Health Outcomes in Mid-Ages: Multistate time to event Statistical Models versus Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) Models[*]

Lakshmi K. Raut
Visiting Fellow, University of Chicago
Center for the Economics of Human Development
1126 E. 59th St., Chicago, IL 60637,USA
mailto:lakshmiraut@gmail.com

December 31, 2019

Abstract

Health outcomes such as serious diseases, disability and death develop sequentially over one's life span. Genetic and epigenetic factors, and health related behaviors throughout one's life condition those health outcomes. Some individuals develop one or more diseases and their health deteriorate faster, worsening their quality of lives and survival probabilities. The models in health economics literature formulate progression of health outcomes over the life span using the multistate time to event framework, and estimate the effects of the above mitigating factors on probability of transitions from one health state to another. The sequential nature of health progression is captured in the Markovian structure. Markov chain models have short memory, i.e., these models assume that given the current health outcome, the past does not influence the probability of transition to another health outcome. Many chronic diseases such as cancers and heart diseases manifest as a result of long lagged past health behaviors. Markov chain models are limited in capturing these effects. More recently, long short term memory (LSTM) recurrent neural network (RNN) models are developed in the machine (deep) learning literature that keep track of important features from the past

[*] Draft prepared for presentation at the ASSA 2020 Annual Meeting, January 3 - 5, 2020, San Diego, California, USA.

Disclaimer: The views, thoughts, and opinions expressed in the paper belong solely to the author, and do not necessarily represent the views of any institution, other group or individual.

inputs in its memory cells, which the model determine important in the ‘context’ of the future outcomes. These models are generally applied to natural language processing, creating audio streams, and video subtitling. In this paper, I adapt the existing LSTM-RNN models for the prediction of sequential health outcomes in mid-ages. I first compare the merits and shortcomings of these two approaches and then use the Health and Retirement Study (HRS) data to compare their performances in predicting sequential health outcomes.

JEL Classifications: I12, C41, C51.

Keywords: Forecasting health outcomes, multistate time-to-event model, long short term memory (LSTM) recurrent neural network (RNN) model.

One-liner: The paper compares performances of a statistical multistate time to event model and a long short term memory (LSTM) recurrent neural network (RNN) model in forecasting sequentially the health outcomes over the life span.

1 Introduction

The path through various health outcomes such as serious diseases, disability and death develop sequentially over one’s life span. Many factors such as genetic and epigenetic factors, and health related behaviors throughout one’s life condition those health outcomes. Some individuals develop one or more diseases and their health deteriorate faster, worsening their quality of lives and survival probabilities, while others have slower aging process. The biology of aging process determines how those factors affect aging at the cellular level. In this paper, with insights from the biology literature, I use statistical and neural network models for predicting and estimating the transition probabilities between health states and the times they spend in a health state before transiting to another health state for individuals in their mid ages. While the framework I develop in the paper is applicable to many other heath outcomes such as incidence of cardiovascular and immune diseases, I focus on one health outcome of interest—any of the disabilities that qualifies for a public disability insurance program—with death as a competing risk health outcome. Death is a competing risk heath outcome in the sense that once dead, one cannot be at risk of disability, or as a matter of fact, any other health outcome of interest.

The main issues I address in this paper are: What kind of aging process and what conditioning factors for it the biology literature recommends? Which type of model—a multistate statistical model, a feedforward multilayer perceptron (MLP) type of neural network model, or a long short memory (LSTM) recurrent neural network (RNN) model—that is better suited for prediction and estimation of transition probabilities among health states, taking into account the effect of various covariates (the conditioning factors)? Which model can estimate the effect of various conditioning factors and their relative importance on the type of health trajectory one follows?

I will not get into the details of the biomedical literature on these issues. Here I will point out the main insights from my reviews elsewhere, Raut, 2019a; Raut, 2019b. Similar to the literature of behavioral genetics of personality and intelligence, the *nature-nurture* controversy exists in the health literature: Is it all nature (i.e., all genetics or genome) or is it all nurture (i.e., all epigenetics or epigenome modulated by the environment and health related individual behaviors) that determines the progression of health over the life span of an individual? The consensus so far is that it is neither the nature nor the nurture, it is a combination of the two that determines health developments over one's life-span. The research so far found that certain genetic make-ups (i.e., certain sequences of DNA) predispose one to certain diseases, (Barondes (1999); Khoury et al. (2009); Bookman et al. (2011)), but the epigenetic inputs—especially at the very early stage of life, i.e. in the womb, but not the least at later stages of life—are also very important determinants of life expectancy and quality of life. The biomedical research so far has not found genes that are responsible for aging and age related diseases, leading to early disability and mortality. The twenty-first century biomedical research emphasizes more on the epigenetic factors than the genetic factors to explain the pattern of health developments over the life-span.

At the cellular level, aging means cellular senescence—i.e., after a certain number of cell divisions, it stops dividing or have defective replications, causing tissues or organs to increasingly deteriorate over time. Senescence leads to incidence of degenerative diseases. It is generally observed that women live longer than men and those with better life styles in terms of smoking, exercising and diets delay the aging process (for evidence, see Blair et al. (1989); Vaupel (2010); Austad and Fischer (2016); Zarulli et al. (2018)). This line of biological inquiry led to explore the (cellular) molecular mechanism of aging process and to find biomarkers of aging that can be used to diagnose, monitor, and improve the age related physiological decline and disease. A good indicator of the aging process at the cellular level is the rate of decay in the telomere length. Telomeres are the caps at the end of chromosomes in a DNA sequence. They look like the plastic caps at the end of shoelaces. The main function of telomeres is to protect cells preserving the genetic content within each chromosome during cell divisions. Unfortunately, the telomere length shortens in the course of each cycle of chromosomal replication during cell division, reaching the

Hayflick limit (about 40 to 60 cell divisions, Hayflick, 1965) with a critically short telomere length, after which the cells stop dividing or divide with chromosomal abnormalities. The rate of shortening of the telomere length is modulated by telomerase enzyme. Why the rate of decay in telomere length varies for individuals is an active area of biomedical research and the mechanism for it is not yet fully understood. Many studies find that higher stress of any kind— psychological, financial, social and chemical—is strongly associated with higher oxidative stress, lower level of telomerase enzyme, and shorter telomere length. Furthermore, shorter telomere length is associated with health related phenotypes of poorer health and higher risks for cardiovascular and immune diseases (see, Epel et al. (2004); Shalev, Entringer, et al. (2013); DiLoreto and Murphy (2015); Shalev and Belsky (2016); Simons et al. (2016)).

More recently emerged second line of biomedical research on aging and aging related diseases explores the epigenetic (which literally means on top of genetic) mechanism for these life-cycle processes. (See for instance, Alisch et al. (2012); Barres and Zierath (2011); Boks et al. (2009); Esteller (2008); Hannum et al. (2013); Horvath (2013)).

The above literature emphasizes that aging and age related diseases are associated with shortening of telomere length and changes in global methylation, and that stress, smoking, drinking, chemical misuse, physical exercising, and diet are important modulators for these changes. The question remains, what are the critical periods or the developmental milestones in life-cycle that program the motions of health developments over the life-span of an individual?

Research along this line began with the striking findings of Barker (1990); Barker (1998) and later of Gluckman et al. (2008). They found strong associations between birth weight and many later life chronic diseases, including hypertension, coronary artery disease, type 2 diabetes, and osteoporosis. Many other studies find that much of health developments in later life is determined very early in life, during the prenatal period right after conception, i.e. in the womb. Sometimes it is said in social sciences that inequality begins in the womb. The effect of an environmental stress in the womb on later life diseases and developmental outcomes is known as *programming*. Gluckman et al. (2008) observes that "like the long latency period between an environmental trigger and the onset of certain cancers, the etiology of many later life diseases such as cardiovascular disease, metabolic disease, or osteoporosis originate as early as in the intrauterine development and the influence of environments that created by the mother." For more empirical evidence for the developmental origin of later life diseases, see Barker (2007); Thornburg et al. (2010). The papres by Kanherkar et al. (2014); Barbara et al. (2017) provide detailed descriptions of the biological process of development of life and health, starting from conception. They explain how the global DNA demethylation of the fertilized egg right after conception creates an epigenetic "clean slate" to start a new life, followed by rapid remythylation to reprogram the maternal and paternal genomes to create epigenetic configurations in the fetus which rapidly produce specialized cells of the body with cell divisions. The environment provided in mother's womb during those times has long-term effects on the child's later cognitive and other health developments. While inputs at early milestone ages are important for later age health, healthy living and good healthcare are still important for maintaining health in mid ages.

Studies in social sciences find that low socio-economic status (SES) are associated with inflammation, metabolic dysregulation, and various chronic and age-related diseases such as type 2 diabetes, coronary heart disease, stroke, and dementia, and that low SES create epigenetic changes in individuals that lead to faster biological aging even after controlling for health-related behaviors such as diet, exercise, smoking, alcohol consumption, or having health insurance, see for evidence, Simons et al. (2016). The study by Karakus and Patton (2011) uses the Health and Retirement Study data and after controlling for education, race, income, health risk indicators like BMI and smoking, functional limitations like gross motor index, health limitations for work, and income, they find depression at baseline leads to significantly higher risk for developing diabetes, heart problems, and arthritis and no significant effect on developing cancer during the 12 years follow-up period. Renna (2008) uses National Longitudinal Survey of Youth data to find no significant effect of alchohol use on labor market outcomes such as on earnings or hours of work. Seib et al. (2014) collected data on a sample of older women in Australia and found that severe traumatic life events create strong stress levels that influence them to have unhealthy living and diet measured by BMI and develop stronger and earlier health problems. Conti et al. (2009) utilize the CES-D data in the Health and Retirement Study dataset to construct a measure of depression, and find that depression of men and women have significant negative effect on employment status, early retirement, and application for DI/SSI benefits.

Using insights from the above literature, I formulate a finite state continuous time stochastic process model of disablement process. I postulate that as individuals age, the homeostatic regulatory mechanism that controls physiological systems—respiratory, cardiovascular, neuroendocrine, immune, and metabolic—becomes more and more fragile in its ability to face internal and external stressors, leading to early occurrence of disease, disability and death. I use available bio-markers (such as BMI, CES-D, cognition) and health related behaviors such as smoking, and vigorous exercising along the life-course to explain how they affect the risk of chronic diseases, disabilities and death. A multistate stochastic process framework is useful to study the effects of various covariates—the covariates that may be different for different intermediate health states—on the risks of disability and death.

The statistical models for estimation of transition probabilities are based on Markov processes and assume that transition intensities of these processes have Cox proportional hazard specifications, see for instance (Aalen and Johansen (1978); Andersen, Borgan, et al. (1993); Andersen and Perme (2008); Crowther and Lambert (2017); Fleming (1978)). Markov models have short memories and proportionality assumption imposes serious limitations on the structure of the model, which could be far from the functional form of the true data generating process. Neural network models relax these limitations.

It is known that a feedforward multilayer perceptron (MLP), also known as a feedforward deep learning model, with a sufficient number of neurons in the hidden layer can approximate any function as closely as desired. That is, an MLP is one of the best 'universal function approximator' (Hornik et al., 1989). A few papers—Faraggi and Simon, 1995; Biganzoli et al., 1998; Katzman et al., 2018; Lee et al., 2018; Ranganath et al., 2016– used feedforward MLP networks to compute the survival probabilities when there is only one possible transition between two health states—alive and death– with the exception of Lee et al. who studied competing risks, by breaking death into various causes of death. Katzman et al. introduced more general non-linearity of of the covariate effects, but still kept the Cox proportionality assumption. Ranganath et al. assumed parametric form for the baseline hazard function as compared to the non-parametric form in Cox model, but they made the covariate effects nonlinear. Ren et al., 2019 consider a recurrent neural network, but the covariates are time-fixed at the initial time step. They also restricted to only one transition, i.e., a two-state model. None of these models deals with sequential framework where new information arise with time steps and update the previously estimated transition probability estimates. In these models, all the inputs from the past, present and the future times in the sample determine current probabilities. These models, with the exception of Ren et al. although with other serious limitations, have no ways to store information learned from the past inputs. After training these models, when new data arrive, these models cannot use this new data to update the predicted probabilities without losing information in the early periods.

