Multi-state Survival Analysis using Pseudo value-based Deep Neural Networks

Md Mahmudur Rahman* Sanjay Purushotham†

Abstract

Multi-state survival analysis (MSA) uses multi-state models for analyzing time-to-event data collected from subjects who may transition to different states before experiencing the final event of interest over time. A key challenge in MSA is the accurate subject-specific prediction of multi-state model quantities such as transition probability and state occupation probability in the presence of censoring. Censoring is another crucial challenge in MSA, leading to the overestimation of multi-state model quantities. The traditional statistical multi-state models typically do not use covariates, which renders them infeasible for making subject-specific predictions. Moreover, they assume a strict Markov stochastic process while modeling transition probabilities along with proportional hazard or linear covariate effect assumptions - that may not hold in real-world data. The current MSA methods have not investigated the impact of different types of censoring on the multi-state model quantities estimation. Recently proposed state-of-the-art neural ordinary differential equation (ODE) models for MSA relax statistical assumptions, but they do not handle the censoring mechanism well. To fill the gap in the MSA literature, we propose a new class of pseudo-value based deep learning models for MSA, where we show that pseudo values - designed to handle censoring - can be a natural replacement for estimating the subject-specific multi-state model quantities when derived from Aalen-Johansen (AJ) or Landmark AJ consistent estimators. We systematically study our proposed models' performance under different censoring settings and when Markovianity or linearity assumptions get violated. Empirical results on both the simulated and real-world MSA datasets show that our proposed models perform better or comparably to existing MSA methods under various censoring settings.

Keywords: Multi-state survival analysis, pseudo values, censoring, neural networks

1 Introduction

Multi-state survival analysis (MSA) is the problem of analyzing time-to-event data using multi-state models (MSM). Multi-state models [20] are models of a continuous-time stochastic process that capture the movement of subjects among a finite number of healthy and/or disease states. Thus, multi-state modeling can provide insights into the disease progression by providing a detailed view of the disease or recovery trajectory in patients. This helps to predict the probability of future events after a given history and, thus, can improve the clinician's decision-making ability for survival analysis. Figure 1 shows an example of a multi-state model for breast cancer progression. Here, a patient who is disease-free or had surgery can transition to locoregional relapse or distant relapse before reaching the death state. Even though death is considered as the ultimate outcome in a study of survival analysis, disease relapse also indicates a treatment failure, thus, considered as an important outcome of the study [4]. Therefore, MSA is crucial when intermediate states are also outcomes of interest in a study.

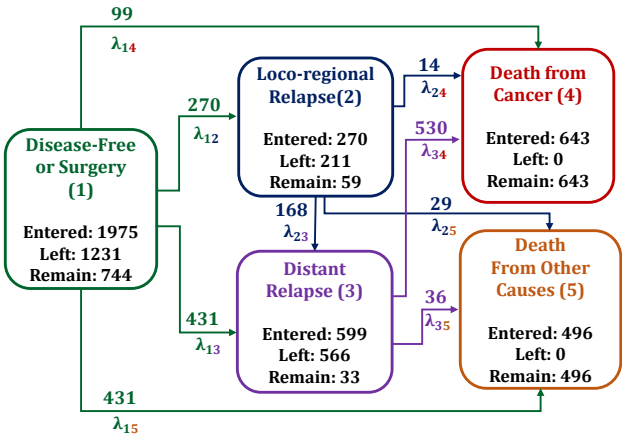

Figure 1: Transitions of patients from a Disease-free state to clinically relevant intermediate health states before experiencing one of the death states for all subjects. The rectangles represent states ($S = \{1, 2, ..., 5\}$), and the number along with arrows represents the number of transitions to a particular direction. *Entered* indicates the number of subjects entering each state, *Left* indicates the number of subjects leaving each state, and *Remain* indicates the number of subjects remaining at the end of follow-up. λ_{jk} denotes the transition intensity functions.

Multi-state survival analysis deals with the estimations of the multi-state quantities such as (a) **state oc-**

*University of Maryland, Baltimore County, USA. Email: mrahman6@umbc.edu
†University of Maryland, Baltimore County, USA. Email: psanjay@umbc.edu

cupation probability (SOP) (the probability that a subject will be in a state k at time t), (b) the **transition probability** (the probability of transition to a state k at time t from another state j at time s), and (c) **dynamic SOP** - which is the state occupation probabilities at some future time point t, given the event history (such as clinical information) is available up to a given time point s. The direct prediction of these probability-based quantities are desirable as they have simple interpretation, and thus, easily accessible to the researchers with limited mathematical background [5].

A variety of statistical and machine learning approaches have been developed for MSA, including non-parametric Aalen-Johansen (AJ) estimators [2], Cox-based semi-parametric methods [14], parametric multi-state methods [21], and neural network-based methods -SurvNODE- [18] to estimate these multi-state quantities. However, these existing MSA methods do not address one or more of these challenges: (a) **violation of linearity assumption:** Most existing MSA methods [14, 6, 21] assume a linear relationship between multi-state quantities and covariates, which rarely holds in a real-world scenario. These models, therefore, cannot capture the complex nonlinear representation from the multi-state data. (b) **violation of Markov assumption:** The Markov assumption, i.e., the transition intensities only depend on the past history via the current state, frequently made in many MSA methods, often get violated [24]. The violation of the Markov assumption leads to biased estimation of multi-state quantities; however, the assumption is rarely checked. (c) **presence of censoring:** Multi-state data are subject to censoring, which leads to an overestimation of the predictions of the outcome of interest [29]. Although it is essential to understand how different types and amounts of censoring impact the prediction of multi-state quantities, very few MSA works have studied the censoring impact. Furthermore, the existing MSA methods perform poorly for non-Markov data because finding consistent estimators for non-Markov data has been understudied in the literature. Thus, new approaches that overcome these issues for multi-state survival analysis are in great demand.

