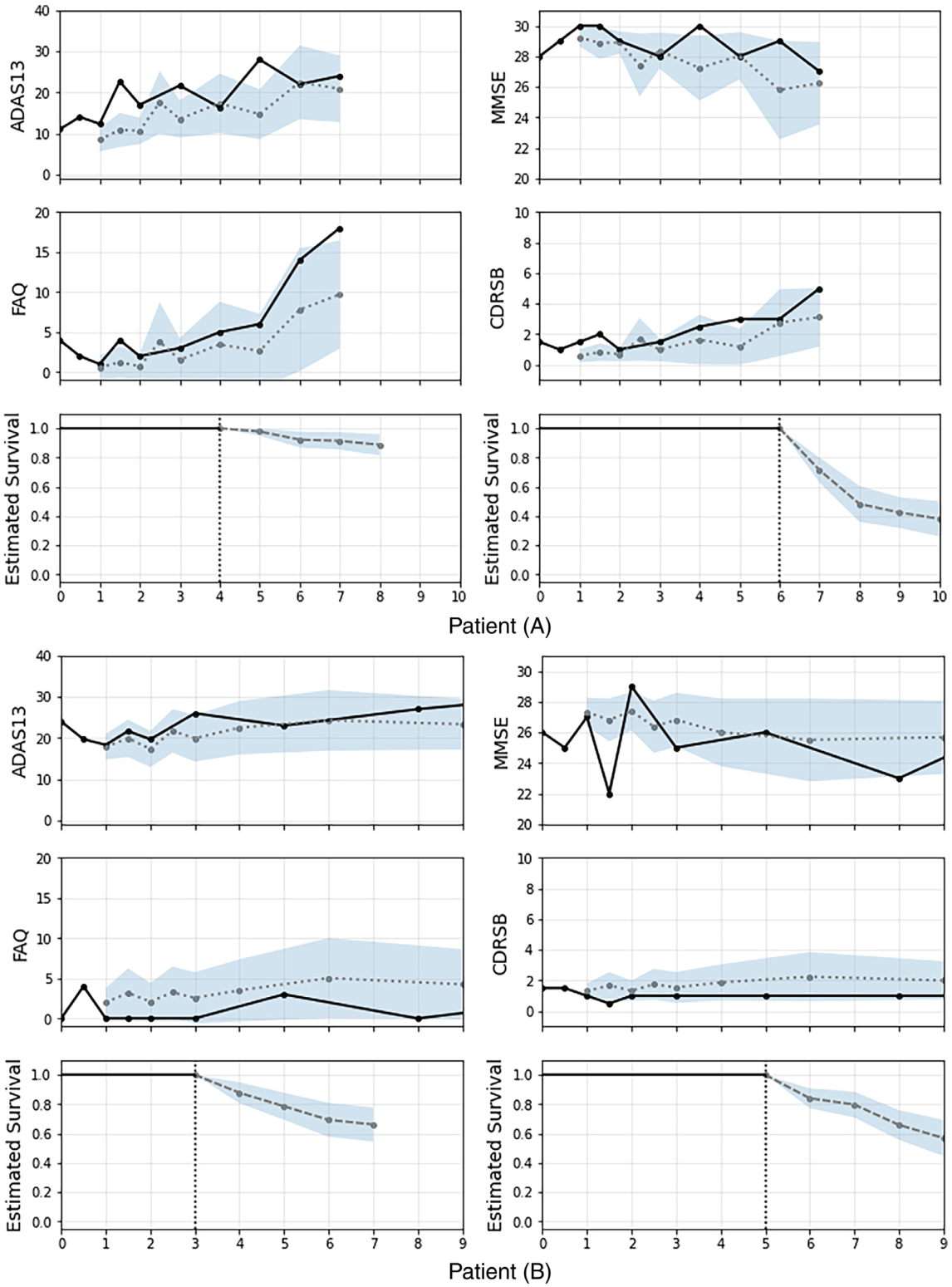

FIGURE 2 Individualized predictions were made for two Alzheimer's Disease Neuroimaging Initiative patients using 100 bootstrap samples of the data. One-year predictions ($\Delta t = 1$) were made for the longitudinal outcomes (dotted) and compared to observed values (solid). Predicted survival probabilities, $P(T > t + \Delta t | T > t)$, were made for $\Delta t = 1, 2, 3, 4$ at selected landmark times, t, marked with a vertical line. The mean trajectories are plotted with their corresponding pointwise 95% confidence intervals