A recurrent neural network (RNN) uses feedback connections or self connection of neurons in the hidden layer, and thus is capable of storing important information learned in the past in these recurrent neurons. Like an MLP is a universal function approximator, an RNN has the similar nice property that with sufficiently large number of hidden recurrent neurons, an RNN can approximate any sequence-to-sequence mapping (Graves et al., 2014; Hammer, 2000; Siegelmann and Sontag, 1992). These models have shared weights between time-steps and in the input and output layers, as a result when new data arrive after training the model, it can use all the past important information learned from the past to this new data point and predict the future probabilities in the light of this new data. Since training such models involve computation of gradients using backpropagation through time, it involves multiplication of numbers less than one many times, leading to vanishing gradient problem. In these scenarios, it cannot keep useful information in memory from the long time back. Overcoming these problems led to a few modifications of the RNN framework. The most successful of them is the long short memory (LSTM) RNN model introduced by Hochreiter and Schmidhuber, 1997. For more on LSTM-RNN models, see Graves, 2012. I use this LSTM-RNN model for prediction of time-to-event probabilies of health outcomes. Another problem is with the training data size. To obtain good predictive performance, these models require a large number of training examples. In drug discovery problems or with surveys or lab experiements, obtaining large number of examples is costly. To overcome small data problem, Altae-Tran, 2016; Altae-Tran et al., 2017 introduced further refinement of the LSTM-RNN framework. I do not adopt such modifications. In this paper, I use the original LSTM-RNN model specified in Graves, 2012 and implemented in Keras module of Tensorflow 2.0 for Python package.

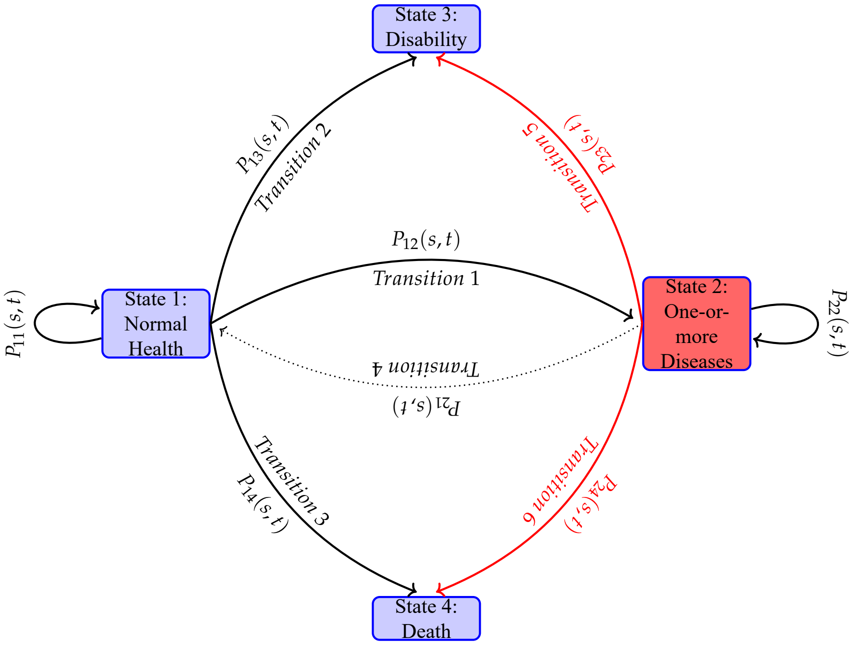
The rest of the paper is organized as follows. In Section 2, I describe the econometric specification of the maultistate stochastic process and describe estimation issues. In Section 3, I describe the Health and Retirement Study data set and the variables that I use in estimation in both frameworks. In Section 4, I describe the recurrent neural network with LSTM memory cells that I use to extend the multistate Cox model of health outcomes process. In Section 5, I describe the performance criteria that I use to compare the performance of the models used in this study. Section 6 concludes the paper.

2 Statistical multi-state model of health outcomes process

The goal is to formulate and then estimate an econometric model of paths to enter disability rolls. An individual can be on the disability rolls if the individual has a qualifying disability before reaching age 65 and has not died before applying for disability benefits. I assume that an individual's getting on the disability rolls is a terminal event, i.e., the individual does not move to normal or diseased health states. After reaching this state, the individual is not followed any further. A competing risk for getting on the rolls is death before age 65. This is a competing risk because an individual cannot be at risk for disability enrollment if the individual is already dead and thus not at risk to get on the disability rolls. The individual is not followed after the event of death because our primary interest is the event of getting on to the disability roll. In the technical terms defined below, we treat the health states—disability and death—as absorbing states, i.e., once in that health state, an individual remains in that health state and removed from the sample for later considerations. An individual can be in normal health and then become disable or die before becoming disabled or may first become diseased with one or more diseases and again from that health state become disabled or die before becoming disabled. Various factors affect individual risks of various transitions of health states and the time they stay in each health state along the life-span. Both, in turn, determine the timing of getting on to the disability rolls.

I model the paths through various health states that individuals follow along their life-spans as a continuous-time finite-state Markov process $X(t), t \in T$, where at each time point t during the study period T, the random variable $X(t)$ takes a value from a finite number of health states in S. In the present study, we take $T = [0, 7]$, treating age 51 as time period $t = 0$ and age 65 as time period $t = 7$. The unit of time is 2 years, as HRS collected data every two years. The state space S contains states, 1 = ”healthy or normal health”, 2 = ”ill with one or more chronic disease”, 3 = ”disabled with DI-or SSI-qualifying disability” and 4 = ”Death”. Sometimes I will use $S = \{h, i, d, D\}$ in place of $\{1, 2, 3, 4\}$.

Typically in the study period, an individual along the path to disability or death before age 65 may be in the normal health state for a length of time, and then moves to another health state, say diseased health state, and remain there for some time, and then jump to the health state of disability or to death, or reach 65 and censored. There are many possible paths that an individual can follow. Even when the health states they pass through are the same, the duration of stay in each health state (also known as the *waiting time* in stochastic process literature) could vary. Each configuration of visited states and the waiting times in those states constitute one path. When time is continuous, an uncountably large number of paths are possible. From the diagram below one can see various paths that an individual may follow during the study period. The focus of the paper is to study the probabilities of various transitions and the duration of stay in each health state.



Let the transition probabilities of our Markov process $X\left(t\right)$ be given by

$$P_{hj}(s,t) = Prob(X(t) = j|X(s) = h), \tag{1}$$

for all $h, j \in S$, and $s, t \in T,$, $s \leq t$. Denote the matrix of transition probabilities by

$$P\left(s,t\right) \equiv \left(P_{hj}(s,t)\right)_{h,j=1\ldots p}.$$

An individual may be in any of the health states in S at time t, the probability of which, known as the *occupation probability*, depends on the occupation of the previous health states. and the transition probabilities among health states. Let $\pi_j(t)$ be the occupation probability of an individual in health state j at time t. Occupation probability can be also viewed as the proportion of population of age t who are in health state j. Denote all the occupation probabilities as a column vector $\pi(t) \equiv (\pi_j)(t), j \in S$. Then the occupation probabilities move over time recursively as follows,

$$\pi(t) = \pi'(s)P(s,t), 0 \leq s < t.$$

The goal is to estimate for each time time $s, s \in T$ the transition probability matrices $P(s,t), t > s, \, t \in T$, taking into account the effects of (i.e., conditioning on) the past materialized values of the determining factors, that is covariates, and the stochastic process $X(.)$ up to time s.

It is known that the transition probabilities of a Markov process satisfies the Chapman-Kolmogorov equation

$$P\left(s,t\right) = P\left(s,u\right) \cdot P\left(u,t\right), \text{ for all } s,u,t \in T \text{ with } s < u < t \tag{2}$$

I will use the above to point out how the transition probabilities are parameterized and statistically estimated by the Aalen-Johnson non-parametric estimation method that plugs in the Nelson-Aalen estimates of the integrates hazards for each transition separately in an unified way for both continuous time and discrete time Markov processes.

I assume that the transition probabilities $P\left(s,t\right), s,t \in T, s < t$ are absolutely continuous in s and t. A *transition intensity*—also known as the *hazard rate* in the survival analysis literature when there is only one possible transition (alive to death), and as the *cause-specific hazard rate* in the competing risk analysis when death is split into various causes of death[1]—of the health process X_t from health state h to health state j at time t is the derivative

$$\begin{aligned}
\lambda_{hj}(t) &= \lim_{\Delta t \to 0} \frac{P_{hj}(t,t+\Delta t) - P_{hj}\left(t,t\right)}{\Delta t}, \text{for } j \in S, \text{ which for } j \neq h \text{ becomes} \\
&= \lim_{\Delta t \to 0} \frac{P_{hj}(t,t+\Delta t)}{\Delta t}, \text{ and for } j = h \text{ becomes} \\
\lambda_{hh}(t) &= \lim_{\Delta t \to 0} \frac{P_{hh}(t,t+\Delta t) - 1}{\Delta t} \qquad (3) \\
&= -\lim_{\Delta t \to 0} \frac{\sum_{j \neq h} P_{hj}(t,t+\Delta t)}{\Delta t} \\
&= -\sum_{j \neq h} \lambda_{hj}\left(t\right)
\end{aligned}$$

For absorbing states $h = 3,4$, the transition intensities $\lambda_{hj}(t) = 0$, for all $j, j \in S$. Denote

[1] See for instance, Raut, 2017 for a competing risk analysis in a similar context using the SSA Administrative data and compare that with the present framework.

the matrix of transition intensities by

$$\Gamma\left(t\right)=\begin{pmatrix} -\left(\lambda_{12}\left(t\right)+\lambda_{13}\left(t\right)+\lambda_{14}\left(t\right)\right) & \lambda_{12}\left(t\right) & \lambda_{13}\left(t\right) & \lambda_{14}\left(t\right) \\ 0 & -\left(\lambda_{23}\left(t\right)+\lambda_{24}\left(t\right)\right) & \lambda_{23}\left(t\right) & \lambda_{24}\left(t\right) \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{pmatrix}. \tag{4}$$

Every Markov stochastic process has an associated intensity process. In the theory of stochastic processes, it is known that given an intensity function $\Gamma\left(t\right)$, there exists a continuous time Markov chain satisfying the Chapman-Kolmogorov equation Eq. (2) and conversely, given a Markov chain with transition probabilities $P\left(s,t\right),s,t\in T,s<t$, that satisfies the Chapman-Kolmogorov equation Eq. (2), one can derive the transition intensities of the form in Eq. (4). For this reason, the intensity function $\Gamma\left(t\right)$ is also referred as *infinitesimal generator of the Markov process*. While $P_{hj}(s,t)$ is an unconditional probability, the transition intensity or the hazard rate $\lambda_{hj}(t)\Delta t$ is the conditional (instantaneous) probability. More specifically, $\lambda_{hj}(t)\Delta t$ is the probability that an individual experiencing event j in a very small interval of time $[t,t+\Delta t)$ given that he has been in state h at time t. This conditional probability may depend on time t and other characteristics up to time t.

In statistical and neural network models, the dependence of transition probabilities on individual characteristics is generally done through parametric or semi-parametric specification of the transition intensity functions $\Gamma\left(t\right)$. One then estimates the transition probabilities from the non-parametric or semi-parametric estimates of the integrated transition functions. An *integrated transition intensity function* $\Lambda_{hj}\left(t\right)$ for a transition $h\rightarrow j$ is defined by $\Lambda_{hj}\left(t\right)=\int_0^t\lambda_{hj}\left(u\right)du$. Just like for a continuous random variable, it is easier to estimate its cumulative density function nonparametrically than its density function, for the time-to-event data with censoring, it is easier to estimate the integrated hazard function than the intensity function. While for the discrete case this problem does not arise, estimation of transition probabilities via nonparametric estimates of the integrated hazard function is an unified approach encompassing both discrete time and continuous time data. I follow this strategy.

