Deep Neural Network Machine Learning Models for Prediction of Risks of Various Diseases in Mid-Ages*

Lakshmi K. Raut[1]

[1]Visiting Fellow, Universityof Chicago
[1]lakshmiraut@gmail.com

2024-10-17

Abstract

Prediction of risks of various diseases and identification of factors that influence these risks are important for public policies and disease diagnosis in healthcare. The biomedical literature suggests that much of an individual's later life health outcomes are programmed at early stages of life. The programming is strongly modulated by epigenetic inputs throughout life such as psychological, financial, social or chemical stress, diets, smoking, substance use, and exercising, with stronger effects imparted in early stages of life. Traditionally predicting effects of these factors on risks of diseases is statistically examined using the logistic regression framework. Deep neural network models have shown superior predictive performances in other fields and can be used in the present context. This paper compares the effectiveness of these two approaches in quantitatively predicting these risks as a function of the observable variables and in identifying the influential variables that strongly affect the risks with the Health and Retirement Studies (HRS) data. I compare its predictive performance with performances of statistical procedures using confusion matrix and other indicators and then compare their predictions of policy outcomes.

Contents

*An earlier draft was presented at the 85th birthday celebration conference of Professor Kirith Parikh. Disclaimer: The views, thoughts, and opinions expressed in the paper belong solely to the author, and do not necessarily represent the views of any group or organization.

List of Figures

List of Tables

1 Introduction

The prediction of risks of various chronic diseases and identification of important factors for such risks are important issues in healthcare industry and public policies. For instance, in healthcare industry, the caregivers such as doctors and nurses may look at symptoms and medical history of an incoming patient to determine what diseases are most likely for the patient and then follow-up further tests and treatments. For public policy, prediction of risks of various diseases in communities can help build up infrastructure to deal with the health problems of the community. Various diseases affect the probability of disability, work loss and early death. Estimating those risks can help better policies for social insurance programs such as disability insurance programs and for private insurance programs to calculate the insurance premiums or for the regulators to regulate the private health insurance companies. What factors are important for development of diseases over a lifespan? Should we be using statistical models or machine learning predictive models for predictions? Public policies influence many of the individual characteristics that determine disease risks. From public policy perspective, it is also important to estimate the quantitative effect of such characteristics on the likelihood of various diseases, which can help designing policies to reduce health inequality among social groups or improve health of the general population. I use the Health and Retirement Surveys data for empirical exploration of the above issues in this paper. Determination of important factors for prediction or diagnosis of diseases with machine learning models is gaining importance in healthcare industry, see for instance, discussions in Ahmad et al. (2018).

The biomedical literature suggests that much of an individual's later life health outcomes are programmed at early stages of life. The programming is strongly modulated by the epigenetic inputs created by the environment in mother's womb at prenatal stage and by the environment at early postnatal ages. The most important epigenetic factor is stress of any kind–psychological, financial, social and chemical. Other significant factors are medical care, diets, smoking, substance use, and exercising. These modulating factors are important throughout life, with stronger effects imparted in early stages of life. At the cellular level, aging and the incidence of age-related diseases occur due to cellular senescence—i.e., after a certain number of cell divisions, it stops dividing or has defective replications, causing tissues or organs to increasingly deteriorate over time. What are the critical periods or the developmental milestones in life-cycle that program the motions of health developments over the lifespan of an individual?

Research along this line began with the striking findings of Barker (Barker, 1990; Barker, 1998) and later of Gluckman et al. (2008). They found strong association between birth weight and many later life chronic diseases, including hypertension, coronary artery diseases, type 2 diabetes, and osteoporosis. Many other studies find that much of health developments in later life is determined very early in life–specifically 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." Many studies in social sciences find that low socioeconomic 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 access to quality health care, see for evidence, Simons et al. (2016).