To address the above challenges and limitations of existing MSA methods, we introduce a new class of MSA models called pseudo value-based deep learning models for multi-state survival analysis, denoted as **msPseudo**. msPseudo estimates the multi-state quantities by treating the complex multi-state survival modeling as a regression analysis problem using pseudo values as response variables. (a) In order to learn complex nonlinear representations, msPseudo uses deep neural network architecture. In particular, msPseudo consists of a deep neural network, which takes covariates as inputs and estimates a multi-state quantity (such as state occupation probability) via a pseudo value prediction task. Furthermore, we propose **MT-msPseudo**, an extension of msPseudo, to perform a multi-task regression to jointly predict all the multi-state quantities using a single neural network model by leveraging correlations among task-specific pseudo values. (b) To efficiently handle censoring, our proposed models, msPseudo and MT-msPseudo, use pseudo values [28, 8] to replace the incompletely observed (censored) multi-state model quantities since they can be viewed as the contribution of a particular subject to the estimate of the multi-state quantities on a sample. We conduct extensive experiments on both simulated and real-world datasets to study how our proposed models perform under various censoring settings. (c) To address the non-Markovianity present in the real-world data, we introduce a simple algorithm to derive the pseudo values from consistent estimators such as AJ or Landmark AJ [32] estimators by testing the Markovianity of the data using statistical significance tests - Commenges-Andersen (CA) test [9] and log-rank statistic based tests [34]. Our algorithm provably obtains pseudo values from consistent estimators for both Markov and non-Markov data. We want to highlight that pseudo values derived from the Aalen-Johansen and Nelson–Aalen estimators used in earlier works [35, 28] are inconsistent for the real-world non-Markov data and can result in large estimation errors. Below, we briefly summarize our contributions:

- We propose first-of-its-kind, pseudo value-based deep learning models for multi-state survival analysis. Our proposed models handle censoring using consistent pseudo values and can obtain accurate subject-specific predictions of the multi-state quantities.
- We provide a simple algorithm to derive the pseudo values for multi-state quantities from consistent estimators. To the best of our knowledge, ours is the first work to estimate and predict the pseudo values for the transition probability using consistent estimators.
- We conducted extensive experiments on simulated and real-world datasets to show that our proposed models mostly outperform the state-of-the-art MSA methods. Furthermore, we empirically show that our proposed models achieve significant performance over other MSA methods under a variety of censoring settings for both Markov and non-Markov data.

2 Related Works

MSA has been well-studied in statistics and machine learning community. Statistical approaches such as non-parametric Aalen-Johansen (AJ) estimators [2] are commonly used for multi-state survival modeling. However,

AJ estimators have been shown to be inconsistent when the Markov assumptions are violated [13]. Therefore, researchers have proposed non-Markov estimators for non-Markov multi-state models [26, 13]. However, these methods make restrictive independent censoring support assumptions, which typically are untrue in medical applications (due to limited follow-ups). Recently, Putter et al. [27] proposed a non-parametric landmark AJ (LMAJ) estimator for non-Markov multi-state models and showed that the LMAJ is a consistent estimator and more efficient compared to other non-Markov estimators [33] while making no assumptions on the support of the censoring distribution. However, landmarking results in data reduction, which leads to a loss of power and undesirable condition for the less traveled Markov transitions [24]. All these non-parametric approaches do not use covariates for multi-state quantity estimation and, thus, cannot provide subject-specific predictions. Semi-parametric multi-state approaches such as Cox-based methods [14] fit a regression model on the covariates using a Cox Proportional Hazards (PH) model [11] or an additive hazard regression model [31] to make patient-specific multi-state quantity predictions. However, the PH assumption may be violated in real data because of the incorrect functional form for a covariate [23]. Some researchers have proposed linear parametric models based on pseudo observations derived from AJ estimators for the estimation of multi-state quantities [7, 16]. While these pseudo value-based approaches can handle censored observations, they are limited linearity and Markovian data assumptions.

In the ML community, neural network-based approaches [10, 18], which capture non-linear covariate effects without making any underlying assumptions, have achieved state-of-the-art results for multi-state survival analysis. [18] proposed SurvNODE to estimate state occupation probability using neural ordinary differential equations, and [10] proposed a transition-specific deep model for specifically solving a simple three-state illness-death model. However, these methods do not show any experimental results on transition probability predictions for non-Markov data and have not studied censoring impact on the model performance. Inspired by the promise of pseudo observations to handle censoring and deep learning-based approaches to capture non-linear co-variate effects, we propose pseudo value-based deep learning models for MSA.

3 Multi-state Survival Analysis: Overview

A multi-state process is a continuous time stochastic process $\{X(t), t \in \mathcal{T}\}$, taking values in the (discrete-state) finite state space $S = \{1, 2, ..., K\}$, where $\mathcal{T} = [0, \tau], \tau < \infty$. Multi-state survival analysis deals with the estimation of the multi-state quantities: transition probability (TP), state occupation probability (SOP), and dynamic SOP, which are described below.

Transition probability: The multi-state process is an interconnected network with a set of $K \times K$ possible transitions. This multi-state model can be understood by the *transition intensities* [32], $\lambda_{jk}(t|\mathcal{F}_{t^-})$ given by:

$$\lambda_{jk}(t|\mathcal{F}_{t^-}) = \lim_{dt \to 0} \frac{P(X(t+dt) = k | X(t) = j, \mathcal{F}_{t^-})}{dt}$$

$\forall j, k \in S : j \neq k$, and $\mathcal{F}_{t^-}$ is the history of the multi-state process prior to the time t. The **transition probabilities** are defined as

$$\mathcal{P}_{jk}(s, t|\mathcal{F}_{s^-}) = P(X(t) = k | X(s) = j, \mathcal{F}_{s^-}), \forall j, k \in S$$

and for two time points s and t, such that $s \leq t$. The $K \times K$ matrix of transition probabilities is referred to as the **transition probability matrix**, $\mathcal{P}(s, t)$, and is derived from a matrix product integral of the transition intensities [2] and is given by: $\mathcal{P}(s, t) = \prod_{s<u\leq t}(\mathbf{I} + d\mathbf{\Lambda}(u))$, where $\mathbf{\Lambda}(t)$ be a $K \times K$ matrix of cumulative transition intensities, with $(j, k)^{th}$ off-diagonal element $\Lambda_{jk}(t) = \int_0^t \lambda_{jk}(u)du$ and with $(j, j)^{th}$ diagonal element $\Lambda_{jj}(t) = -\sum_{k\neq j} \Lambda_{jk}(t), \forall j, k \in S$. When the multi-state model is Markov, the transition probabilities become $\mathcal{P}_{jk}(s, t) = P(X(t) = k|X(s) = j)$. In the non-Markov multi-state model, the transition probability $\mathcal{P}_{jk}(s, t|\mathcal{F}_{s^-})$ will depend on the past history prior to time s, $\mathcal{F}_{s^-}$.