To explain and gain insights into this estimation strategy, denote the matrix of all the cumulative hazard functions as $\Lambda\left(t\right)=\left(\Lambda_{hj}\left(t\right)\right)_{h,j=1,2,3,4}$, and the matrix of the derivatives of the component functions by $d\Lambda\left(t\right)$. Let the time interval $[s,t]$ is subdivided into a partition of m sub-intervals with cut-off points $s=t_0<t_1<...<t_m=t$. Denote the partition by $P(m)$. Denote the largest size of the sub-intervals by $|P(m)|\equiv$

$\max\{|t_i - t_{i-1}|, i = 1, ..., m\}$. Applying repeatedly the Chapman-Kolmogorov equation Eq. (2) on the sub-intervals of the partition, we have

$$P(s,t) = P(t_0, t_1) \cdot P(t_1, t_2) \cdot ... \cdot P(t_{m-1}, t_m) = \prod_{i=1}^{m} P(t_{i-1}, t_i) \tag{5}$$

Note that as $|t_{i-1} - t_i| \to 0$, the transition probability matrix $P(t_{i-1}, t_i) \to P(t_{i-1}, (t_{i-1} + dt) = I + \Gamma(t)dt$.[2] As $|P(m)| \to 0$, the right hand side of Eq. (5) converges to a matrix, known as the *the product integral*[3]of the integrated hazard functions $\Lambda(s,t)$, denoted as $\prod_s^t (I + d\Lambda(u))$. Or in other words, the transition probabilities of a stochastic process parameterized via an intensity process is given by the product integral of integrated hazard function.

$$P(s,t) = \prod_s^t (I + d\Lambda(u)). \tag{6}$$

The above product-integral solution is a generalization of the Kaplan-Meier (Kaplan and Meier, 1958) product-limit formula for the survival function in survival analysis. The product integral formula unifies both discrete time and continuous time Markov processes, and is an extremely useful apparatus for statistical analysis of Markov processes.

The most widely used statistical procedure to estimate the transition probabilities $P(s,t)$, $s, t \in T, s < t$ is to plug in an estimate $\Lambda(u)$ in Eq. (6). The effect of covariates is incorporated by conditioning the transition intensity functions $\Gamma(t; X(t))$ on the covariates process $X(t)$. There are many ways to get these estimates. I will follow two approaches in this paper: First, I will explore the more widely used non-parametric Aalen-Johnson-Fleming method via Nelson Aalen estimates for each-component of the $\hat{\Lambda}(u; X(u))$ with Cox proportional hazard model to incorporate the time-varying covariate effects in the next sub-section. Second, the Neural network approach explored in a later section.

2.1 Aalen-Johansen Estimator of Transition Probabilities

Most widely used statistical procedure incorporates the time-varying covariates for the transition probabilities by specifying a semi-parametric functional forms for the intensity hazard

[2] From definition of transition intensities above and writing it in matrix form, we have $\Gamma(t) = \lim_{\Delta t \downarrow 0} \frac{P(t,t+\Delta t) - P(t,t)}{\Delta t} = \lim_{\Delta t \downarrow 0} \frac{P(t,t+\Delta t) - I}{\Delta t}$. From this it follows that for small Δt, we have $P(t, t + \Delta t) = I + \Gamma(t)\Delta t$.

[3] For more formal treatment of product integral see Gill and Johansen, 1990 and for a lucid exposition with some applications, see Gill, 2005.

functions

$$\lambda_{hj}\left(t;X\left(t\right)\right)=\lambda_{hj}^{0}\left(t\right)e^{\beta_{hj}'X(t)} \tag{7}$$

$\lambda_{hj}^{0}\left(t\right)$ is known as the *baseline hazard function*. The specification of transition intensity in Eq. (7) is known as the *proportional hazard model*. It aggregates the effects the regressors linearly as a measure of some kind of latent factor, and that latent factor shifts the baseline hazard proportionately, i.e., the effect on hazard is uniform over time. Two papers (Fleming, 1978 and Aalen and Johansen, 1978) independently extended the Kaplan-Meier nonparametric product limit estimator from survival analysis to the multi-state time to event models. While Fleming gave the estimator for complete data, Aalen and Johansen gave the estimator for censored data. To describe the Aalen-Johansen estimator, let me introduce some concepts and notation. For each individual $i, i=1,2,...,n$ and corresponding to each transient health state, $h, h=1,2$, define two types of stochastic processes: (1) the counting processes $N_{hj,i}(t)$ denoting the ***observed*** number of transitions from health state h to health state j that the individual i has made by time t—which in our case is either 0 or 1, since by assumption when an individual exits a health state, the individual does not return to it in future ; and (2) $Y_{h,i}(t)$, taking value 1 if individual i is at risk at time t for transition to another possible health state, and taking value 0 otherwise.

Let us focus on one transition $h\rightarrow j$. Denote by $\bar{N}_{hj}(t)=\sum_{i}^{n}N_{hj,i}(t)$, a counting process measuring the number of transitions of the $h\rightarrow j$ in the sample at time t, $\bar{Y}_{h}(t)=\sum_{i}^{n}Y_{h,i}(t)$, a counting process measuring the number of individuals in the sample at risk for a transition at time t, and $\bar{M}_{hj}(t)=\sum_{i}^{n}M_{hj,i}(t)$. In any empirical study the data will be at the discrete times, say in ordered times $0=t_0<t_1<...<t_m$. At each time t_i, we calculate

$$\hat{\lambda}_{hj}\left(t_i\right)=\frac{\triangle\bar{N}_{hj}\left(t_i\right)}{\bar{Y}_{h}\left(t_i\right)},j\neq h, \tag{8}$$

Without covariates, the *Nelson-Aalen non-parametric estimate* of the integrated intensity functions is given by, for each $h=1,2$

$$\begin{aligned}\hat{\Lambda}_{hj}\left(t\right) &= \sum_{i:t_i\leq t}\hat{\lambda}_{hj}\left(t_i\right),j\neq h,\\ \hat{\Lambda}_{hh}\left(t\right) &= -\sum\hat{\Lambda}_{hj}\left(t\right)\\ \hat{\Lambda}_{hj}\left(t\right) &= 0\text{ for all other }h,j\text{ combinations}\end{aligned} \tag{9}$$

The *Aalen-Johansen estimator* $\hat{P}(s,t), s,t,\in T, s<t$ for the transition probabilities is obtained by substituting for each component hj the Nelson-Aalen estimates $\hat{\Lambda}_{hj}(t)$ and

then applying the product integral formula Eq. (6) as follows

$$\hat{P}(s,t) = \prod_{s<u<t} \left(I + d\hat{\Lambda}(u)\right) = \prod_{i:t_i \leq t} \left(I + \left[\hat{\Lambda}(t_i) - \hat{\Lambda}(t_{i-1})\right]\right). \tag{10}$$

With covariates one obtains the Cox partial likelihood estimate for $\hat{\beta}_{hj}$ for each transition $h \to j$ separately and then computes an weighted risk set defined by

$$\bar{Y}^*_{hj}(t) = \sum_{i=1}^{n} Y_{hj,i}(t) \exp\left(\hat{\beta}'_{hj} X^0_{h,i}\right). \tag{11}$$

The estimates of cumulative intensities with covariates are obtained from Eq. (8) by replacing, $\bar{Y}_h(t)$ with $\bar{Y}^*_{hj}(t)$.

Nelson-Aalen estimator has nice statistical property. For instance, using Martingale calculus, it can be shown that the estimator is asymptotically unbiased. Using the results from Martingale theory, one can derive the formula for variance-covariance estimates of parameter estimates and the normalized estimate is normally distributed (central limit theorem holds for normalized parameter estimates), see for details, Aalen, Borgan, et al., 2008; Andersen, Borgan, et al., 1993; Fleming and Harrington, 2005.

The likelihood of the sample (for details, see Andersen, Borgan, et al., 1993; Andersen and Perme, 2008; Commenges, 2002),

$$L(\theta) = \prod_{i} \prod_{\substack{h=1,2 \\ j=2,3,4 \\ h \neq j}} \left(\prod_{t} \lambda_{hj,i}(t|X_{h,i})^{\triangle N_{hj,i}(t)} \right) \exp\left(- \int_0^{T^*_{h,i}} \lambda_{hj,i}(u|X_{h,i}) \right) du \tag{12}$$

I will use this likelihood function in construction of the log-likelihood loss function for the neural network models.

I use the R package, *mstate*, developed and described by the authors in Wreede et al. (2010) for the estimation of the parameters and their standard errors?

The Aalen-Johnson nonparametric estimates without any covariates of the transition probabilities, are shown in Table 3 and plotted in Figure 1.

For the Cox regression parameter estimates, I have used both the R package *mstate* (see, Wreede et al., 2010 for details) and also used the SAS procedure *phreg* (both produced the same estimates) and used the *mstate* package to estimate all the transition probabilities (SAS does not have readily available procedure for this purpose). The parameter estimates are shown in Table 4 and Table 5.

3 The data set and the variables

I use the Health and Retirement Study (HRS) dataset for empirical analysis. A lot has been reported on the family of HRS datasets—about its structure, purpose, and various modules collecting data on genetics, biomarkers, cognitive functioning, and more, see for instance Juster and Suzman (1995); Sonnega et al. (2014); Fisher and Ryan (2017). The first survey was conducted in 1992 on a representative sample of individuals living in households i.e., in non-institutionalized, community dwelling, in the United States from the population of cohort born during 1931 to 1941 and their spouses of any age. "The sample was drawn at the household financial unit level using a multistage, national area-clustered probability sample frame. An oversample of Blacks, Hispanics (primarily Mexican Americans), and Florida residents was drawn to increase the sample size of Blacks and Hispanics as well as those who reside in the state of Florida", Fisher and Ryan (2017). The number of respondents were 13,593. Since 1992, the survey were repeated every two years, each is referred to as a wave of survey. New cohorts were added in 1993, 1998, 2004 and 2010, ending the survey up with the sample size of 37,495 from around 23,000 households in wave 12 in 2014. RAND created many variables from the original HRS data for ease of use. I create my dataset and all the variables with a few exceptions mentioned below from the RAND HRS dataset version P. The details of the Rand HRS version P can be found in Bugliari et al. (2016).

As mentioned in the introduction, I define the disability health state to be one that qualifies one to be on the disability programs OASDI or SSI. The data on disability is self-reported. Later I plan to use the Social Security Administration's matched administrative data on this variable and earnings variables not included here. The matched data will, however, reduce the sample size to half, as only 50 percent of the respondents are used for matching HRS with SSA Administrative data. The HRS data collected information on if and when the doctor diagnosed that the respondent has any of the severe diseases such as high blood pressure, diabetes, cancer, lung disease, heart attack, stroke, psychiatric disorder and severe arthritis. I drop respondents who received disability before the first survey year 1992 and I also drop the spouses in the sample who were not born between 1931 to 1941, that is the respondents in our sample are between age 51 to 61 and not disabled or dead in 1992. I ended up with the final sample size of 9,493 for this analysis.

Table 1 and Table 2 provide a few characteristics of the data.

Table 1: Percentage distribution of the pooled sample population by health status by age

Age	#obs	Percentage distriubtion of HealthStatus			
		Normal	With Diseases	Disabled	Died at age
51	945	47.62	52.38	0.00	0.00
52	936	47.65	52.35	0.00	0.00
53	1906	42.71	56.19	0.73	0.37
54	1850	41.95	57.08	0.76	0.22
55	2791	39.59	58.47	1.29	0.64
56	2684	37.30	61.07	0.97	0.67
57	3572	35.39	62.46	1.12	1.04
58	3469	32.92	64.72	1.33	1.04
59	4240	31.04	66.11	1.75	1.11
60	4182	29.94	67.62	1.51	0.93
61	4894	27.42	69.55	1.45	1.57
62	4080	25.51	70.20	2.16	2.13
63	4746	23.68	72.48	1.85	1.98
64	3905	21.95	75.29	0.72	2.05
65	4564	20.86	76.40	0.66	2.08

Source: The author.