Generally prediction and estimation of risks of disease are carried out in statistical multinomial framework. In recent years machine learning techniques, especially the deep neural network predictive models are producing much superior predictions compared to the statistical models. For instance, Chen et al. (2017) used patient data from various hospitals in China to fit a neural network model and achieved high prediction accuracy rates. Machine learning techniques are used for prediction and selection of factors in cancer research, see Kourou et al. (2015) for a survey of these models. It is known that a feed-forward deep neural network model with a sufficient number of neurons in the hidden layers can approximate any function as closely as desired. That is, an MLP is one of the best universal function approximator, see (Hornik et al., 1989; Cybenko, 1989). To many social scientists, the machine learning techniques are mysterious. The purpose of this paper is to explain the structure of deep feed-forward neural network models, how they differ from statistical models in terms of prediction of risks and identification of important factors of the risks.

The rest of the paper is organized as follows. In Section 2, I describe two modeling paradigms, the multinomial logistic regression model from the statistics literature and the deep neural network model from the machine learning literature. In Section 3, I describe the Health and Retirement Study data set and the common set of variables that I use for fitting models in both frameworks. In Section 4, I report the numerical findings on predictive performance and variable importance. In Section 6, I discuss the issues in the dynamic context. Section 7 concludes the paper.

2 Two modeling paradigms

Empirical investigation of our main issues—prediction of risks of various diseases and identification of factors that influence these risks–involves two types of statistical models: explanatory models and predictive models. To explain it briefly using a unified general framework covering problems in other fields, I use the terminology and notation in Shmueli (2010). Most problems postulate a relationship, $\mathcal{Y} = \mathcal{F}(\mathcal{X})$, where $\mathcal{X}$ is a vector of certain input constructs related to a certain set of output constructs $\mathcal{Y}$ through some theoretical hypothesis or theories represented by $\mathcal{F}$. In most situations including the present, the underlying theory such as biology of aging and disease developments of humans provide qualitative relationships, not quantitative specifications in terms of measurable variables and a mathematical function. A statistical or econometric model operationalizes this relationship in terms of a vector of observable input random variables X and a vector of output random variables Y with a mathematical function $Y = f(X)$. Depending on data availability or collection problems, one may have many choices for the set of input variables X, and outcome variables Y and the functional form f relating them, each configuration gives one model (see for instance various types structural equation model specifications in Hair et al. (2017), and Pearl (2009)).

In the present context, $\mathcal{X}$ consists of individual genetic make-up and epigenetic factors over the life-span such as stressors that govern the cell divisions producing various health outcomes over time. These are not directly observable or we do not have information on individual genomes. I use various observable input variables X, some of which can be improved with public policies and some with individual behaviors. Our outcome health variable is a single discrete random variable Y denoting an individual's health status in middle age. The set of health statuses is {1,2,3,4,5}, where 1 = normal health, 2 = Cardiovascular disease, 3 = Cancer, 4 = Other single disease, 5 = comorbid diseases. These are discrete values without distinguishing the degree of severity. In an explanatory model, f is a causal function, i.e., it assumes X causes Y. In a predictive model f denotes the association between X and Y.

Both statistical models and neural network models approximate the unknown function $Y = f(X)$ with a parametric function $Y = f(X; \beta)$ from some parametric family of functions and chooses the best one that minimizes the expected value of a loss function, $\mathcal{L}(Y, f(X))$. Statistical models generally specify parametric functions in such a way that parameters can directly provide the quantitative effect of an input variable on the outcome variables. In the next subsection, I describe the most widely used such model, multinomial logistic regression model.This the statistical model I use in this paper.

2.1 The Multinomial logistic regression model

The multinomial logistic regression model assumes that the conditional probability distribution of Y given X belongs to a family of exponential distributions with parameters β, giving the following specification,

$$log\frac{P\ (Y = k|X)}{P\ (Y = 1|X)} = X'\beta_k, \quad k = 2, ..., 5. \tag{1}$$

Taking $\beta_1 = 0$, i.e., taking the normal health as the baseline or reference outcome, the above is equivalent to,

$$Y = \arg \max_{k=1,\ldots,5} \frac{\exp\left(X'\beta_k\right)}{\sum_{j=1}^{5} \exp\left(X'\beta_j\right)} \equiv f(X; \beta) \tag{2}$$