State Occupation Probability (SOP): The state occupation probability (SOP) at time t for state k is defined as $\pi_k(t) = P(X(t) = k); k \in S$. Given an initial distribution at time 0, $\pi_j(0) = P(X(0) = j)$, we can obtain the SOP, $\pi_k(t)$, using the following equation: $\pi_k(t) = \sum_{j=1}^{K} \pi_j(0)\mathcal{P}_{jk}(0, t)$, where $\mathcal{P}_{jk}(0, t)$ is the transition probability from the initial state j to state k at time t. Matrix form of SOP is given by

$$\pi(t) = \pi(0) \prod_{0<u\leq t} (\mathbf{I} + d\mathbf{\Lambda}(u))$$

Dynamic SOP: In many instances, the clinical interest is to predict the state occupation probability for a particular state k for a patient at some future time points t given that the patient is alive at an earlier time s. This prediction task is referred to as dynamic state occupation probabilities (dynamic SOP) prediction [32]. The **dynamic SOP** for k^{th} state starting at time s can be defined as

$$\pi_k(t|s) = P(X(t) = k|\mathcal{F}_{s^-})$$

In this paper, the multi-state quantities are defined by conditioning on baseline covariates vector $\mathbf{Z}$. Thus, TP,

SOP, and dynamics SOP are respectively denoted as $\mathcal{P}_{jk}(s,t|\mathbf{Z}), \pi_k(t|\mathbf{Z})$, and $\pi_k(t|s,\mathbf{Z})$.

Estimators for multi-state model quantities: The Aalen-Johansen (AJ) estimator is a widely used non-parametric estimator for estimating SOP and TP and is given by $\hat{P}^{AJ}(s,t) = \prod_{s<u\leq t}(\mathbf{I}+\Delta\hat{\mathbf{\Lambda}}(u))$. Datta et al. showed that the AJ estimator is consistent for SOP for non-Markov models *(Theorem 2 in [12])*; however, Putter et al. [27] showed that it is inconsistent for TP for non-Markov data. So, Putter et al. proposed a landmark Aalen–Johansen (LMAJ) estimator following the same consistency criteria of the AJ estimator of the SOP for the estimation of TP and showed that it is consistent for a non-Markov model. LMAJ estimator for TP is given by $\hat{P}^{LMAJ}_{lm}(s,t) = \hat{\pi}^{(LM)}(s)\prod_{s<u\leq t}(\mathbf{I}+\Delta\hat{\mathbf{\Lambda}}^{(LM)}(u))$, where $\hat{\pi}^{LM}(s)$ is a $1\times K$ vector with $\hat{\pi}^{LM}_l(s)=1$ and 0 otherwise [27].

4 Our Proposed Pseudo value-based Models

We first describe the derivation of pseudo values for multi-state quantities before discussing our proposed pseudo value-based deep neural networks.

Pseudo values for multi-state quantities: Multi-state survival datasets are subject to censoring, i.e., incomplete information about the stochastic process (for example, event or transition information missing due to loss to follow-up). Therefore, direct modeling of the event time or status with respect to covariates is challenging for censored observations. Recent research has shown that pseudo values [3] can be used to handle censoring for survival analysis [35], competing risk analysis [28] and MSA [7]. Inspired by these works, we propose to use pseudo values as a substitute for the estimation of subject-specific multi-state model quantities in the presence of censoring. When there is no censoring, for all subjects in state j at time s, the indicator $Y_{ki}(t) = \mathbf{I}(X_i(t)=k)$ of being in state k at time $t>s$ can be observed and used as pseudo values for transition probabilities (TP). In the presence of censoring, the pseudo values for TP for subject i can be derived as

$$J_i = n_s\hat{P}_{jk}(s,t) - (n_s-1)\hat{P}^{-i}_{jk}(s,t) \tag{4.1}$$

where $\hat{P}_{jk}(s,t)$ is the consistent estimator for TP based on a sample of size n_s and the same estimator based on a leave-one-out sample, obtained by omitting the subject i. n_s could be the size of the full landmark dataset for the LMAJ estimator or the complete dataset or at risk (at time s) dataset for AJ estimator [5]. For SOP, we simply replace the estimator of TP $\hat{P}_{jk}(s,t)$ in 4.1 by $\hat{\pi}_k(t)$ to obtain the pseudo values for SOP.

Consistent pseudo value derivation via Markov assumption testing: Pseudo values can be derived from an unbiased and consistent estimator such as AJ estimator [34]. However, the AJ estimator for TP is inconsistent for non-Markov data and can result in large estimation errors. Thus, recently, researchers have proposed Landmark AJ (LMAJ) [27] as a consistent and robust estimator for estimating the pseudo value for non-Markov data. However, the AJ estimator is known to be more efficient than LMAJ [34] (when the Markov assumptions are valid), and in any practical scenario, the appropriateness of the Markov assumptions for a specific dataset remains unknown in advance, thus, making it infeasible to use just one estimator for pseudo value estimation. It is common to make Markov assumptions on the model; however, the assumption often may not hold in the real-world. Even though the violation of the Markov assumption can lead to biased estimation of multi-state quantities, the assumption is rarely checked [24]. To address this important challenge, we introduce and describe a pseudo value derivation algorithm, shown in **Algorithm 1**, to efficiently derive pseudo values by selecting consistent estimators by testing the underlying Markovian assumptions.