Table 2: Distribution of the sample by health status in various survey rounds

Age	#obs	Percentage Distribution of HealthStatus			
		Normal	With Diseases	Disabled	Died in period
1992	9493	39.65	60.35	0.00	0.00
1994	9493	34.16	62.73	1.75	1.36
1996	9198	30.17	66.72	1.52	1.59
1998	7461	27.48	69.09	1.81	1.62
2000	5791	25.38	71.14	1.49	1.99
2002	4106	22.80	74.35	1.32	1.53
2004	2437	20.68	75.91	1.40	2.01
2006	785	17.83	79.75	0.38	2.04

Source: The author.

3.1 Variables

The demographic variables **White** and **Female** have the standard definition. The variable **College** is a binary variable taking value 1 if the respondent has education level of college and above (does not include some college), i.e., has a college degree and more and taking value 0 otherwise.

cesd: I used a score on the Center for Epidemiologic Studies Depression (CESD) measure in various waves that is created by RAND release of the HRS data. RAND creates the score as the sum of five negative indicators minus two positive indicators. "The negative indicators measure whether the Respondent experienced the following sentiments all or most of the time: depression, everything is an effort, sleep is restless, felt alone, felt sad, and could not get going. The positive indicators measure whether the Respondent felt happy and enjoyed life, all or most of the time." I standardize this score by subtracting 4 and dividing 8 to the RAND measure. The wave 1 had different set of questions so it was not reported in RAND HRS. I imputed it to be the first non-missing future CESD score. In the paper, I refer the variable as cesd. Steffick (2000) discusses its validity as a measure of stress and depression.

cogtot: This variable is a measure of cognitive functioning. RAND combined the original HRS scores on cognitive function measure which includes "immediate and delayed word recall, the serial 7s test, counting backwards, naming tasks (e.g., date-naming), and vocabulary questions". Three of the original HRS cognition summary indices—two indices of scores on 20 and 40 words recall and third is score on the mental status index which is sum of scores "from counting, naming, and vocabulary tasks"—are added together to create this variable. Again due to non-compatibility with the rest of the waves, the score in the first wave was not reported in the RAND HRS. I have imputed it by taking the first future non-missing value of this variable.

bmi: The variable body-mass-index (BMI) is the standard measure used in the medical field and HRS collected data on this for all individuals. If it is missing in 1992, I impute it with the first future non-missing value for the variable.

behav_smoke: This variable is constructed to be a binary variable taking value 1 if the respondent has reported yes to ever smoked question during any of the waves as reported in the RAND HRS data and then repeated the value for all the years.

behav_vigex: The RAND HRS has data on whether the respondent did vigorous exercise three or more days per week. I created this variable in each time period to take the value 1 if the respondent did vigorous exercise three or more days per week.

3.2 Statistical Results

Table 3: Estimated transition probabilities for transition $i \to j$ by duration of stay in state 1 and 2

duration	Tr1_1	Tr2_2	Tr1_2	Tr1_3	Tr2_3	Tr1_4	Tr2_4
0	1.0000	1.0000	0.0000	0.0000	0.0000	0.0000	0.0000
1	0.8711	1.0000	0.1289	0.0000	0.0000	0.0000	0.0000
2	0.7543	0.9767	0.2367	0.0062	0.0127	0.0027	0.0106
3	0.6570	0.9430	0.3105	0.0201	0.0292	0.0123	0.0279
4	0.5786	0.9082	0.3651	0.0331	0.0477	0.0232	0.0441
5	0.5126	0.8730	0.4087	0.0446	0.0659	0.0341	0.0610
6	0.4761	0.8421	0.4212	0.0562	0.0792	0.0465	0.0787
7	0.4667	0.8229	0.4116	0.0681	0.0896	0.0536	0.0874

Note: Time is in the unit of 2 years.

With only demographic covariates (most that can be done with the Administrative data) the parameter estimates in Table 4 show that significantly lower risks of transitions 1 → 4; 2 → 3; 2 → 4 for whites and 1 → 4 and 2 → 4 for women. This may entail that the genetic make-up of being white or female sex yield favorable genetic predisposition to have better health outcomes and longer life. This is a misleading inference as we will see next that when we control for epigenetic factors that the biomedical literature pointed out to have significant effects on aging process and health outcomes, the above effects disappear.

Table 5 shows the Cox regression coefficient estimates of the effects of various factors on the risk of having transitions $h \to j$ from health status $h = 1, 2$ to health status $j, j = 2, 3, 4$.

These estimates show that the parameter estimates for the demographic covariates in Table 4 are biased as they are capturing the effects of excluded epigenetic and behavioral factors in that model. After controlling for these epigenetic and behavioral factors, the significance of those effects disappear. Furthermore, women show significantly lower probability of transition from diseased health states onto the disability health state.

Most important factors in Table 5 are cesd, measuring depression and stress and college graduation or higher level of education, with positive effect on all transitions with the

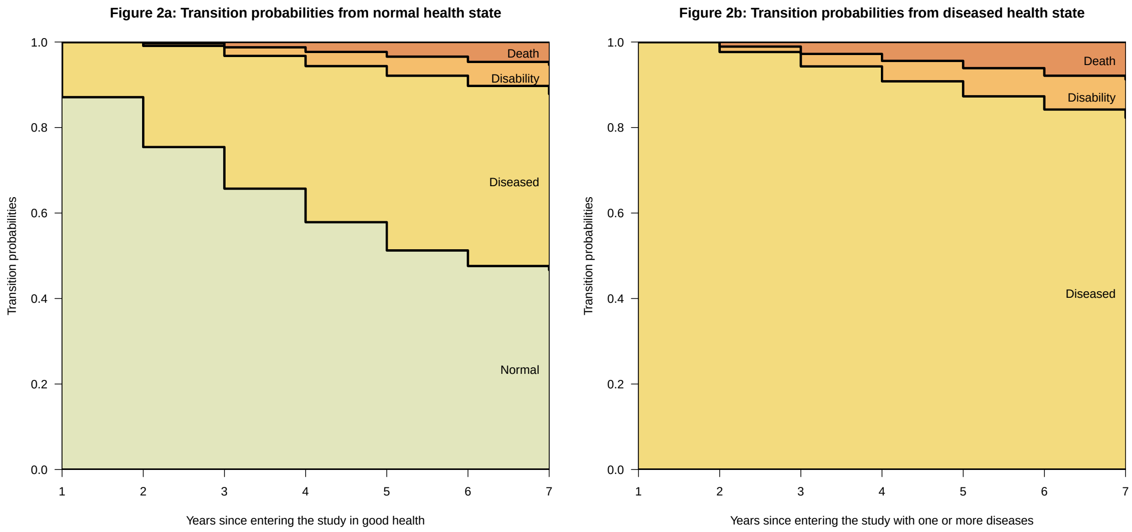


Figure 1: Transition probabilities (a) from normal health state and (b) from diseased health state

exception of no effect on transition from normal health to death.

Other important factors are smoking, with significant adverse effect on transitions, and exercising three or more times regularly has significant favorable effect on most transitions.

Table 4: Estimates of Cox regression models separately for each transition with demographic variables only

	1->2	1->3	1->4	2->3	2->4
White	-0.0280	-0.3279	-0.7162^{**}	-0.3867^{***}	-0.4259^{***}
	(0.0637)	(0.2293)	(0.2623)	(0.1005)	(0.1000)
Female	0.0515	-0.2163	-0.5220^{*}	-0.0285	-0.4927^{***}
	(0.0470)	(0.1899)	(0.2493)	(0.0909)	(0.0918)
AIC	25346.8833	1713.1509	1041.6019	8080.1715	7993.8108
R^2	0.0003	0.0009	0.0031	0.0020	0.0063
Max. R^2	0.9992	0.3709	0.2496	0.6927	0.6902
Num. events	1602	112	69	476	475
Num. obs.	3583	3695	3652	6856	6855
Missings	0	0	0	0	0
PH test	0.0426	0.2507	0.4483	0.1161	0.6312

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 5: Estimates of Cox regression models separately for each transition with health measures

	1->2	1->3	1->4	2->3	2->4
cesd	0.5612^{***}	1.9802^{***}	0.1265	1.2950^{***}	0.5467^{*}
	(0.1095)	(0.3460)	(1.0202)	(0.1587)	(0.2461)
bmi	0.0423^{***}	-0.0051	0.0118	0.0250^{**}	-0.0213
	(0.0055)	(0.0313)	(0.0499)	(0.0088)	(0.0168)
cogtot	-0.0029	-0.0671^{**}	0.0152	-0.0353^{***}	-0.0091
	(0.0058)	(0.0256)	(0.0407)	(0.0094)	(0.0157)
behav_smoke	0.0454	0.2577	2.5107^{*}	0.3814^{***}	0.8173^{***}
	(0.0508)	(0.2228)	(1.0166)	(0.1078)	(0.1777)
behav_vigex	-0.1966^{**}	-0.9995^{***}	-1.1687^{*}	-0.6103^{***}	-1.1988^{***}
	(0.0712)	(0.2438)	(0.4997)	(0.1039)	(0.1431)
White	0.0452	0.0982	-0.4942	-0.1172	-0.3252^{*}
	(0.0685)	(0.2823)	(0.5103)	(0.1080)	(0.1576)
College	-0.1112	-0.8871^{*}	-1.1140	-0.6839^{***}	-0.7443^{**}
	(0.0648)	(0.4167)	(0.7564)	(0.1937)	(0.2571)
Female	0.0851	-0.2734	-0.6858	-0.0801	-0.2518
	(0.0515)	(0.2335)	(0.4694)	(0.1006)	(0.1445)
AIC	23311.3089	1358.1202	318.1039	7268.4078	3268.0150
R^2	0.0294	0.0252	0.0084	0.0366	0.0217
Max. R^2	0.9993	0.3483	0.0961	0.7029	0.4349
Num. events	1500	95	23	446	207
Num. obs.	3239	3334	3262	6165	5926
Missings	344	361	390	691	929
PH test	0.0237	0.2294	0.0447	0.0001	0.1660

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

4 Long-short memory (LSTM) recurrent neural network (RNN) model of health outcomes process

I first give an overview of the neural network framework. I then briefly review a selected few papers that used the neural network framework to estimate and predict the time-to-event probabilities in a single event set-up. The main focus of the papers, including this paper, is to relax the proportionality and linearity assumption of the statistical survival models, see Eq. (7) for the proportionality assumption. I then explain why a long-short memory (LSTM) recurrent neural network (RNN) framework is more appropriate for predictive models of time-to-event probabilities in sequential set-up. I describe the *c-index* criterion that I use to compare the performance of various models and compare the performance of the multistate statistical model and a feed-forward neural network model of competing risk model by Lee et al., 2018 with the LSTM-RNN model of this paper. I code these three models using Keras and Tensorflow 2.0 modules in Python and estimate them using the Health and Retirement Survey data mentioned above. In the following section, I report the findings.