In the above specification, the parameter β_{jk}, the j^{th} component of β_k corresponding to the input variable X_j has the interpretation that a unit increase in X_j will change the log-odd of disease k by β_{jk}. That is, $exp(\beta_{jk})$ is the odds-ratio of outcome k and outcome 1 associated with a one-unit increase in the input X_j. In other words, a unit increase in X_j will change the likelihood of disease k by $exp(\beta_{jk})$ times the likelihood of normal health. This specification has the advantage that statistical estimate of the β_{jk} using a dataset provides both statistical and numerical significance of this variable, and thus helps one to decide which input variables are most significant for outcomes.

2.2 The deep neural network model

Neural network is a highly parameterized universal function approximator of the form $y = f(x; w)$, where 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 $\{(x_i, y_i), i = 1, ..., n\}$, where x_i is a vector of individual i's characteristics, and y_i is a vector of the individual's output levels and w is a set of parameters common to all individuals. 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 the unknown data-generating function using the observed data. This is the problem that both statistics and neural network deal with. In neural network, the problem is to design a neural network architecture of the approximating function $y = f(x, w)$ and find a suitable learning algorithm to learn the parameters w of the network using a training set of examples to minimize a loss function. This trained network can then be used to predict y for given characteristics x of any individual.

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 the approximated function to the true data-generating function. Most neural network models have the following type of hierarchical functional form:

$$y = f(x; w) \equiv f^L_{w^L} \circ .... \circ f^1_{w_1}(x). \tag{3}$$

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

For graphical illustration, consider a simple neural network architecture depicted in Figure 1. 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:

$$f(x; w) = \sigma^2 \left(z^2 \left(\sigma^1 \left(z^1(x, w^1)\right), w^2\right)\right) \equiv f^2_{w^2} \circ f^1_{w^1}(x). \tag{4}$$

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 i-th 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. An activation function squashes the value of z^i to a range such as to the range $(0, 1)$ by the sigmoid activation function ($f(x) = 1/(1+e^{-x})$), to the range $(-1, 1)$ by the tanh activation function ($\tanh(x) = 2\text{sigmoid}(2x) - 1$) and to the range $[0, \infty)$ by the most widely used ReLu function ($f(x) = \max(0, x)$). In many situations, better and faster results emerge from the ReLu function because it does not activate all the nodes in the following layers during training. I have mentioned about only a few widely used ones. There are many other ones. In fact, any function can be an activation function.

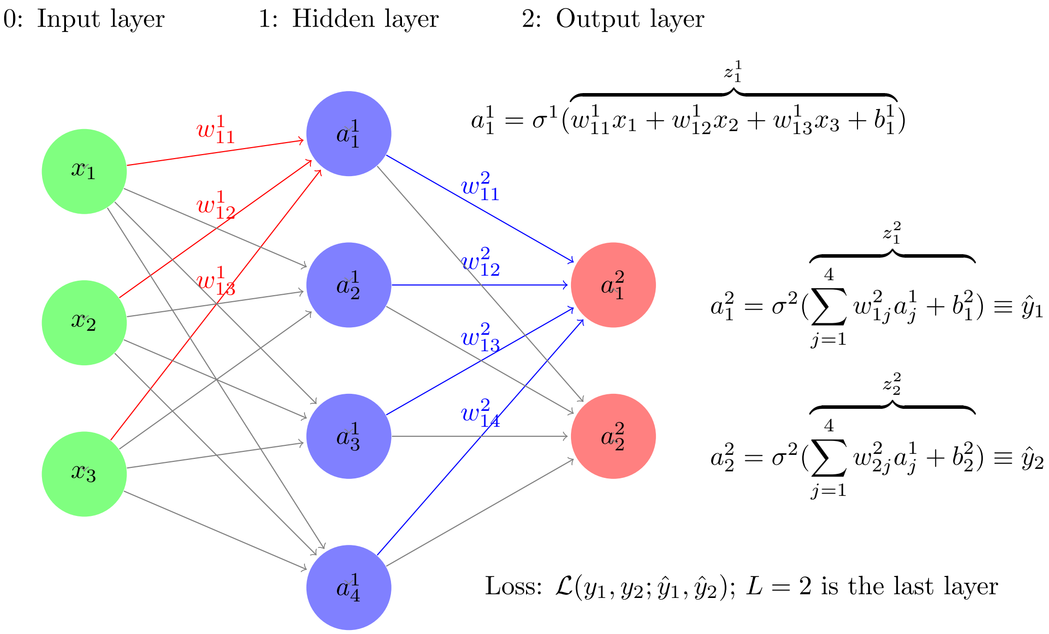
For the output layer, which is the last layer of the network, when the goal is to estimate the probability or the value of a binary outcome variable, one uses the sigmoid function; when the goal is to estimate the probability or the outcome of a categorical output variable, one uses the softmax function. Denoting a k dimensional vector as $\tilde{z} = (z_1, ..., z_k)$, the softmax function is a k variate function $\sigma(\tilde{z}) = (\sigma_1(\tilde{z}), ..., \sigma_k(\tilde{z}))$ defined as