able to account for the accelerating rise of FAQ and CDRSB measurements. The changes in the longitudinal outcomes are reflected in the survival predictions with the estimated survival probability dropping sharply after landmark year 6. This contrasts with the predictions made at year 4 which showed that the patient would remain in stable condition. Patient A was eventually diagnosed with AD by year 8, confirming the prognosis made by the model. This example demonstrates how the model adjusts its prediction in light of new data which reflected a downturn in the patient's condition. In contrast, Patient B is an example of a patient who remains in relatively stable condition. Because the model accounts for multiple longitudinal outcomes, the survival predictions at year 3 are not sensitive to the initial noisy measurements of MMSE due to the values of the other longitudinal outcomes. The prediction made at year 5 suggests that Patient B continues to remain on a similar risk trajectory after accounting for additional data.

5 | DISCUSSION

In this article, several methods for modeling time-to-event data using multiple longitudinal outcomes were presented. First, we described a functional data approach using MFPCA[10] to capture informative features from longitudinal outcomes. These features can be used directly as covariates in subsequent survival modeling. We demonstrated this pipeline using the Cox model and the DeepSurv neural network as possible survival models. Next, we reviewed MATCH-net,[14] a neural network which accounts for multiple longitudinal outcomes using convolutional layers. Lastly, we developed a novel transformer neural network, named TransfomerJM, to jointly model longitudinal and survival trajectories. Analysis on both simulated and real datasets showed excellent performance across all models in general. To help practitioners determine an appropriate model for their particular use case, we discuss the advantages and disadvantages of these models.

First, TransformerJM demonstrated predictive performance that matched or improved upon other state-of-the-art methods in both the simulation study and real data analysis. The results of the simulations showed that TransformerJM has better performance in terms of discrimination and calibration in two scenarios not handled by the traditional Cox model: nonlinear combinations of covariates and nonproportional hazards. Even when these two issues are not present, TransformerJM achieves AUC and Brier score metrics close to the true values. This advantage becomes especially evident when working with real life data as it is difficult to know which particular interaction terms should be included or the extent to which nonproportional hazards are present. In the same manner as previous dynamic prediction models, TransformerJM allows for predictions to be updated as patients return for follow-up visits. Using an individual patient from the ADNI dataset, we demonstrated how dynamic prediction can provide meaningful insight to a clinician as a patient is followed from year to year. One limitation of TransformerJM that we would like to address in our future work is the inability to handle asynchronous missingness (some longitudinal measurements may be missing at some visits). While TransformerJM can deal with irregular observation times (the time interval between consecutive visits may vary), it requires all longitudinal outcomes to be observed per visit. This requirement does not pose a large problem in our real data analysis using clinical assessments, as all relevant clinical assessments are typically carried out during the same visit. However, it may be an issue when considering multiple data modalities. For example, training a model using both clinical assessments and MRI data may be problematic as these two types of data are gathered on different visits. In this situation, we recommend using MFPCA or the convolutional dual-stream architecture of MATCH-net to handle the asynchronous missingness in the longitudinal data. Alternatively, an imputation step could be used to generate missing values prior to using TransformerJM.

Both MFPCA and TransformerJM can be used to obtain predictions of future trajectories of the longitudinal outcomes, while MATCH-net is designed to predict only the survival probabilities. In the results, both MFPCA and TransformerJM were able to predict the longitudinal outcomes with low MSE. While MFPCA had a slightly lower MSE than TransformerJM, either method would be a suitable choice when future trajectories of the longitudinal outcomes are needed in addition to the survival probabilities.

The practitioners may also give consideration to the computational requirements and data size. The two-staged MFPCA models are computationally fast to build in comparison to the end-to-end neural network models. It took MATCH-net and TransformerJM approximately 10-fold longer to train than either MFPCA-Cox or MFPCA-DS for the tasks described in this work. On the other hand, MFPCA requires the training data to fit in the computer's memory. This may not be feasible for many large-scale datasets. MATCH-net and TransformerJM, however, process the data in batches and do not require all the data to be loaded at once during training.

In both the simulation and real data analysis, patients were assessed at fixed observation times. However, the survival event could occur at any point between observation times, leading to interval censoring. One proposed method for handling interval censored data presented MLE and Continuous Ranked Probability Score loss functions which accounts for both right and interval censoring.[28] The method was applied to a feed-forward neural network for cross-sectional survival data as well as a RNN for survival data with longitudinal observations. The proposed loss functions could be extended to other neural networks for longitudinal and survival data, such as MATCH-net and TransformerJM, in the future.