4.1 The basics of neural network

Neural network is a highly parameterized universal function approximator of the form $\hat{y} = f(x; w)$, x is a set of inputs, and w is a vector of parameters. This is of the same nature as a statistical model. More precisely, suppose we have data on a set of individuals of the type (x, y), where x is a vector of individual characteristics, and y is a vector of output levels and w is a set of parameters. The output could be a categorical variable for classification problems, it could be a probability distribution over finite classes, as in our case, or it could be a continuous variable for regression problems. The data generating process for y as a function of x, is not known. The goal is to approximate that unknown data generating function. This is the problem that both statistics and neural network find computational solutions to. In neural network, the problem is to design a *neural network architecture* of the approximating function $\hat{y} = f(x, w)$ and find a suitable *learning algorithm* to learn the parameter values w of the network using a training set of examples. This trained network can then be used to predict y for an individual given his characteristics x. The popularity and wide applicability of neural network lies in the fact that it designs the approximator in a hierarchy of functions, joined together by compositions of functions, that renders good properties in terms of ease of computation and closeness of function approximation. Most neural network models have the following type of hierarchical functional form:

$$\hat{y} = f(x;w) \equiv f^L_{w^L} \circ .... \circ f^1_{w_1}(x). \tag{13}$$

Each function corresponds to a layer of artificial neurons. The role of each neuron is to perform simple calculations and then pass on that to the next layer of neurons.

Neurons in each layer get signals which are the outputs of the neurons of the previous layer (also known as activation levels) that it is connected with. It sums them, I denote this sum with z and apply an activation function to produce an output also known as *activation level* which I denote by a. The activation level a will then be passed on as an input to a neuron that it is connected to in the next layer. The neurons of the last layer will compute the output level taking the activation levels of the connected neurons of the previous layer.

Consider a simple neural network architecture depicted in Figure 2. It has three layers—layer 0: input layer, layer 1: hidden layer, and layer 2: output layer. Last layer in the text is denoted by L, and hence $L = 2$. Layer 0 has three input neurons. The second layer has 4 neurons. and last layer has two neurons corresponding to the two output levels, in our case probability of two events. In this neural network, the hierarchical function specification is of the form:

$$f(x;w) = \sigma^2\left(z^2\left(\sigma^1\left(z^1(x,w^1)\right), w^2\right)\right) \equiv f^2_{w^2} \circ f^1_{w^1}(x). \tag{14}$$

The function $z^i(a^i, w^i) = w^i \cdot a^{i-1}$ at each layer i is a linear aggregator. In the notation, z^i is a vector of functions, each component of which corresponds to a neuron of the ith layer. The function σ^i is a squashing function of the same dimension as z^i, each component having the same function real valued function of one variable, known as *activation function*, which squashes the value of z^i to a range such as $(0,1)$ for a sigmoid activation function, or $(-1,1)$ for the *tanh* activation function. The value of the activation function, $a^i = \sigma^i(z^i)$ is known as the *activation level* of the neurons of the ith layer. The activation levels of the 0th layer, $a^0 = x$, the inputs, fed to the neural network from outside. The operation on the right is also performed component-wise for each neuron at the ith layer it computes the weighted sum of the activation levels (outputs) of the neurons of the previous layer that the neuron of the ith layer is connected to. The weights used are specific to the neuron of the ith layer. An activation function σ^i which generally taken to be same for all the neurons of the ith layer) is applied to this aggregated value z^i. These activation functions do not have any unknown parameters that need to be estimated. These two computations—aggregation and activation—are shown as a composite mapping $f^i_{w^i}$ for the neurons of the ith layer.

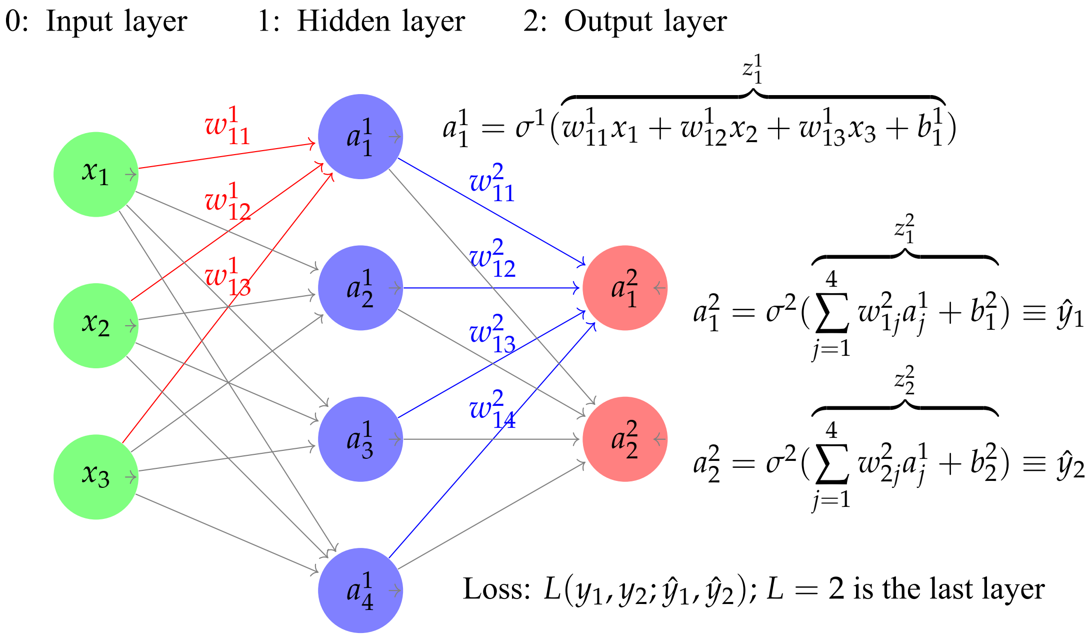

0: Input layer
1: Hidden layer
2: Output layer
x_1
x_2
x_3
w_{11}^1
w_{12}^1
w_{13}^1
a_1^1
a_2^1
a_3^1
a_4^1
w_{11}^2
w_{12}^2
w_{13}^2
w_{14}^2
a_1^2
a_2^2
$a_1^1 = \sigma^1(\overbrace{w_{11}^1 x_1 + w_{12}^1 x_2 + w_{13}^1 x_3 + b_1^1}^{z_1^1})$
$a_1^2 = \sigma^2(\overbrace{\sum_{j=1}^{4} w_{1j}^2 a_j^1 + b_1^2}^{z_1^2}) \equiv \hat{y}_1$
$a_2^2 = \sigma^2(\overbrace{\sum_{j=1}^{4} w_{2j}^2 a_j^1 + b_2^2}^{z_2^2}) \equiv \hat{y}_2$
Loss: $L(y_1, y_2; \hat{y}_1, \hat{y}_2)$; $L = 2$ is the last layer

Figure 2: MLP architechture

There are different types of neural network, depending on the functional form of f in Eq. (13). A*deep feed forward neural network*, also known as a *feedforward neural network with hidden layers* or as a *multilayer perceptron (MLP)* is a network architecture in which there is no feedback of any neurons to itself or to others in the same layer. The activation levels of the neurons only feed forward to the neurons in the next layer. This is the reason also why these type of neural networks are called feed forward, as opposed to the recurrent neural network consider later that allow feedback. The MLP has good computational properties and an MLP is a great universal functional approximator: It is shown (Hornik et al., 1989) that with a sufficient number of layers in a hidden layer can approximate a function to any level of precision desired. So a MLP can be used to approximate the true data generating process as closely as one wants. How does one find such a network, i.e., how does one choose the weights of the network.

To get a good approximation, the artificial neural network contains hundreds of thousands of deep parameters w, how does one train the network, i.e., how to learn the the parameter values. The learning is done by choosing the weights to minimize a loss function together with a nonnegative regularization term (in statistical term a regulerization term corresponds to shrinkage estimator).

$$L(y, f(x, w)) + \lambda C(w). \tag{15}$$

In the present context of learning about a probability distribution of time-to-events, an appropriate loss function is to take the negative log-likelihood of the sample and the additive regularization term $C(w)$ to be $||w||_2$. In this case the loss function with regularization term has a Bayesian interpretation—it is the posterior log-likelihood with normal prior distribution for the parameters w. The choice of w to minimize the loss is done by a gradient descent method. The neural network architecture Eq. (13) yields a very convenient fast and automatic computation of the gradients $\partial L/\partial w$ using an algorithm known as back-propagation algorithm, used pretty much in all types of neural networks. Two steps in this algorithm are assume an initial value of the weights w, first compute all the activation levels starting at the input layer, i.e., $f_w^1(x)$ forward through the layers $2, 3, ..., L$ in Eq. (13), i.e., go through layer superscripts forward. In the next step, compute the gradients of the weight parameters w of various layers, starting from the last layer, backward in layers, i.e., decreasing order in the subscripts in Eq. (13). Once all gradients are computed, weights are adjusted using a steepest descent algorithm. Further details are omitted, since I do not implement any type of back propagation algorithm in this paper.

4.2 Previous neural network models of survival analysis

The main objective of the neural network approach to survival analysis is the relaxation of restrictive proportional hazard assumption in Eq. (7). The following papers use a MLP architecture to estimate the survival probabilities with covariates fixed at the beginning of the study period. Faraggi and Simon, 1995; Biganzoli et al., 1998; Fernandez et al., 2016; Katzman et al., 2018; Lee et al., 2018; Ranganath et al., 2016; Ren et al., 2019; Zhao et al., 2019. Except for Lee et al., 2018 who consider a competing risk model, all other papers consider two-state, alive-death, models.

Two main limitations of the previous literature is that they use feed-forward neural network architecture to specify the functional approximator for the hazard function in Eq. (7). These are static model in the sense that the input vectors of an individual are given at the begin of the period and the neural network model predict the probabilities of the event for a fixed number of periods. This is problematic because prediction of probabilities at a given period use all the information from the future periods. It cannot handle training using data in which some individuals do not have information for all periods. More importantly, these models cannot incorporate new information that may come after training the model to update the time-to-event prediction probabilities. Another shortcoming of these models is

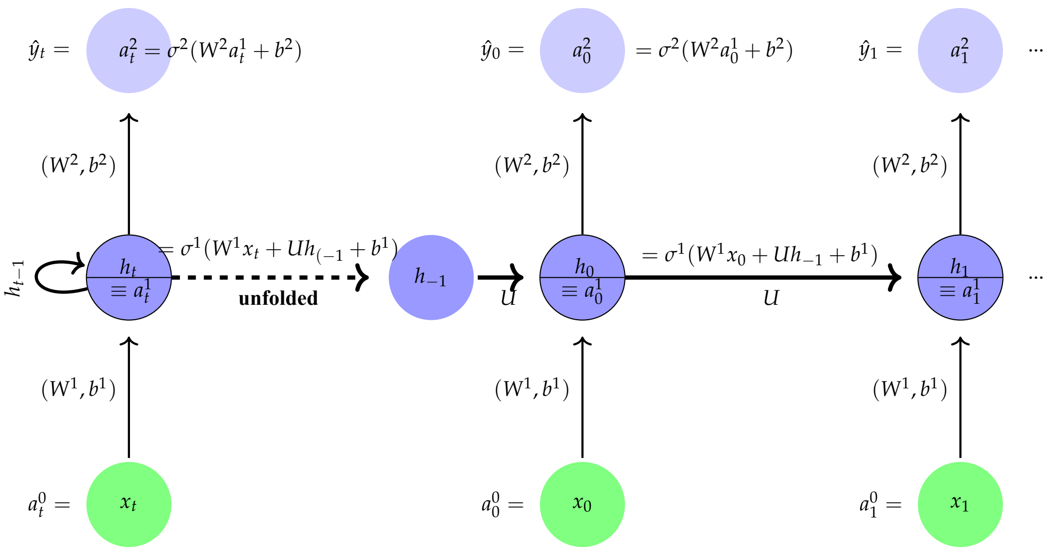


Figure 3: Showing a simple recurrent neural, with layer 0 is the input layer, layer 1 is the recurrent hidden layer, layer 2 is another pure hidden layer, layer 3 is the output layer, for each time point (t), layers are denoted as superscipt, and the subscript in parenthesis is time.

that they are restricted to analyzing life histories of only one transition between two health states—from alive to death or at the most to various causes of death in the language of survival analysis. As I have mentioned in the introduction, to better predict the probabilities of time to certain health come of interest, it is important to analyze dynamic paths through other intermediary health outcomes and the factors that affect the dynamics of those paths. I argue that a long-short memory recurrent neural network is more appropriate model to that end. I explain this framework next.