$$\sigma_j(\tilde{z}) = \frac{e^{z_j}}{\sum_1^k e^{z_i}}, j = 1, .., k. \tag{5}$$

This is the function used in the multinomial logit model, see Eq. (2). Here z_j's are non-linear functions of input variables, e.g. the composite vector function $z_j \equiv z^2_j(X; w)$ in our example above and in the Figure 1, whereas in the multinomial logit model Eq. (2), it is a linear function $z_j \equiv X'\beta_j, j = 1, ..., k$.

The value of an activation function, $a^i = \sigma^i(z^i)$ is known as the activation level of the neurons of the i-th layer. The activation levels of the 0-th 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 i-th layer it computes the weighted sum of the activation levels (outputs) of the neurons of the previous layer that the neuron of the i-th layer is connected to. The weights used are specific to the neuron of the i-th layer. An activation function σ^i which generally taken to be same for all the neurons of the i-th 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 i-th layer.

Figure 1: A feed-forward deep neural network architecture.


0: Input layer
1: Hidden layer
2: Output layer
x_1
x_2
x_3
a^1_1
a^1_2
a^1_3
a^1_4
a^2_1
a^2_2
w^1_{11}
w^1_{12}
w^1_{13}
w^2_{11}
w^2_{12}
w^2_{13}
w^2_{14}
$a^1_1 = \sigma^1(\overbrace{w^1_{11}x_1 + w^1_{12}x_2 + w^1_{13}x_3 + b^1_1}^{z^1_1})$
$a^2_1 = \sigma^2(\overbrace{\sum_{j=1}^{4} w^2_{1j}a^1_j + b^2_1}^{z^2_1}) \equiv \hat{y}_1$
$a^2_2 = \sigma^2(\overbrace{\sum_{j=1}^{4} w^2_{2j}a^1_j + b^2_2}^{z^2_2}) \equiv \hat{y}_2$
Loss: $\mathcal{L}(y_1, y_2; \hat{y}_1, \hat{y}_2)$; $L = 2$ is the last layer

There are different types of neural networks, depending on the functional form of f in Eq. (3). A deep feed-forward neural network, also known as a feed-forward neural network with hidden layers or as a multi-layer perceptron (MLP) is a network architecture in which there is no feedback of any neurons to itself or to others in the same layer. See Goodfellow et al. (2016) for detailed descriptions of these terminologies and workings of the feed-forward deep neural network models of various types and Graves (2012) for recurrent neural models of various types, which I will not describe or use in this paper. The activation levels of the neurons only feed-forward to the neurons in the next layer. This is the reason also why these types of neural networks are called feed-forward neural network, as opposed to the recurrent neural network which allows feedback of a neuron to itself. The MLP has good computational properties and an MLP is a great universal functional approximator: It is shown by 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.

I take the categorical cross entropy as the loss function. I have tried the Kullback-Leibler divergence as the loss function and findings are similar and not reported.