Algorithm 1 Pseudo value derivation algorithm

Inputs: Multi-state data, Selection of Commenges-Andersen (CA) test/log-rank test, ϵ
Output: Pseudo values for multi-state quantities
For SOP:
Choose either AJ estimator or LMAJ estimator to derive pseudo values since both are consistent and same quantity.
For Dynamic SOP:
Perform CA test.
if *P-value of the test is statistically significant for the violation of overall Markov assumption in the data* **then**
  Pseudo values ← LMAJ estimator
else
  Pseudo values ← AJ estimator
For TP:
Perform log-rank test to check transition-specific Markovianity.
if *P-value of the test is statistically significant for the violation of Markov assumption of a transition in the data* **then**
  if *landmark population* $< \epsilon$ **then**
    Pseudo values ← AJ estimator
  else
    Pseudo values ← LMAJ estimator
else
  Pseudo values ← AJ estimator

Our algorithm takes the multi-state survival data as input and selects a consistent estimator to obtain pseudo values by testing the Markovianity assumptions in the dataset by using statistical significance tests such as CA test [9], or log-rank statistic-based tests [34]. CA and log-tank tests use a test statistic (usually χ^2

statistic) and its corresponding p-values to identify the violation of Markov assumptions in the data. Please note that the choice of statistical tests- CA test or log-rank test- to check the Markovian-ity of data is still an important, open, and unresolved problem. CA test is known to be a more powerful test than the log-rank test only if applied to the subset of positive dependent transitions (i.e., $1 \rightarrow 2$, $1 \rightarrow 3$, and $2 \rightarrow 3$), while the log-rank tests are most useful in the scenario where the non-Markov behavior does not hold or is not uniform across the transition times of the patients. Note that the local log-rank test statistic can be undefined or unstable for time intervals in which very few patients are available in a particular qualifying state. In that case, we can use the stratified CA test to check the Markov assumption across different transitions. Similar to Titman et al. [34], we found (via experiments) that the nature of the datasets (whether Markov or not) dictates the choice of statistical test. If the Markovianity of data is unknown, LMAJ is preferable over AJ for consistency reasons; however, the AJ estimator is more efficient than LMAJ when the Markov assumption holds [34]. Theoretical analysis of the consistency of the AJ and LMAJ estimators can be found in [27]. As shown in Algorithm 1, we test Markovian assumptions via statistical tests and choose a consistent estimator for pseudo value derivation. In Algorithm 1, ϵ is chosen based on the minimum size of the population in a landmark state. Thus, Algorithm 1 provides a viable solution for a small-sized sample where LMAJ estimator is not applicable. We fix $\epsilon = 1$ in our experiments.

msPseudo model: We propose **msPseudo** - a first-of-its-kind, pseudo value-based deep learning model for multi-state survival analysis. msPseudo is a deep neural network trained to directly predict the multi-state quantities, such as state occupation probability (SOP), dynamic SOP, and transition probability (TP), using pseudo values as the response variables given the covariates. Our model captures the complex non-linear hidden relationship between the patient's characteristics, i.e., the baseline covariates and the multi-state model quantities. msPseudo is a simple feed-forward deep neural network with an input layer, multiple hidden layers, and an output layer. A $n \times p$ matrix of p baseline covariates with n individuals are fed as input to the input layer. The output layer of msPseudo returns the predictions of a multi-state quantity (SOP, dynamic SOP, or TP). For a multi-state dataset with K states, predicted output of SOP and dynamic SOP for a subject at a prespecified vector of M time points $\mathbf{t} = \{\tau_1, \tau_2, ..., \tau_M\}$ is a $K \times M$ matrix. In the TP prediction task, the output is a $Q \times M$ matrix, where Q is the number of transitions. A sigmoid activation function is used in the output layer to directly predict multi-state quantities such as SOP or TP. Note that pseudo values can be greater than 1 or less than 0. Thus, using a sigmoid activation function bounds the predicted values in the [0,1] range, similar to how the inverse logit link function works in a generalized linear model.

Training objective: As we formulated MSA as a regression problem, we can use the mean squared error [35] (the mean squared difference between pseudo values (ground truth) and the predicted multi-state quantity) loss function (L) to train our msPseudo model. The loss is given by:

$$L = \frac{1}{n * M * K} \sum_{k=1}^{K} \sum_{j=1}^{M} \sum_{i=1}^{n} (J_{ik}(t_j) - \hat{y}_{ik}(t_j|\mathbf{Z_i}))^2$$

Here, $J_{ik}(t_j)$ are the pseudo values for a multi-state quantity for state k (e.g., SOP or dynamic SOP) or for transition k (TP) at a pre-specified time point t_j. $\hat{y}_{ik}(t_j|\mathbf{Z_i})$ is the corresponding prediction given the covariates for i^{th} individual.

MT-msPseudo model: Instead of predicting the multi-state quantities separately using different msPseudo models, we can leverage the relatedness between the multi-state quantities and predict them jointly. Thus, we propose **MT-msPseudo** model as a natural extension of msPseudo. MT-msPseudo model uses three task-specific fully connected networks for the prediction of multi-state quantities and uses a fully connected shared subnetwork to learn the shared representation among the task-specific networks using covariates as inputs. The task-specific networks predict the three multi-state quantities, i.e., SOP, dynamic SOP, and TP, at M pre-specified time points. To train our MT-msPseudo model, we optimize a total loss function L_{Total} to jointly train the task-specific networks. The total loss is defined as: $L_{Total} = L_{SOP} + L_{dynSOP} + L_{TP}$, where L_{SOP}, L_{dynSOP}, and L_{TP}, respectively, are the mean squared error losses for predicting SOP, dynamic SOP, and TP. Algorithm 2 provides the pseudo code for learning the parameters of our proposed models.

5 Experiments

We conducted experiments on both simulated and real-world datasets to answer the following questions: (1) How do our proposed models compare against the existing multi-state survival models for the following prediction tasks: (a) SOP prediction, (b) Dynamic SOP prediction at a given time point s, (c) TP prediction? (2) Compared to baseline models, how well do our proposed models perform under linearity and Markov assumptions violation? (3) How well do our proposed models perform under different types and amounts of

Algorithm 2 Training of msPseudo or MT-msPseudo

Inputs : Observed covariates $\mathbf{Z}$, transition time and status, Pre-specified time points $t = 1, ..T$, learning rate η, mini-batch size M, the number of epochs E, early stopping patience p.