4.3 Formulation within Recurrent Neural Network with LSTM

Let x_t be a set of measurements, or characteristics and y_t, a set of health outcomes are measured repeatedly over time $t, t = 0, 1, ..T$. A recurrent neural network (RNN) allows a feedback link among neurons in a hidden layer. A typical RNN is shown in the left panel and the unfolded version in the right panel of Figure 3. This neural network is a simple three layer architecture similar to MLP architecture shown earlier, with one main difference is that the hidden layer has a feedback connection linked to itself, i.e., a cyclical or recurrent connection and there is weight sharing, i.e., the weights of all the layers for a given time period t is constant for all time periods. The recurrent connections in the RNN architecture effectively creates an internal memory of the effects of previous inputs that can affect the current and future outcomes. It is also interesting to note that the functional forms of the RNN are similar to Kalman filtering models, but more general in some respects and restrictive in some other respects.

The hidden layer 1: The activation levels of the hidden recurrent layer at time-step t is produced by combining activation levels h_{t-1} of the previous period's hidden recurrent layer and input levels x_t and then applying the activation function σ^1 as follows: For $t = 0, 1, 2, ...T$

Layer 1 (hidden recurrent layer):

$$\begin{aligned} h_t &= \sigma^1(W^1 \cdot x_t + U \cdot h_{t-1} + b^1) \\ a_t^1 &\equiv h_t \text{ (for notational convenience)} \qquad (16) \\ h_{-1} &= \text{ a user supplied initial activation level} \end{aligned}$$

Layer 2 (Output):

$$\begin{aligned} a_t^2 &= \sigma^2(W^2 \cdot a_t^1 + b^\ell) \\ \hat{y}_t &\equiv a_t^2 \text{ (notational convenience)} \qquad (17) \end{aligned}$$

The pseudo computational graph of the network is shown in Figure 3. It may seem a minor difference between the architectures of an MLP and an RNN, but difference in their capabilities are very important and crucial for sequential learning. For instance, an RNN can map the entire history of previous inputs to each output as compared to an MLP which can map only current input to output vectors. Similar to the result that an MLP is an universal function approximator mentioned earlier, with a sufficiently large number of hidden self-connected neurons, an RNN can approximate extremely closely any sequence-to-sequence mapping, see Graves et al., 2014; Hammer, 2000; Siegelmann and Sontag, 1992. There are various types of RNN, for details see Graves, 2012; Lipton et al., 2015. In many situations, especially in drug-discovery, the study samples may have small size, an important extension of an RNN framework that handles this problem is Altae-Tran, 2016; Altae-Tran et al., 2017.

The computation of the gradients is done using a modification of the back propagation method known as back propagation through time, which involves multiplication of gradients less than one many times over the time steps tending to vanishing gradient problem, meaning the gradient of the weights representing memories of the far back in time tend to

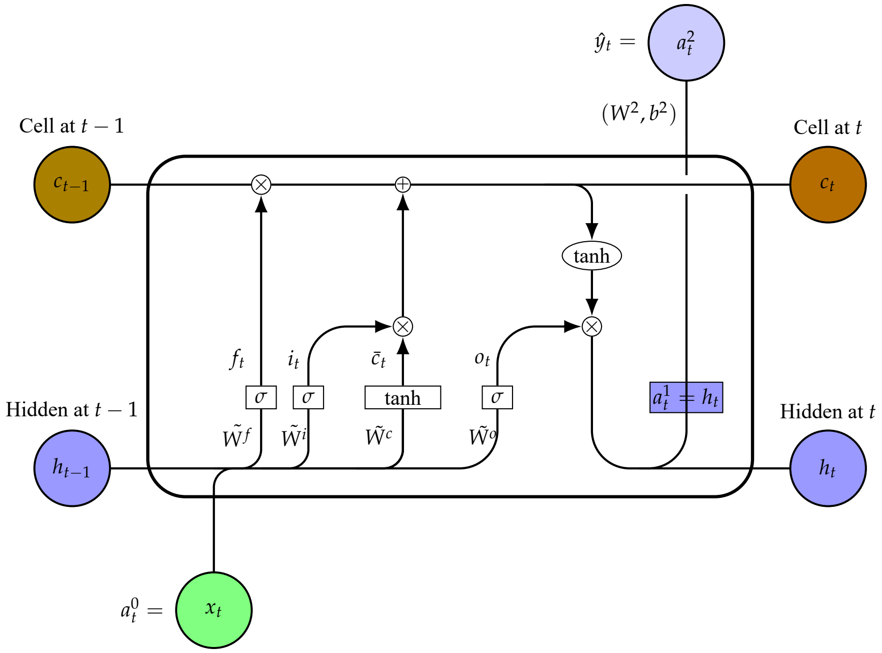


Figure 4: Showing a typical memory cell at time-step t

zero or vanish. A similar problem arises when the gradients larger than 1 are multiplied many times leading to exploding gradient problem.

Quite a few other fixes are proposed for the vanishing and exploding gradient problems of RNN models, such clipping of gradients, and various other types of extensions such as Elman, 1990 are proposed. But the long-short memory extension of the RNN framework by Hochreiter and Schmidhuber, 1997 proved to be very useful and successful in many applications which I use for this paper.

4.4 LSTM memory cell

The main innovation in this approach is to replace the hidden recurrent cells in an RNN with a memory cell with three three gates—an input gate, an output gate and a forget get—and a recurrent cell memory together with the original hidden layer memory cell. The functional form of a typical LSTM memory cell is as follows

Layer 1 (Hidden layer with recurrent neurons):

$$\begin{aligned} f_t &= \sigma\left(W^f \cdot x_t + U^f \cdot h_{t-1} + b^f\right) \\ i_t &= \sigma\left(W^i \cdot x_t + U^i \cdot h_{t-1} + b^i\right) \\ o_t &= \sigma\left(W^o \cdot x_t + U^o \cdot h_{t-1} + b^o\right) \\ \bar{c}_t &= tanh(W^c \cdot x_t + U^c \cdot h_{t-1} + b^c) \\ c_t &= f_t * c_{t-1} + i_t * \bar{c}_t \\ h_t &= o_t * tanh(c_t) \\ a_t^1 &\equiv h_t \text{ (for notational convenience)} \\ h_{-1}, c_{-1} &= \text{ user supplied initial activation levels} \end{aligned} \tag{18}$$

Layer 2 (Output): This is same as the RNN output layer.

$$\begin{aligned} a_t^2 &= \sigma^2(W^2 \cdot a_t^1 + b^\ell) \\ \hat{y}_t &\equiv a_t^2 \text{ (notational convenience)} \end{aligned} \tag{19}$$

The pseudo computational graph of the network is shown in Figure 4.

5 Prediction and performance of the model

As a multistate time to event model has censored data, the standard metrics such as R^2, root mean square error are not applicable for performance measurement. Harrell et al., 1982 introduce the measure of concordance index, known as c-index, extending the concept of the area under the ROC curve, which was further refined by Antolini et al., 2005 to be applicable for survival models with time-varying covariates. Most machine learning models use this criteion to compare the performance of various neural network models of survival analysis.

To get the basic idea behind the c-index, suppose two individual in the dataset are in a particular health state, both of whom are at risk for exits to another possible health state. Suppose the first individual exits at time t and the second individual did not exit at time t or earlier. The first individual's estimated survival probability at time t should be smaller than the second individual who did not exit. A good model should have this maintained for most or all such comparable pairs. The c-index measures the proportion people for whom this is true out of all the people who could be compared.

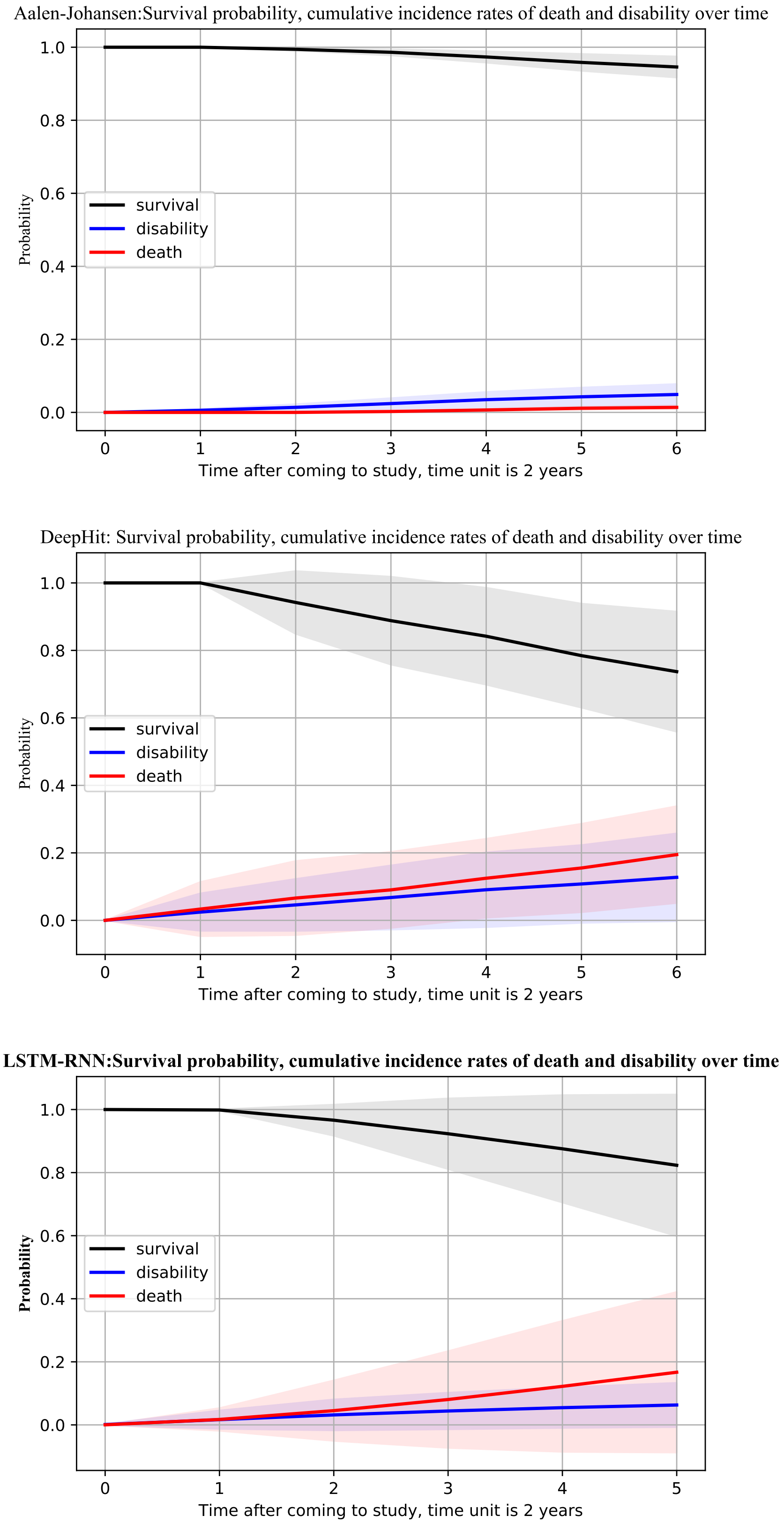
I use the c-index criterion to discriminate the performance of various models in predicting time-to-event probabilities. The c-index estimates for three three models are shown in the last row of Table 6. Judging from the c-index estimates, it appears that neural network models do better job in prediction of time to event probabilities than Aalen-Johansen estimators of the multistate statistical model. Furthermore, the LSTM-RNN model of this paper does perform slightly better.

I also show in Table 6 the average of the predicted time to event probabilities of the individuals in the test data set, and in Figure 5, I plotted the average probabilities together with the confidence intervals. Looking at the variability in the graph, and direct comparison of probabilities for the individual cases, it appears that neural network models might be over-fitting the data, in spite of using the L_2 regularization. More investigation is needed.