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 best set of parameter values using the data at hand. The learning is done by choosing the weights to minimize a loss function together with a non-negative regularization term (in statistical term a regulerization term corresponds to a shrinkage estimator) to avoid over-fitting.

$$\frac{1}{n}\sum_{1}^{n}(\mathcal{L}(y_i - f(x_i; w))) + \lambda C(w). \tag{6}$$

In the present context of learning about a probability distribution over the discrete set of health outcomes, it is appropriate to take the loss function to be negative log-likelihood of the sample and the additive regularization term $C(w)$ to be $\|w\|_1$ known as $L1$ regularizer, or to be $\|w\|_2$, known as the L_2 regularizer or take a convex combination of the two known as the elastic-net or L_1L_2 regularizer. The choice of w to minimize the loss is done by a gradient descent method. The neural network architecture Eq. (3) yields a very convenient fast and automatic computation of the gradients $\partial\mathcal{L}/\partial w$ using an algorithm known as the back-propagation algorithm, used pretty much in all types of neural networks. The steps in this algorithm are as follows.

Backpropgation Algorithm

* Step 0: Assume an initial value for the weights w.
* Step 1: 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. (3), i.e., go through layer superscripts forward.
* Step 2: 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. (3).
* Step 3: Adjust weights using a steepest descent rule.
* Step 4: If the difference between the initial weights and these adjusted weights is within a tolerance limit, stop, and take these adjusted weights as the optimal weight estimate; otherwise, reset the initial weights to these adjusted weight, and go to Step 1.

2.3 Estimation and prediction methods in two paradigms

In a parametric model such as in Eq. (1), a parameter will provide the effect of the input variable on an outcome variable, only if the model is statistically identified in the sense that given $P(Y|X, \beta)$, there exists a unique β. By parameterizing the odds-ratio in Eq. (1), the multinomial logit model is statistically identified.

Whether a parametric model is statistically identified or not, it can produce predictions of outcomes, by predicting the outcome to be that which has the highest probability, i.e., $\arg\max_k P(Y = k|X, \beta)$. Other rules are possible too.

In statistics, to estimate parameters of a model from a sample of observations, one takes negative the log-likelihood function as the loss function $\mathcal{L}(Y, f(X; \beta)$ and minimizes the expected value of the loss function taken with respect to the empirical distribution, i.e., one maximizes $\sum_1^n(\mathcal{L}(y_i, f(x_i; \beta)))$, where $y_i, x_i, i = 1, ..., n$ is a random sample. This optimum estimator $\hat{\beta}$ is the maximum likelihood estimator. Under the assumption that the true data-generating process, i.e., the true f is a member of the parametric family, the maximum likelihood estimator $\hat{\beta}$ is asymptotically efficient, consistent, and asymptotically unbiased. In some cases, these properties hold in the small sample as well. These properties are utilized to carry various hypothesis testing about the parameters such as if a parameter or a set of parameters is statistically significant. For instance, we can look at the p-values to determine if a variable has statistically significant effect. This assumption about the true data-generating process is hardly true in real-life data. Breiman (2001b) provides a strong critique of this statistical modeling approach. He introduced algorithmic tree based methods such as CART and Random Forests. Other models along this line are the linear additive models (Hastie and Tibshirani (2006)), Support Vector Machines, and neural network model. These models can be found in Hastie, Tibshirani, and Friedman (2009) and Efron and Hastie (2016).

For the machine learning models including the deep neural network models, the estimated parameters do not have such interpretations; the models are generally not identified; and there is no straightforward procedure to identify the input variables with significant effects on the outcomes. For our deep neural network model, I adapt the permutation importance technique introduced by Breiman (2001a) for random forests models.

2.4 Computational details

For both maximum likelihood estimation of the multinomial logit model in Eq. (1) and the learning algorithm for the deep feed-forward neural network model in Eq. (3), I split the dataset into the training dataset containing a random sample without replacement of 80 percent of the original sample and the remains as the test dataset.