An important future work is improving the interpretability and visualization of these models. As the models are able to incorporate a larger number of longitudinal outcomes, it is of interest to know which variables are more influential and how they contribute to the prediction. While MFPCA and the neural network methods can encode the longitudinal outcomes into informative features, it is challenging to assess the contribution of each outcome in the encoded features. Some methods have been proposed to visualize the input components that contribute the most to the neural network models when making predictions. Examples of these include saliency maps,[29] deconvolution,[30] and activation maximization.[29,31] These methods provide a good start to understanding the model behavior, however, they only highlight the areas of interest and do not explain why a model makes a certain decision. A recent idea in interpretable image classification is to have the model reason in a similar fashion to how humans recognize patterns by identifying comparable past examples.[32] For instance, radiologists often compare scans of a suspected tumor to images of known tumors in determining a cancer diagnosis. Extending these methods to longitudinal and survival prediction would allow clinicians to not just know what the prediction is, but also identify previous examples with matching patterns in certain longitudinal outcomes. Doing so would help determine important longitudinal outcomes and reveal the underlying reasoning that led to such a prediction. While we focused on the predictive aspect of the models, we believe that our work lays a strong foundation for which longitudinal and survival models can be used to help enhance decision-making in the clinical setting.

ACKNOWLEDGEMENTS

The research of Sheng Luo was supported by National Institute on Aging (grant number: R01AG064803, P30AG072958, and P30AG028716). Data collection and sharing for this project was funded by the ADNI (National Institutes of Health Grant U01 AG024904) and DOD ADNI (Department of Defense award number W81XWH-12-2-0012). The NACC database is funded by NIA/NIH Grant U24 AG072122. NACC data are contributed by the NIA-funded ADCs.

DATA AVAILABILITY STATEMENT

Data used in the preparation of this article were obtained from the ADNI database (adni.loni.usc.edu). The ADNI was launched in 2003 as a public-private partnership, led by Principal Investigator Michael W. Weiner, MD. The primary goal of ADNI has been to test whether serial magnetic resonance imaging (MRI), positron emission tomography (PET), other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and early Alzheimer's disease (AD). For up-to-date information, see https://www.adni-info.org. NACC is funded by the National Institute on Aging (U24 AG072122) and located in the Department of Epidemiology at the University of Washington School of Public Health. For up-to-date information, see https://naccdata.org/requesting-data/nacc-data.

ORCID

Jeffrey Lin https://orcid.org/0000-0003-0180-0117
Sheng Luo https://orcid.org/0000-0003-4214-5809

REFERENCES

1. Alzheimer's Association. Alzheimer's disease facts and figures. *Alzheimer's Dement*. 2021;2021(17):327-406.
2. Veitch DP, Weiner MW, Aisen PS, et al. Understanding disease progression and improving Alzheimer's disease clinical trials: recent highlights from the Alzheimer's disease neuroimaging initiative. *Alzheimer's Dement*. 2019;15(1):106-152.
3. Cummings J, Lee G, Ritter A, Sabbagh M, Zhong K. Alzheimer's disease drug development pipeline: 2020. *Alzheimer's Dement Transl Res Clin Intervent*. 2020;6(1):e12050.
4. Gauthier S, Reisberg B, Zaudig M, et al. Mild cognitive impairment. *Lancet*. 2006;367(9518):1262-1270.
5. Ward A, Tardiff S, Dye C, Arrighi HM. Rate of conversion from prodromal Alzheimer's disease to Alzheimer's dementia: a systematic review of the literature. *Dement Geriatr Cogn Disorders Extra*. 2013;3(1):320-332.
6. Rizopoulos D. Dynamic predictions and prospective accuracy in joint models for longitudinal and time-to-event data. *Biometrics*. 2011;67(3):819-829.