So I based my comparison of the models on the c-index.

Table 6: Average of the predicted cumulative incidence rates of disability and death in the test sample

up to time	Statistical multistate model		Lee etal Deephit model		LSTM-RNN model	
	Disability	Death	Disability	Death	Disability	Death
0	0.00000	0.00000	0.00000	0.00000	0.00119	0.00028
1	0.00573	0.00000	0.03487	0.00001	0.01658	0.01725
2	0.01383	0.00000	0.05179	0.00396	0.03165	0.04515
3	0.02420	0.00250	0.12043	0.02575	0.04405	0.08047
4	0.03489	0.00666	0.24674	0.04990	0.05461	0.12219
5	0.04276	0.01136	0.27139	0.13317	0.06310	0.16697
6	0.04906	0.01379	0.28676	0.15479	0.07025	0.20781
7	0.05171	0.01705	0.49585	0.50415	0.07734	0.24754
c-index	0.476290706		0.74824416		0.755676489	


Aalen-Johansen:Survival probability, cumulative incidence rates of death and disability over time

DeepHit: Survival probability, cumulative incidence rates of death and disability over time

LSTM-RNN:Survival probability, cumulative incidence rates of death and disability over time

Figure 5: Means and mean ± standard deviation intervals of the predicted cumulative incidence rates of disability and death in the test sample

6 Conclusion

Individuals follow different health trajectories over their lives. Many factors along a health trajectory up to a given time period affect their health state at the time and the future paths that the individual will follows. Such sequential health developments conditioned by the health related behaviors along the path determine the risks of diseases, disability and death. This paper studies statistical and artificial neural network models for predicting time to an event of interest—disability—with death as a competing risk event during the mid ages and how different factors along the health trajectory affect these probabilities. The paper uses the Health and Retirement Study data for estimation of the models. The paper discusses the merits and demerits of these two approaches and compare their performances.

The paper formulates the life histories of individuals as a stochastic process through a finite set of health states. The main problem is to estimate the transition probabilities between health states and the effects of various time varying covariates on transition probabilities. The effect of covariates is parameterized through 'transition intensity functions' in both statistical and neural network models. Statistical multistate models assume that in each transition, the covariates linearly affect a latent health measure and that latent health measure shifts a baseline hazard function (or transition intensity function) proportionately up and down by a multiplicative factor. This type of specification of transition intensities is known as Cox proportional hazard model. In many situations, including the present, this proportionality assumption imposes strong structure and could be far from the structure of the true data generating process. The multi layer perceptron (MLP), a kind of deep neural network (explained in the paper), is known to be an excellent 'universal function approximator'. Utilizing this insight, recently a few papers use MLP models to relax the proportionality assumption. This paper argues that the prediction of progression through health states and estimation of the probabilities of health events of interest such as disability and death are better done in a recurrent neural network (RNN) model with long-short memory cells to capture the effects of long lagged health related behaviors and outcomes that affect the current and future health outcomes. The paper compares the performances of a statistical multistate model, and an MLP based competing risk model of Lee et al., 2018 and the LSTM-RNN model proposed in this paper. The paper finds that performance, measured by c-index, is much better for both the neural network models and is slightly better for the LSTM-RNN model. Since an LSTM-RNN neural model is more suitable for prediction in a sequential set-up, the findings prescribe that it is better to use an LSTM-RNN type of neural network model to estimate and predict transition probabilities than an MLP based neural network model.

For all three models, the paper included some epigenetic factors (that include health related behaviors), demographic factors, education level, some biomarkers like BMI, CESD and cognition and depression and stress level—time varying for the LSTM-RNN and multi-state statistical model and time fixed at the first period for the Lee et al. model. At present there are no neural network software package that can estimate and compare the relative importance of the covariates on transition probabilities. The statistical models are good at that. More research in neural network along this line will be useful.

From the estimates of the statistical model, the paper finds that college graduates have significantly lower probability of all the transitions. The variable CESD measuring the level of depression and stress has significant positive effects on transiting from normal health to diseased health state, from normal health to becoming disabled and from diseased health state to become disabled or to death. The other most significant behavioral variables are smoking and sufficiently vigorous level of regular exercising. The smoking has significantly adverse effects and exercising has favorable effects on most transitions.

For unfamiliar readers, the paper gives introduction to estimation of multistate stochastic models and the MLP and LSTM-RNN type of neural networks models when the data has censored observations.

While a statistical model is good for studying the effects of various covariates on time to event probabilities, judging from the performance based on the c-index, its performance as a predictive model is much worse than the two neural network models considered. Relaxation of the proportionality assumption to include more general specifications is a useful direction. For neural network models, it will be very useful to study the relative strength and statistical significance of various covariates on the outcomes such as transition probabilities through different health states that the models predict.

References

[1] Aalen, O. O., Ø. Borgan, and H. K. Gjessing (2008). Survival and Event History Analysis. Springer New York. DOI: 10.1007/978-0-387-68560-1 (cit. on p. 14).

[2] Aalen, O. O. and S. Johansen (1978). An Empirical Transition Matrix for Non-Homogeneous Markov Chains Based on Censored Observations, *Scandinavian Journal of Statistics*, **5**, no. 3, 141–150 (cit. on pp. 6, 13).

[3] Alisch, R. S., B. G. Barwick, P. Chopra, L. K. Myrick, G. A. Satten, K. N. Conneely, and S. T. Warren (Feb. 2012). Age-associated DNA methylation in pediatric populations, *Genome Research*, **22**, no. 4, 623–632. DOI: 10.1101/gr.125187.111 (cit. on p. 4).

[4] Altae-Tran, H. (2016). RNNs for Model Predictive Control in Unknown Dynamical Systems with Low Sampling Rates, *working paper, Stanford University* (cit. on pp. 7, 26).

[5] Altae-Tran, H., B. Ramsundar, A. S. Pappu, and V. Pande (2017). Low Data Drug Discovery with One-Shot Learning, *ACS Central Science*, **3**, no. 4, 283–293. DOI: 10.1021/acscentsci.6b00367 (cit. on pp. 7, 26).

[6] Andersen, P. K., Ø. Borgan, R. D. Gill, and N. Keiding (1993). Statistical Models Based on Counting Processes. Springer-Verlag, New York (cit. on pp. 6, 14).

[7] Andersen, P. K. and M. P. Perme (Sept. 2008). Inference for outcome probabilities in multi-state models, *Lifetime Data Analysis*, **14**, no. 4, 405–431. DOI: 10.1007/s10985-008-9097-x (cit. on pp. 6, 14).

[8] Antolini, L., P. Boracchi, and E. Biganzoli (2005). A time-dependent discrimination index for survival data, *Statistics in Medicine*, **24**, no. 24, 3927–3944. DOI: 10.1002/sim.2427 (cit. on p. 28).

[9] Austad, S. N. and K. E. Fischer (2016). Sex Differences in Lifespan, *Cell Metabolism*, **23**, no. 6, 1022–1033. DOI: 10.1016/j.cmet.2016.05.019 (cit. on p. 3).

[10] Barbara, M. A., Y. Abdilla, and J. Calleja-Agius (May 2017). An Introduction to Epigenetics, *Neonatal Network*, **36**, no. 3, 124–128. DOI: 10.1891/0730-0832.36.3.124 (cit. on p. 5).

[11] Barker, D. J. P. (May 2007). The origins of the developmental origins theory, *Journal of Internal Medicine*, **261**, no. 5, 412–417. DOI: 10.1111/j.1365-2796.2007.01809.x (cit. on p. 5).

[12] Barker, D. J. P. (1990). The fetal and infant origins of adult disease. *BMJ: British Medical Journal*, **301**, no. 6761, 1111 (cit. on p. 4).

[13] Barker, D. J. P. (Aug. 1998). In utero programming of chronic disease, *Clinical Science*, **95**, no. 2, 115–128. DOI: 10.1042/cs0950115 (cit. on p. 4).

[14] Barondes, S. (1999). Molecules and Mental Illness. Scientific American Library (cit. on p. 3).

[15] Barres, R. and J. R. Zierath (Apr. 2011). DNA methylation in metabolic disorders, *The American Journal of Clinical Nutrition*, **93**, no. 4, 897S–900S. DOI: 10.3945/ajcn.110.001933 (cit. on p. 4).

[16] Biganzoli, E., P. Boracchi, L. Mariani, and E. Marubini (1998). Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach, *Statistics in Medicine*, **17**, no. 10, 1169–1186. DOI: 10.1002/(SICI)1097-0258(19980530)17:10<1169::AID-SIM796>3.0.CO;2-D (cit. on pp. 6, 24).

[17] Blair, S. N., I. Kohl Harold W., J. Paffenbarger Ralph S., D. G. Clark, K. H. Cooper, and L. W. Gibbons (Nov. 1989). Physical Fitness and All-Cause Mortality: A Prospective Study of Healthy Men and Women, *JAMA*, **262**, no. 17, 2395–2401. DOI: 10.1001/jama.1989.03430170057028 (cit. on p. 3).

[18] Boks, M. P., E. M. Derks, D. J. Weisenberger, E. Strengman, E. Janson, I. E. Sommer, R. S. Kahn, and R. A. Ophoff (Aug. 2009). The Relationship of DNA Methylation with Age, Gender and Genotype in Twins and Healthy Controls, *PLoS ONE*, **4**, no. 8. Ed. by J. Najbauer, e6767. DOI: 10.1371/journal.pone.0006767 (cit. on p. 4).

[19] Bookman, E. B., K. McAllister, E. Gillanders, K. Wanke, D. Balshaw, J. Rutter, J. Reedy, D. Shaughnessy, T. Agurs-Collins, D. Paltoo, and et al. (2011). Gene-environment interplay in common complex diseases: forging an integrative model-recommendations from an NIH workshop, *Genetic Epidemiology*, n/aâ€“n/a. DOI: 10.1002/gepi.20571 (cit. on p. 3).

[20] Bugliari, D., N. Campbell, C. Chan, O. Hayden, M. Hurd, R. Main, J. Mallett, C. McCullough, E. Meijer, M. Moldoff, P. Pantoja, S. Rohwedder, and P. St.Clair (2016). *RAND HRS Data Documentation, Version P*. Tech. rep. RAND Center for the Study of Aging (cit. on p. 15).

[21] Commenges, D. (Apr. 2002). Inference for multi-state models from interval-censored data, *Statistical Methods in Medical Research*, **11**, no. 2, 167–182. DOI: 10.1191/0962280202sm279ra (cit. on p. 14).

[22] Conti, R. M., E. R. Berndt, and R. G. Frank (July 2009). “Early Retirement and DI/SSI Applications: Exploring the Impact of Depression”, *Health at Older Ages: The Causes and Consequences of Declining Disability among the Elderly*. NBER Chapters. National Bureau of Economic Research, Inc, 381–408 (cit. on p. 5).

[23] Crowther, M. J. and P. C. Lambert (2017). Parametric multistate survival models: Flexible modelling allowing transition-specific distributions with application to estimating clinically useful measures of effect differences, *Statistics in Medicine*, **36**, no. 29, 4719–4742. DOI: 10.1002/sim.7448 (cit. on p. 6).

[24] DiLoreto, R. and C. T. Murphy (Dec. 2015). The cell biology of aging, *Molecular Biology of the Cell*, **26**, no. 25. Ed. by W. Bement, 4524–4531. DOI: 10.1091/mbc.e14-06-1084 (cit. on p. 4).

[25] Elman, J. L. (1990). Finding structure in time, *Cognitive Science*, **14**, no. 2, 179–211. DOI: 10.1016/0364-0213(90)90002-E (cit. on p. 27).