For both the logistic regression model and the deep neural network model, I use the training data to fit the model and use the test data to assess the performance of these two fitted models and finally use the fitted models on the training data to compare the predicted disease risks of various social groups and talk about their policy implications. These are reported in a section following the next section that describes the dataset and the variables used in this study.

I estimate the multinomial logit model R using the package nnet (Venables and Ripley (2002)). There are other R packages such as glmnet and mlogit that can also fit multinomial logistic models. The glmnet package, however, does not calculate the standard errors of the parameter estimates and mlogit package invloves complex data transformations.

To fit the feed-forward neural network model, I use the Tensorflow 2.3.0 (TensorFlow Developers (2021)) and Keras 2.4.3 (Chollet et al. (2015)) in Python. While the R package nnet can also do feed-forward neural network, it is limited to only one hidden layer and it does not incorporate early stopping, regularization, random dropout to find the best possible fit of a deep neural network. The Keras and Tensorflow are more versatile. Other machine learning packages such as Pytorch, Mxnet, Scikit Learn, can also be used. I found Tensorflow and Keras to be convenient for the task of this paper.

The feed-forward deep neural network I finally chose had two hidden layers of 32 and 512 neurons. Given we have 13 input variables, and 6 output neurons for the output layer, the network consists of 20,422 parameters, i.e., w's to train. I used categorical cross entropy as the objective function, which is equivalent of the log-likelihood of the multinomial logit model. To handle over fitting, I used the reLu activation function in the two hidden layers, elastic-net or L_1L_2 regularization of parameters in the two hidden layers, and early stopping, i.e., stopping early instead of training it to the maximum iterations (known as epochs) if the objection function stops improving early on. The results are reported and discussed after describing the dataset and the variables.

3 The Dataset and the construction of variables

3.1 The dataset

Table 1: Summary of the health status of the individuals in the sample over the survey years.

Survey Year	Alive:normal health	Alive: diseased	Became disabled	Died before disability	65+: censored	Total
1992	3016	6403	92	0	0	9511
1994	2591	6608	76	144	0	9419
1996	2291	6623	139	146	0	9199
1998	1726	5478	134	123	727	8188
2000	1279	4313	86	113	741	6532
2002	848	3148	54	58	759	4867
2004	468	1893	34	48	781	3224
2006	132	636	4	14	795	1581

I use the Health and Retirement Study (HRS) dataset for empirical analysis. A lot has been written about 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 from 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 was 13,593. Since 1992, the surveys have been repeated every two years, and 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. The RAND created many variables from the original HRS data for ease of use. I created all the variables (with a few exceptions noted below) from the RAND HRS dataset version P. The details of the Rand HRS version P can be found in Bugliari et al. (2016). I use the original cohort first interviewed in 1992 so that we have a homogeneous group of individuals with data for many years to avoid cohort effects in our analysis. This sample has the largest sample size.

The HRS data collected information on if a doctor diagnosed that the respondent had 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 were enrolled on to disability programs before the first survey year 1992 and I also drop the spouses in the sample who were not born between 1931 and 1941, so that the respondents in our sample are between ages 51 to 61 and are not disabled or dead by the first survey year 1992. I ended up with the final sample size of 9511 for this analysis.

The table reports these statistics only up to the survey year 2006, as the individuals exited the study because of disability or death before disability or censored because they are over age 65 after this survey year. The table shows that the first period of this study in 1992 has 3016 individuals, which is 32 percent of the sample, in good health, 6403 individuals (i.e., 67 percent) in diseased health state with one-or-more chronic diseases and 92 individuals (i.e., 1 percent) left the study as they become disabled. No individuals died or were censored because of ages higher than 65–this is the result of sample selection criterion mentioned above. In the next survey year 1994, out of 9419 non-exited individuals, 144 died without any disability. In the survey year 1998 for the first time, 727 individuals in the sample left our study because they reached ages above 65. The total number of individuals during the last survey round of 2006 before they all become older than 65 is 1581, i.e. about 17 percent of the original sample.