Output: Model weights Φ.

for $i \leftarrow 1$ **to** N **do**
  Pseudo values $J_i(t) \leftarrow$ Algorithm 1.
Create the training dataset, $D = \{\mathbf{Z_i}, \mathbf{J_i(t)}\}_{i=1}^{N}$.
Initialize Φ_0
for *epoch* $e \leftarrow 1$ **to** E **do**
  $B \leftarrow$ (split D into batches of size M)
  for *each batch* $b \in B$ **do**
    Update $\Phi \leftarrow \Phi - \eta \nabla L(\Phi; b)$
  if *early stopping counter reaches patience* p **then**
    stop training
  else
    continue
return Φ

censoring compared to other models?

Dataset descriptions:

Simulated datasets: We generated *linear and nonlinear time-homogeneous Markov* simulated datasets as well as *linear and nonlinear reversible Non-Markov* simulated datasets - to obtain different datasets with varying Markov and linearity assumptions. We simulated 5000 examples, each with multiple transitions for all the simulation datasets. The Markov datasets have three states and allow only forward transitions. On the other hand, the Non-Markov multi-state datasets consist of four states: states 1–3 are intermediate and interconnected states, and state 4 is an absorbing state, and they allow reverse transitions [19].

Real-world datasets: We used two publicly available datasets in our experiments - (1) *METABRIC [30]* contains 1975 breast cancer patient data with multiple transitions and 20 covariates. Data were collected during a 360-month study to determine breast cancer subgroups to help clinicians to provide better treatment. The multi-state dataset has five states: Disease-free or Surgery, Locoregional Relapse, Distant Relapse, Death from Cancer, and Death from other causes (shown in figure 1. (2) *EBMT* dataset contains 2279 transplantation patient data collected during 1985-1998 and can be used to study the survival of the patients who had blood cancer after the transplant treatment [15]. In this dataset, an alive patient in remission without recovery or adverse event can move to three possible distinct intermediate states, i.e., recovery, adverse event, and co-occurrence of recovery and adverse event, until one of the two absorbing states, death, and relapse, is observed.

Censoring settings: We investigate the impact of censoring on MSA model performances under these two censoring settings: incremental and induced censoring settings on nonlinear time-homogeneous Markov dataset and nonlinear reversible non-Markov dataset. In **Incremental censoring** setting, we incrementally add 25%, 50%, and 75% censored observations from the censored set to a fixed number of uncensored observations. In **Induced censoring** setting, we induce 25%, 50%, and 75% censored observations by flipping the label of the transition status of the selected uncensored observations. In addition, we incrementally add and induce the same amount of censoring (25%, 50%, and 75%) on each transition. These censoring settings help us to investigate the impact of (a) an increase in censoring, (b) an increase in censoring ratio (on a fixed-size dataset), and (c) censoring on a specific transition on the model performance.

Prediction tasks: Given the covariates, we perform regression tasks for estimating the multi-state quantities such as SOP, dynamic SOP, and TP. We compare the performances of the following multi-state models for these prediction tasks: **Non-parametric models:** LMAJ estimator (`LMAJ`) [27]; **Parametric models:** Weibull parametric model (`msWeibull`) [21], linear Pseudo value model (`LinearPseudo`) [6]; **Semi-parametric model:** Multi-state Cox proportional hazard model (`msCox`) [14]; **Deep learning multi-state model:** `SurvNODE` [18]; **Our proposed models:** `msPseudo` and `MT-msPseudo`.

Evaluations: We evaluate all the models in terms of a calibration measure, **integrated Brier score (iBS)** [32], and a discrimination measure, **integrated AUC (iAUC)** [17]. We conduct experiments on 5 sets of 5-fold cross-validation and report the mean and standard deviation of these evaluation metrics. Our proposed deep learning models use Adam optimizer [22] and are trained to 10000 epochs with an early stopping criterion. Hyperparameter tuning (over batch size, learning rate, drop-out, activation functions, number of layers, etc.) is performed to choose the best-performing deep learning models. A sigmoid activation function is used in the output layer to obtain the multi-state quantities from the predicted pseudo values. We perform the Wilcoxon signed-rank test [25] to evaluate the significance of our models' performance.

6 Results and Discussion

We evaluate the models for each state or each transition and report the overall performance, i.e., the average evaluation metric (iAUC and iBS) scores of all states (for SOP and dynamic SOP prediction tasks) or of all transitions (for TP prediction task), along with statistical significance using Wilcoxon signed-rank test. The state-wise or transition-wise performance compar-

Table 1: Comparison of the models' performance on different censoring settings for **SOP** prediction on **Nonlinear Markov & Non-Markov** dataset.