[26] Epel, E. S., E. H. Blackburn, J. Lin, F. S. Dhabhar, N. E. Adler, J. D. Morrow, and R. M. Cawthon (2004). Accelerated telomere shortening in response to life stress, *Proceedings of the National Academy of Sciences*, **101**, no. 49, 17312–17315. DOI: 10.1073/pnas.0407162101 (cit. on p. 4).

[27] Esteller, M. (Mar. 2008). Epigenetics in Cancer, *New England Journal of Medicine*, **358**, no. 11, 1148–1159. DOI: 10.1056/nejmra072067 (cit. on p. 4).

[28] Faraggi, D. and R. Simon (Jan. 1995). A neural network model for survival data, *Statistics in Medicine*, **14**, no. 1, 73–82. DOI: 10.1002/sim.4780140108 (cit. on pp. 6, 24).

[29] Fernandez, T., N. Rivera, and Y. W. Teh (2016). "Gaussian Processes for Survival Analysis", *Advances in Neural Information Processing Systems 29*. Ed. by D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett. Curran Associates, Inc., 5021–5029 (cit. on p. 24).

[30] Fisher, G. G. and L. H. Ryan (Dec. 2017). Overview of the Health and Retirement Study and Introduction to the Special Issue, *Work, Aging and Retirement*, **4**, no. 1. Ed. by M. Wang, 1–9. DOI: 10.1093/workar/wax032 (cit. on p. 15).

[31] Fleming, T. R. (1978). Nonparametric Estimation for Nonhomogeneous Markov Processes in the Problem of Competing Risks, *The Annals of Statistics*, **6**, no. 5, 1057–1070 (cit. on pp. 6, 13).

[32] Fleming, T. R. and D. P. Harrington (2005). Counting Processes and Survival Analysis. Wiley (cit. on p. 14).

[33] Gill, R. D. (2005). "Product-integration", *Encyclopedia of Biostatistics*. American Cancer Society. DOI: 10.1002/0470011815.b2a11058 (cit. on p. 12).

[34] Gill, R. D. and S. Johansen (1990). A Survey of Product-Integration with a View Toward Application in Survival Analysis, *The Annals of Statistics*, **18**, no. 4, 1501–1555 (cit. on p. 12).

[35] Gluckman, P. D., M. A. Hanson, C. Cooper, and K. L. Thornburg (July 2008). Effect of In Utero and Early-Life Conditions on Adult Health and Disease, *New England Journal of Medicine*, **359**, no. 1, 61–73. DOI: 10.1056/nejmra0708473 (cit. on p. 4).

[36] Graves, A. (2012). Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg. DOI: 10.1007/978-3-642-24797-2 (cit. on pp. 7, 26).

[37] Graves, A., G. Wayne, and I. Danihelka (2014). Neural Turing Machines, *CoRR*, **abs/1410.5401** (cit. on pp. 7, 26).

[38] Hammer, B. (2000). On the approximation capability of recurrent neural networks, *Neurocomputing*, **31**, no. 1, 107–123. DOI: https://doi.org/10.1016/S0925-2312(99)00174-5 (cit. on pp. 7, 26).

[39] Hannum, G. et al. (2013). Genome-wide Methylation Profiles Reveal Quantitative Views of Human Aging Rates, *Molecular Cell*, **49**, no. 2, 359–367. DOI: https://doi.org/10.1016/j.molcel.2012.10.016 (cit. on p. 4).

[40] Harrell Frank E., J., R. M. Califf, D. B. Pryor, K. L. Lee, and R. A. Rosati (May 1982). Evaluating the Yield of Medical Tests, *JAMA*, **247**, no. 18, 2543–2546. DOI: 10.1001/jama.1982.03320430047030 (cit. on p. 28).

[41] Hayflick, L. (1965). The limited in vitro lifetime of human diploid cell strains, *Experimental Cell Research*, **37**, no. 3, 614–636. DOI: 10.1016/0014-4827(65)90211-9 (cit. on p. 4).

[42] Hochreiter, S. and J. Schmidhuber (1997). Long Short-Term Memory, *Neural Computation*, **9**, no. 8, 1735–1780. DOI: 10.1162/neco.1997.9.8.1735 (cit. on pp. 7, 27).

[43] Hornik, K., M. Stinchcombe, and H. White (Jan. 1989). Multilayer feedforward networks are universal approximators, *Neural Networks*, **2**, no. 5, 359â€“366. DOI: 10.1016/0893-6080(89)90020-8 (cit. on pp. 6, 23).

[44] Horvath, S. (Dec. 2013). DNA methylation age of human tissues and cell types, *Genome Biology*, **14**, no. 10, 3156. DOI: 10.1186/gb-2013-14-10-r115 (cit. on p. 4).

[45] Juster, F. T. and R. Suzman (1995). An Overview of the Health and Retirement Study, *The Journal of Human Resources*, **30**, S7. DOI: 10.2307/146277 (cit. on p. 15).

[46] Kanherkar, R. R., N. Bhatia-Dey, and A. B. Csoka (Sept. 2014). Epigenetics across the human lifespan, *Frontiers in Cell and Developmental Biology*, **2**, no. 3, 124–128. DOI: 10.3389/fcell.2014.00049 (cit. on p. 5).

[47] Kaplan, E. L. and P. Meier (1958). Nonparametric Estimation from Incomplete Observations, *Journal of the American Statistical Association*, **53**, no. 282, 457–481. DOI: 10.1080/01621459.1958.10501452 (cit. on p. 12).

[48] Karakus, M. C. and L. C. Patton (Feb. 2011). Depression and the Onset of Chronic Illness in Older Adults: A 12-Year Prospective Study, *The Journal of Behavioral Health Services & Research*, **38**, no. 3, 373–382. DOI: 10.1007/s11414-011-9234-2 (cit. on p. 5).

[49] Katzman, J. L., U. Shaham, A. Cloninger, J. Bates, T. Jiang, and Y. Kluger (Feb. 2018). DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, *BMC Medical Research Methodology*, **18**, no. 1, 24. DOI: 10.1186/s12874-018-0482-1 (cit. on pp. 6, 24).

[50] Khoury, M. J. et al. (June 2009). Genome-Wide Association Studies, Field Synopses, and the Development of the Knowledge Base on Genetic Variation and Human Diseases, *American Journal of Epidemiology*, **170**, no. 3, 269–279. DOI: 10.1093/aje/kwp119 (cit. on p. 3).

[51] Lee, C., W. R. Zame, J. Yoon, and M. van der Schaar (2018). “Deephit: A deep learning approach to survival analysis with competing risks”, *Thirty-Second AAAI Conference on Artificial Intelligence* (cit. on pp. 6, 21, 24, 31, 32).

[52] Lipton, Z. C., J. Berkowitz, and C. Elkan (2015). A critical review of recurrent neural networks for sequence learning, *arXiv preprint* (cit. on p. 26).

[53] Ranganath, R., A. Perotte, N. Elhadad, and D. Blei (2016). “Deep Survival Analysis”, *Proceedings of the 1st Machine Learning for Healthcare Conference*. Vol. 56. PMLR, 101–114 (cit. on pp. 6, 24).

[54] Raut, L. K. (2019a). Health Outcomes in Mid-Ages: Multistate time to event Statistical Models versus Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) Models, *Draft prepared for presentation at the 2020 ASSA Meetings, San Diego, January 3 - 5, 2020* (cit. on p. 3).

[55] Raut, L. K. (2017). Exits from Disability: Estimates from a Competing Risk Model, *Social Security Bulletin*, **77**, no. 3, 15–38 (cit. on p. 10).

[56] Raut, L. K. (2019b). Progression of health, mortality and morbidy in the aging process: genetics, epigenetics and fetal programming, *(mimeo) University of Chicago* (cit. on p. 3).

[57] Ren, K., J. Qin, L. Zheng, Z. Yang, W. Zhang, L. Qiu, and Y. Yu (2019). “Deep recurrent survival analysis”, *Proc. AAAI*, 1–8 (cit. on pp. 6, 24).

[58] Renna, F. (Oct. 2008). Alcohol Abuse, Alcoholism, and Labor Market Outcomes: Looking for the Missing Link, *ILR Review*, **62**, no. 1, 92–103. DOI: 10.1177/001979390806200105 (cit. on p. 5).

[59] Seib, C., E. Whiteside, K. Lee, J. Humphreys, T. H. D. Tran, L. Chopin, and D. Anderson (2014). Stress, Lifestyle, and Quality of Life in Midlife and Older Australian Women: Results From the Stress and the Health of Women Study, *Women's Health Issues*, **24**, no. 1, e43–e52. DOI: 10.1016/j.whi.2013.11.004 (cit. on p. 5).

[60] Shalev, I. and J. Belsky (2016). Early-life stress and reproductive cost: A two-hit developmental model of accelerated aging?, *Medical Hypotheses*, **90**, 41–47. DOI: https://doi.org/10.1016/j.mehy.2016.03.002 (cit. on p. 4).

[61] Shalev, I., S. Entringer, P. D. Wadhwa, O. M. Wolkowitz, E. Puterman, J. Lin, and E. S. Epel (2013). Stress and telomere biology: A lifespan perspective, *Psychoneuroendocrinology*, **38**, no. 9, 1835–1842. DOI: https://doi.org/10.1016/j.psyneuen.2013.03.010 (cit. on p. 4).

[62] Siegelmann, H. T. and E. D. Sontag (1992). "On the computational power of neural nets", *Proceedings of the fifth annual workshop on Computational learning theory - COLT '92*. ACM Press. DOI: 10.1145/130385.130432 (cit. on pp. 7, 26).

[63] Simons, R. L., M. K. Lei, S. R. Beach, R. A. Philibert, C. E. Cutrona, F. X. Gibbons, and A. Barr (2016). Economic hardship and biological weathering: The epigenetics of aging in a U.S. sample of black women, *Social Science & Medicine*, **150**, 192–200. DOI: https://doi.org/10.1016/j.socscimed.2015.12.001 (cit. on pp. 4, 5).

[64] Sonnega, A., J. D. Faul, M. B. Ofstedal, K. M. Langa, J. W. Phillips, and D. R. Weir (Mar. 2014). Cohort Profile: the Health and Retirement Study (HRS), *International Journal of Epidemiology*, **43**, no. 2, 576–585. DOI: 10.1093/ije/dyu067 (cit. on p. 15).

[65] Steffick, D. E. (2000). Documentation of affective functioning measures in the Health and Retirement Study, *Ann Arbor, MI: University of Michigan* (cit. on p. 17).

[66] Thornburg, K. L., J. Shannon, P. Thuillier, and M. S. Turker (2010). "In Utero Life and Epigenetic Predisposition for Disease", *Epigenetics and Cancer, Part B*. Elsevier, 57–78. DOI: 10.1016/b978-0-12-380864-6.00003-1 (cit. on p. 5).

[67] Vaupel, J. W. (Mar. 2010). Biodemography of human ageing, *Nature*, **464**, no. 7288, 536–542. DOI: 10.1038/nature08984 (cit. on p. 3).

[68] Wreede, L. C. de, M. Fiocco, and H. Putter (Sept. 2010). The mstate package for estimation and prediction in non- and semi-parametric multi-state and competing risks models, *Computer Methods and Programs in Biomedicine*, **99**, no. 3, 261–274. DOI: 10.1016/j.cmpb.2010.01.001 (cit. on p. 14).

[69] Zarulli, V., J. A. Barthold Jones, A. Oksuzyan, R. Lindahl-Jacobsen, K. Christensen, and J. W. Vaupel (2018). Women live longer than men even during severe famines and epidemics, *Proceedings of the National Academy of Sciences*, **115**, no. 4, E832–E840. DOI: 10.1073/pnas.1701535115 (cit. on p. 3).

[70] Zhao, R., R. Yan, Z. Chen, K. Mao, P. Wang, and R. X. Gao (2019). Deep learning and its applications to machine health monitoring, *Mechanical Systems and Signal Processing*, **115**, 213–237. DOI: 10.1016/j.ymssp.2018.05.050 (cit. on p. 24).