3.2 Variables

Molecular biology literature mentioned in the introduction points out that the stressors of the body cells are important determinants of the nature of cell divisions during early development and later life health outcomes.[1] While those stressors cannot be directly observed or measured, many socioeconomic factors modulate those stressors and cell developments, and thus affect later life health outcomes. Furthermore, early life health developments together with health-related behaviors are important determinants of later life health outcomes. Health behaviors are partly determined by cognitive and non-cognitive skills. Education level thus can affect health behaviors and health developments in later life. Education also determines earnings, which determine health-related expenditures and thus health outcomes. The HRS dataset does not have prenatal or postnatal data on individuals. It has a few variables on childhood socioeconomic status, which are correlates of the stressors of the cell developments.

How does one quantify childhood SES? There is no consensus on what exactly constitutes Childhood SES. Some studies use different sets of variables to represent Childhood SES. For instance, Heckman and Raut (2016) and a few other studies use parents' education as a measure childhood SES in modeling attainment of college degree. Luo and Waite (2005) used Father's and Mother's education and the Family financial well-being as regressors without aggregating them into a single measure to examine how these variables affect a measure of mid-age health outcomes for the HRS sample. It is useful to have a single measure of SES. Some studies used the latent variable approach to come up with a statistically defined measure of Childhood SES. For instance, Vable et al. (2017) used a number of variables from the HRS dataset to create their Childhood SES measure. Similar to their approach, I use the latent variable statistical procedure IRT on a set of parental characteristics during the childhood of the respondents.

Other important factors are biomarkers and mental health status in the mid-ages and health-related behaviors. Furthermore, the health development may vary by race and sex. I describe the construction of these variables in this subsection.

[1] Genetic make-up also controls gene expressions for producing proteins that create diseases but the epigenetic factors creating the stressors are important as well.

I use the Item Response Theory (IRT) from the latent variable analysis literature to construct an aggregate measure of childhood socioeconomic status, and two health-related behavioral traits, Smoking and Exercising.

IRT techniques are not commonly used in Economics. Originally the IRT techniques were used in the psychometry literature to measure latent traits such as cognitive ability and personality of individuals. More recently this technique has been used in health care fields to measure health status of individuals in clinical trials and treatments. In this procedure, the latent trait, known as score, is assumed to be a continuous variable and individuals differ in the levels of its possession. The procedure uses responses on a number of test items usually with true/false or with multiple choices to estimate the level of the latent trait that an individual possesses. The probability of a particular response to an item depends on the individual's trait level and on item characteristics such as difficulty level to answer objectively a question or the imperfection of the item question to measure the trait, or an individual might be guessing a response. The IRT procedure specifies a probability model of the responses to each item as a function of the level of the latent trait and item characteristics. The procedure uses various statistical methods to estimate the latent trait level and the characteristics of the item. Mainly three statistical estimation procedures are used in the literature–the maximum likelihood (ML) procedure, Bayesian maximum a posteriori (MAP) procedure and expected a posteriori (EAP) procedure. I have used a two parameter model (which includes the well known Rasch model as special case) of the probabilities of item responses and the MAP procedure to estimate the individual scores and the set of item parameters. I did this in SAS. See Embretson and Reise (2000) for a lucid exposition of the basic one-dimensional IRT models and the above three estimation procedures, see Cai et al. (2016) for a survey of IRT models of multi dimensional traits and extensions to dynamic scoring, and see An and Yung (2014) for details on the SAS IRT procedure and general introduction to various IRT procedures that SAS can perform.

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 completed college and above (does not include some college), i.e., has a college degree and more and taking value 0 otherwise.

CES-D: I used the score on the Center for Epidemiologic Studies Depression (CES-D) 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 CES-D score. In the paper, I refer the variable as CES-D. Steffick (2000) discusses its validity as a measure of stress and depression.

Cognitive scores: 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.

HIGH BMI : The variable body-mass-index (HIGH 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. Following the criterion in the literature, I create the variable HIGH BMI taking value 1 if HIGH BMI $>$ 25 and value 0 otherwise.

Now I describe the construction of the behavioral variables.

Smoking: 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.