Censoring Setting	Model	Nonlinear Markov						Nonlinear Non-Markov					
		iAUC (↑)			iBS (↓)			iAUC (↑)			iBS (↓)		
		25%	50%	75%	25%	50%	75%	25%	50%	75%	25%	50%	75%
Incremental Censoring	LMAJ [27]	0.50	0.50	0.50	0.19	0.18	0.16	0.50	0.50	0.50	0.13	**0.12**	0.12
	msCox [14]	0.52	0.52	0.53	0.22	0.22	0.21	0.62	0.62	0.62	**0.12**	**0.12**	0.12
	msWeibull [21]	0.52	0.52	0.51	0.19	0.18	0.17	0.59	0.54	0.56	0.22	0.22	0.20
	SurvNODE [18]	0.63	0.66	0.64	0.31	0.30	0.27	0.62	0.62	0.61	0.16	0.15	0.14
	LinearPseudo [6]	0.52	0.51	0.52	0.19	0.18	0.17	**0.63**	**0.63**	**0.63**	**0.12**	**0.12**	**0.11**
	msPseudo	**0.74**	**0.75**	**0.77**	**0.17**	**0.16**	**0.15**	**0.63**	**0.63**	**0.63**	**0.12**	**0.12**	**0.11**
Induced Censoring	LMAJ [27]	0.50	0.50	0.50	0.20	0.19	0.20	0.50	0.50	0.50	0.24	0.24	0.18
	msCox [14]	0.53	0.52	0.52	0.22	0.21	0.20	0.59	0.60	**0.62**	0.18	0.17	0.37
	msWeibull [21]	0.51	0.51	0.52	0.21	**0.17**	**0.16**	**0.62**	**0.62**	0.61	0.29	0.28	0.23
	SurvNODE [18]	0.67	0.67	0.65	0.33	0.28	0.26	0.60	0.60	0.59	0.18	0.19	0.20
	LinearPseudo [6]	0.52	0.53	0.52	0.20	0.19	0.20	0.60	0.59	0.61	0.16	**0.14**	**0.14**
	msPseudo	**0.74**	**0.72**	**0.72**	**0.18**	0.18	0.18	0.59	0.58	0.60	**0.14**	**0.14**	**0.14**
Transition-specific Incremental Censoring	LMAJ [27]	0.50	0.50	0.50	0.19	0.18	0.17	0.50	0.50	0.50	**0.13**	**0.12**	0.12
	msCox [14]	0.52	0.51	0.52	0.22	0.22	0.21	**0.64**	0.61	**0.61**	**0.13**	**0.12**	0.24
	msWeibull [21]	0.52	0.51	0.52	0.19	0.17	0.17	0.54	0.60	0.58	0.20	0.22	0.20
	SurvNODE [18]	0.65	0.66	0.67	0.31	0.29	0.27	0.63	0.61	**0.61**	0.16	0.15	0.14
	LinearPseudo [6]	0.51	0.52	0.52	0.19	0.18	0.17	**0.64**	0.63	**0.61**	**0.13**	**0.12**	**0.11**
	msPseudo	**0.75**	**0.74**	**0.76**	**0.16**	**0.16**	**0.15**	0.63	**0.64**	**0.61**	**0.13**	**0.12**	**0.11**
Transition-specific Induced Censoring	LMAJ [27]	0.50	0.50	0.50	0.20	0.20	0.20	0.50	0.50	0.50	0.24	0.18	0.18
	msCox [14]	0.51	0.51	0.56	0.22	0.21	0.20	0.59	**0.63**	0.58	0.17	0.18	0.17
	msWeibull [21]	0.51	0.52	0.52	0.21	**0.18**	**0.16**	**0.61**	0.62	**0.60**	0.27	0.28	0.28
	SurvNODE [18]	0.64	0.65	0.65	0.33	0.29	0.25	0.60	0.60	0.59	0.18	0.19	0.20
	LinearPseudo [6]	0.52	0.52	0.53	0.20	0.20	0.20	0.60	0.60	0.59	**0.15**	0.15	**0.14**
	msPseudo	**0.72**	**0.72**	**0.72**	**0.18**	**0.18**	0.18	0.59	0.59	0.59	**0.15**	**0.14**	**0.14**

Table 2: Comparison of the models on real-world (METABRIC & EBMT) datasets.

Dataset	Metric	Prediction Task	Model					
			msCox [14]	msWeibull [21]	SurvNODE [18]	LinearPseudo [6]	msPseudo	MT-msPseudo
METABRIC	iAUC (↑)	SOP	**0.69 (0.027)**	0.68 (0.041)^d	0.67 (0.051)	0.68 (0.016)^c	**0.69 (0.012)**	**0.69 (0.020)**
		Dynamic SOP	**0.70 (0.016)**	0.69 (0.037))	0.68 (0.048)	0.68 (0.016)	0.66 (0.057)	0.69 (0.029)
		TP (S=1 yr)	0.65 (0.028)^a	**0.71 (0.046)**	NA	0.67 (0.013)^b	0.70 (0.065)	0.67 (0.055)
	iBS (↓)	SOP	0.14 (0.069)^a	0.09 (0.037)^c	0.10 (0.004)^a	**0.08 (0.002)**	0.09 (0.008)	**0.08 (0.002)**
		Dynamic SOP	0.15 (0.063)^a	0.10 (0.044)	0.10 (0.002)^a	**0.09 (0.001)**	**0.09 (0.001)**	**0.09 (0.001)**
		TP (S=1 yr)	0.13 (0.015)^a	0.14 (0.083)^b	NA	0.12 (0.003)^a	**0.09 (0.028)**	0.10 (0.004)
EBMT	iAUC (↑)	SOP	0.59 (0.032)^d	0.59 (0.038)	0.58 (0.016)^c	0.58 (0.043)^c	**0.60 (0.045)**	0.58 (0.041)
		Dynamic SOP	**0.80 (0.028)**	0.59 (0.038)	0.57 (0.013)^a	0.59 (0.041)	0.60 (0.030)	0.61 (0.025)
		TP (S=1 month)	0.61 (0.042)^d	0.63 (0.063)	NA	0.63 (0.091)	**0.64 (0.075)**	0.64 (0.076)
	iBS (↓)	SOP	0.12 (0.009)^b	0.13 (0.009)^a	0.16 (0.016)^a	0.12 (0.008)^b	**0.11 (0.009)**	0.12 (0.009)
		Dynamic SOP	**0.09 (0.005)**	0.11 (0.011)^a	0.11 (0.008)^b	0.10 (0.011)	0.10 (0.012)	0.10 (0.011)
		TP (S=1 month)	0.06 (0.004)^a	0.10 (0.002)^a	NA	0.07 (0.013)^a	0.06 (0.014)	**0.03 (0.007)**

Wilcoxon signed-rank test - statistically significant codes: 0 'a' 0.001 'b' 0.01 'c' 0.05 'd' 0.1 ' ' 1, (Read '*' p as significant at $(p \times 100)\%$ level of significance)

ison results for all the multi-state models are provided as additional results in the supplementary materials [1].