7. Rizopoulos D, Ghosh P. A Bayesian semiparametric multivariate joint model for multiple longitudinal outcomes and a time-to-event. *Stat Med*. 2011;30(12):1366-1380.
8. Li K, Chan W, Doody RS, Quinn J, Luo S. Prediction of conversion to Alzheimer's disease with longitudinal measures and time-to-event data. *J Alzheimer's Dis*. 2017;58(2):361-371.
9. Li K, Luo S. Dynamic prediction of Alzheimer's disease progression using features of multiple longitudinal outcomes and time-to-event data. *Stat Med*. 2019;38(24):4804-4818. doi:10.1002/sim.8334
10. Happ C, Greven S. Multivariate functional principal component analysis for data observed on different (dimensional) domains. *J Am Stat Assoc*. 2018;113(522):649-659.
11. Lin J, Li K, Luo S. Functional survival forests for multivariate longitudinal outcomes: dynamic prediction of Alzheimer's disease progression. *Stat Methods Med Res*. 2021;30(1):99-111.
12. Faraggi D, Simon R. A neural network model for survival data. *Stat Med*. 1995;14(1):73-82.
13. Katzman JL, Shaham U, Cloninger A, Bates J, Jiang T, Kluger Y. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network. *BMC Med Res Methodol*. 2018;18(1):24.
14. Jarrett D, Yoon J, van der Schaar M. Dynamic prediction in clinical survival analysis using temporal convolutional networks. *IEEE J Biomed Health Inform*. 2019;24(2):424-436.
15. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need; 2017. arXiv preprint arXiv:1706.03762.
16. Song H, Rajan D, Thiagarajan JJ, Spanias A. Attend and diagnose: clinical time series analysis using attention models. Paper presented at: Proceedings of the 32nd AAAI Conference on Artificial Intelligence; 2018; New Orleans, LA.
17. Li Y, Rao S, Solares JRA, et al. BEHRT: transformer for electronic health records. *Sci Rep*. 2020;10(1):1-12.
18. Yao F, Müller HG, Wang JL. Functional data analysis for sparse longitudinal data. *J Am Stat Assoc*. 2005;100(470):577-590.
19. Happ C. MFPCA: multivariate functional principal component analysis for data observed on different dimensional domains; 2019. R package version 1.3-2.
20. Zheng Y, Liu Q, Chen E, Ge Y, Zhao JL. Time series classification using multi-channels deep convolutional neural networks. Proceedings of the International Conference on Web-Age Information Management; 2014:298-310; Springer, New York, NY.
21. Wang Z, Yan W, Oates T. Time series classification from scratch with deep neural networks: A strong baseline. Proceedings of the 2017 International Joint Conference on Neural Networks (IJCNN); 2017:1578-1585; IEEE, Anchorage, AK.
22. Chen W, Shi K. Multi-scale attention convolutional neural network for time series classification. *Neural Netw*. 2021;136:126-140.
23. Li L, Greene T, Hu B. A simple method to estimate the time-dependent receiver operating characteristic curve and the area under the curve with right censored data. *Stat Methods Med Res*. 2018;27(8):2264-2278.
24. Li L, Wu C. tdROC: nonparametric estimation of time-dependent ROC curve from right censored survival data; 2016. R package version 1.0.
25. Blanche P, Proust-Lima C, Loubère L, Berr C, Dartigues JF, Jacqmin-Gadda H. Quantifying and comparing dynamic predictive accuracy of joint models for longitudinal marker and time-to-event in presence of censoring and competing risks. *Biometrics*. 2015;71(1):102-113.
26. Green RC, Cupples LA, Go R, et al. Risk of dementia among white and African American relatives of patients with Alzheimer disease. *JAMA*. 2002;287(3):329-336.
27. Tang MX, Stern Y, Marder K, et al. The APOE-e4 allele and the risk of Alzheimer disease among African Americans, whites, and Hispanics. *JAMA*. 1998;279(10):751-755.
28. Avati A, Duan T, Zhou S, Jung K, Shah NH, Ng AY. Countdown regression: sharp and calibrated survival predictions. Proceedings of The 35th Uncertainty in Artificial Intelligence Conference; Vol. 115, 2020:145-155; PMLR, Tel Aviv, Israel.
29. Simonyan K, Vedaldi A, Zisserman A. Deep inside convolutional networks: visualising image classification models and saliency maps. Proceedings of the Workshop at International Conference on Learning Representations; 2014; Citeseer, Banff, Canada.
30. Zeiler MD, Fergus R. Visualizing and understanding convolutional networks. Proceedings of the European Conference on Computer Vision; 2014:818-833; Springer, New York, NY.
31. Selvaraju RR, Cogswell M, Das A, Vedantam R, Parikh D, Batra D. Grad-cam: visual explanations from deep networks via gradient-based localization. Proceedings of the IEEE International Conference on Computer Vision; 2017:618-626; Venice, Italy.
32. Chen C, Li O, Tao D, Barnett A, Rudin C, Su JK. This looks like that: deep learning for interpretable image recognition. *Adv Neural Inf Process Syst*. 2019;32:8930-8941.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Lin J, Luo S. Deep learning for the dynamic prediction of multivariate longitudinal and survival data. *Statistics in Medicine*. 2022;1-14. doi: 10.1002/sim.9392