Various censoring settings results: In table 1, we compare the iAUC and iBS results of different multi-state models on the nonlinear time-homogeneous Markov dataset and nonlinear reversible non-Markov dataset generated with different types of censoring settings with censoring amounts 25%, 50%, and 75%, respectively, for the SOP prediction task. Our `msPseudo` performs significantly better than other models in almost all cases on the nonlinear Markov dataset. In the nonlinear non-Markov dataset, both msPseudo and LinearPseudo (linear version of msPseudo) give similar or better performance than other models, especially in terms of iBS. The msPseudo model performs similarly to the LinearPseudo on nonlinear non-Markov dataset because the dataset is simple and contains only three covariates (two are binary). Therefore, our deep learning-based multi-state models did not show much improvement for this dataset. On the other hand, the nonlinear Markov dataset is more complex and contains 12 covariates with a large number of interaction terms, on which deep learning-based models significantly outperform the linear multi-state models. Our proposed

Table 3: Comparison of the models in terms of iAUC (↑ better) and iBS (↓ better) scores on simulated datasets.

Dataset	Prediction Task	Metric	Model				
			msCox [14]	msWeibull [21]	SurvNODE [18]	LinearPseudo [6]	msPseudo
Nonlinear Markov	**SOP**	**iAUC (↑)**	0.82 (0.008)^a	0.77 (0.013)^a	0.84 (0.008)^a	0.81 (0.010)^a	**0.87 (0.011)**
		iBS (↓)	0.21 (0.008)^a	0.30 (0.112)^a	**0.13 (0.005)**	0.15 (0.005)^a	**0.13 (0.007)**
Nonlinear Non-Markov	**SOP**	**iAUC (↑)**	0.67 (0.006)^a	0.59 (0.021)^a	0.71 (0.006)^a	0.79 (0.006)^a	**0.83 (0.005)**
		iBS (↓)	0.11 (0.009)^a	0.19 (0.018)^a	0.15 (0.007)^a	0.08 (0.006)^a	**0.07 (0.005)**
Linear Markov	**SOP**	**iAUC (↑)**	**0.56 (0.027)**	0.54 (0.014)^a	0.53 (0.018)^a	0.55 (0.026)^d	**0.56 (0.016)**
		iBS (↓)	0.18 (0.006)^a	**0.13 (0.004)**	0.15 (0.007)^a	**0.13 (0.004)**	**0.13 (0.005)**
	Dynamic SOP	**iAUC (↑)**	**0.66 (0.070)**	0.59 (0.019)	0.54 (0.012)^c	0.56 (0.021)	0.56 (0.016)
		iBS (↓)	0.18 (0.005)^a	**0.12 (0.004)**	0.17 (0.013)^a	0.13 (0.006)	0.13 (0.006)
	TP (S=1 yr)	**iAUC (↑)**	0.55 (0.012)^b	0.55 (0.012)^b	NA	0.55 (0.003)^a	**0.57 (0.013)**
		iBS (↓)	0.13 (0.004)^a	0.13 (0.004)^a	NA	0.14 (0.005)^a	**0.09 (0.004)**
Linear Non-Markov	**SOP**	**iAUC (↑)**	**0.63 (0.029)**	0.57 (0.025)^a	0.61 (0.039)^c	0.61 (0.028)^d	**0.63 (0.030)**
		iBS (↓)	0.14 (0.003)^a	0.19 (0.002)^a	0.19 (0.016)^a	0.14 (0.005)^a	**0.13 (0.005)**
	Dynamic SOP	**iAUC (↑)**	**0.72 (0.025)**	0.58 (0.018)^a	0.63 (0.015)^a	0.64 (0.014)^a	0.67 (0.011)
		iBS (↓)	**0.10 (0.007)**	0.14 (0.021)^a	0.12 (0.006)^a	0.11 (0.005)	0.11 (0.005)
	TP (S=1 yr)	**iAUC (↑)**	**0.65 (0.023)**	0.64 (0.032)	NA	0.63 (0.044)	0.64 (0.023)
		iBS (↓)	0.06 (0.009)	0.12 (0.004)^a	NA	**0.05 (0.002)**	0.07 (0.004)

Statistically significant codes: 0 'a' 0.001 'b' 0.01 'c' 0.05 'd' 0.1 ' ' 1, (Read '*' p as significant at $(p \times 100)\%$ level of significance)

msPseudo model, which uses pseudo values to handle censoring, achieves significant improvement over the other deep learning model, SurvNODE, in different censoring settings.

Real-world data results: The predictive performances for the real-world clinical data: METABRIC and EBMT are shown in Table 2. On the METABRIC dataset, our proposed models show overall better performance, especially in terms of iBS compared to the baselines. msCox shows better discriminative ability (better iAUC) in SOP and dynamic SOP prediction tasks; however, it suffers from loss in calibration (iBS) performance. Our models show a balance between iAUC and iBS scores. On the EBMT dataset, our models show outperformance over baselines in SOP and TP prediction tasks in terms of both iAUC and iBS, whereas msCox shows significant improvement in dynamic SOP prediction tasks.

Simulated data results: In table 3, it is observed that our proposed model, msPseudo, performs significantly better than the baseline models on nonlinear Markov and non-Markov datasets in terms of both iAUC and iBS. msPseudo consists of deep learning architecture along with pseudo values derived from consistent estimators for both Markov and non-Markov data, which helps the models to achieve outperformance even when the linearity and Markov assumptions get violated. On linear Markov and non-Markov datasets, msPseudo performs similar to or better than the statistical SOTA model for MSA, msCox. Linear models, such as msCox and LinearPseudo, perform similarly to the msPseudo on linear datasets since the datasets are generated assuming a linear relationship between covariates and the multi-state quantities of interest.

Limitations: We identify two limitations of our work: (a) Our models are computationally expensive compared to other non-deep learning-based MSA methods due to the pseudo values computation via leave-one-out approach. (b) The choice of statistical tests for testing the Markovianity of the data needs further investigation. We believe more research is needed to address these limitations, and in our future work, we plan to tackle them as follows: for (a) we will investigate using infinitesimal Jackknife approach for pseudo value calculation; and for (b) we will perform rigorous experiments to provide a robust Markov assumption testing tool.

7 Conclusion

Multi-state survival analysis (MSA) is an important yet understudied problem in time-to-event literature. Finding consistent estimators for non-Markov data is still an open problem in this field. In this paper, we proposed first-of-its-kind pseudo value-based deep learning models, **msPseudo** and **MT-msPseudo**, for estimating multi-state survival quantities in the presence of censoring without making any underlying assumptions on the multi-state processes. We showed that pseudo values can be a replacement for the estimation of multi-state quantities when derived from a consistent estimator. Through empirical experiments on simulated and real datasets, we demonstrated that our proposed models outperform other multi-state survival models under various censoring settings and for both Markov and non-Markov datasets. We believe this work lays the foundation for future investigations on the use of deep learning models for MSA, including explaining survival predictions and state-specific transition probabilities in real-world datasets.

Acknowledgment

This work is supported by grant IIS–1948399 from the US National Science Foundation (NSF).

References

[1] https://github.com/umbc-sanjaylab/SDM2023.

[2] O. O. Aalen and S. Johansen, *An empirical transition matrix for non-homogeneous markov chains based on censored observations*, Scandinavian Journal of Statistics, (1978).

[3] P. K. Andersen et al., *Generalised linear models for correlated pseudo-observations, with applications to multi-state models*, Biometrika, 90 (2003), pp. 15–27.

[4] P. K. Andersen et al., *Inference for outcome probabilities in multi-state models*, Lifetime data analysis, (2008).

[5] P. K. Andersen et al., *Inference for transition probabilities in non-markov multi-state models*, Lifetime Data Analysis, (2022).

[6] P. K. Andersen and J. P. Klein, *Regression analysis for multistate models based on a pseudo-value approach, with applications to bone marrow transplantation studies*, Scandinavian Journal of Statistics, 34 (2007), pp. 3–16.

[7] P. K. Andersen and M. P. Perme, *Pseudo-observations in survival analysis*, Statistical methods in medical research, (2010).

[8] N. Binder et al., *Pseudo-observations for competing risks with covariate dependent censoring*, Lifetime data analysis, (2014).

[9] D. Commenges and P. K. Andersen, *Score test of homogeneity for survival data*, Lifetime data analysis, (1995).

[10] A. Cottin et al., *Idnetwork: A deep illness-death network based on multi-states event history process for versatile disease prognostication*, in Survival Prediction-Algorithms, Challenges and Applications, PMLR, 2021, pp. 1–21.

[11] D. R. Cox, *Regression models and life-tables*, Journal of the Royal Statistical Society: Series B (Methodological), (1972).

[12] S. Datta and G. A. Satten, *Validity of the aalen–johansen estimators of stage occupation probabilities and nelson–aalen estimators of integrated transition hazards for non-markov models*, Statistics & probability letters, 55 (2001), pp. 403–411.

[13] J. de Uña-Álvarez and L. Meira-Machado, *Nonparametric estimation of transition probabilities in the non-markov illness-death model: A comparative study*, Biometrics, 71 (2015), pp. 364–375.

[14] L. C. De Wreede et al., *The mstate package for estimation and prediction in non-and semi-parametric multi-state and competing risks models*, Computer methods and programs in biomedicine, (2010).

[15] L. C. de Wreede et al., *mstate: an r package for the analysis of competing risks and multi-state models*, Journal of statistical software, (2011).

[16] S. Dutta et al., *Temporal prediction of future state occupation in a multistate model from high-dimensional baseline covariates via pseudo-value regression*, Journal of statistical computation and simulation, (2017).

[17] T. Fawcett, *An introduction to roc analysis*, Pattern recognition letters, 27 (2006), pp. 861–874.

[18] S. Groha et al., *A general framework for survival analysis and multi-state modelling*, arXiv preprint arXiv:2006.04893, (2020).

[19] R. Hoff et al., *Landmark estimation of transition probabilities in non-markov multi-state models with covariates*, Lifetime data analysis, (2019).

[20] P. Hougaard, *Multi-state models: a review*, Lifetime data analysis, (1999).

[21] C. H. Jackson, *flexsurv: a platform for parametric survival modeling in r*, Journal of statistical software, (2016).

[22] D. P. Kingma et al., *Adam: A method for stochastic optimization*, arXiv preprint arXiv:1412.6980, (2014).

[23] H. Lin and D. Zelterman, *Modeling survival data: extending the cox model*, 2002.

[24] N. Maltzahn et al., *A hybrid landmark aalen-johansen estimator for transition probabilities in partially non-markov multi-state models*, Lifetime data analysis, 27 (2021), pp. 737–760.

[25] J. McDonald, *Wilcoxon signed-rank test*, Handbook of biological Statistics, (2014).

[26] L. Meira-Machado et al., *Nonparametric estimation of transition probabilities in a non-markov illness–death model*, Lifetime Data Analysis, 12 (2006), pp. 325–344.

[27] H. Putter and C. Spitoni, *Non-parametric estimation of transition probabilities in non-markov multi-state models: The landmark aalen–johansen estimator*, Statistical methods in medical research, (2018).

[28] M. M. Rahman et al., *Deeppseudo: Pseudo value based deep learning models for competing risk analysis*, in AAAI, 2021.

[29] M. M. Rahman and S. Purushotham, *Fair and interpretable models for survival analysis*, in KDD, 2022.

[30] O. M. Rueda et al., *Dynamics of breast-cancer relapse reveal late-recurring er-positive genomic subgroups*, Nature, (2019).

[31] Y. Shu and J. P. Klein, *Additive hazards markov regression models illustrated with bone marrow transplant data*, Biometrika, 92 (2005), pp. 283–301.

[32] C. Spitoni et al., *Prediction errors for state occupation and transition probabilities in multi-state models*, Biometrical Journal, 60 (2018), pp. 34–48.

[33] A. C. Titman, *Transition probability estimates for non-markov multi-state models*, Biometrics, 71 (2015), pp. 1034–1041.

[34] A. C. Titman et al., *General tests of the markov property in multi-state models*, Biostatistics, (2020).

[35] L. Zhao and D. Feng, *Deep neural networks for survival analysis using pseudo values*, IEEE journal of biomedical and health informatics, (2020).