Exercising: The RAND HRS has data on whether the respondent did vigorous exercise three or more days per week. I created in each time period to be 1 if the respondent did vigorous exercise three or more days per week in any of the waves and then that value is assigned to all the years.

Childhood SES: This variable is a binary variable measuring childhood SES. I constructed it using the IRT procedure as follows. From the HRS data I created four binary variables using the original categorical data on family moved for financial reason, family usually got financial help during childhood, father unemployed during childhood, father's usual occupation during childhood (0 = disadvantaged and 1 = advantaged), and three tertiary variables two on each parent's educational levels (0 = High School dropout, 1 = some college, 2 = completed college and higher) and third on family financial situation (0 = poor, 1 = average, 2 = well-off). I used these seven variables as items in the IRT procedure to first compute a continuous score estimate and then I define Childhood SES = 1 if the score is above mean plus one standard deviation of the scores and 0 otherwise.

Childhood Health is a binary measure of childhood health constructed from the self-reported qualitative childhood health variable in HRS. I define Childhood Health = 1 if the respondent reported very good or excellent, and zero otherwise.

4 Numerical findings

Childhood health status is an important factor for later life health outcomes and educational attainments. Childhood SES influences the stressors of the cells environment and thus will affect Childhood Health. Apart from Childhood SES, other factors such as nutrition and pediatric health care are important factors. We do not have data on those. I estimated a logit model of diseases with right-hand side variables as childhood health, childhood socioeconomic status, college+, and other observable characteristics mentioned above characteristics.

4.1 Predictions from two models

In the next two tables, I report the performance metrics like precision, recall, f1-score confusion matrix of an LSTM model trained on the unbalanced and balanced training data and evaluated on the same test data.

The definition of these performance metrics are: Precision metric measures the percent of positive predictions of a health status is correct, i.e. actually of that health status. More precisely, precision = true positive/(true positive + false positive). Recall measures the proportion of the actual population of a health status is correctly predicted, i.e., recall = true positive / (true positive + false negative). f1-score = 2(Recall · Precision)/(Recall + Precision). A confusion matrix provides a more detailed classification statistics. A diagonal number tells us the fraction of population of health status given in the row label or column label is correctly predicted. This is same as recall metric. The row-wise off-diagonal numbers tell us what percentages of the population of the health status in the row label are wrongly predicted to be of health status in the column labels. The column-wise off-diagonal numbers tell us which fractions of populations of each of the other health statuses are wrongly predicted as the given health status in the column label. In other words, the (i, j)-th element of a confusion matrix gives the number of people having actually disease i are predicted by the model to have disease j, for all $i, j = 1, ...k$. The numbers below in square brackets are percentages. The accuracy is taken as the percentage of observations in the sample are correctly predicted by the model.

Table 2: Confusion matrix on test data using the multinomial Logit model estimates on the training data

Health Status	Normal	Cardiovas	Cancer	Other	Comorbid	Support	Percent	Precision	Recall	f1-score
Normal	0.674	0.134	0.000	0.002	0.191	1,205	31.65	0.388	0.674	0.492
Cardiovas	0.517	0.230	0.000	0.001	0.252	853	22.41	0.359	0.230	0.280
Cancer	0.800	0.022	0.000	0.000	0.178	45	1.18		0.000	
Other	0.576	0.095	0.000	0.010	0.319	675	17.73	0.389	0.010	0.020
Comorbi	0.403	0.121	0.000	0.008	0.468	1,029	27.03	0.419	0.468	0.442
Total	2,093	546	0	18	1,150					
Percent	54.98	14.34	0.00	0.47	30.21				0.39	

Notes: Rows correspond to actual health status and columns to model-predicted health status. The numbers in 'Percent' column denote the percentage distribution of actual health status in the test data, and the numbers in the 'Percent' row denote percentage distrubtion of the predicted race. The last number in the 'Recall' column denotes the accuracy rate for all health statuses together

Confusion matrix on the test data using the parameter estimates from the training data for the multinomial logistic model is shown in Table 2 and for the deep neural network model in Table 3.

The tables show that multinomial logit model has accuracy both models have close to 40 percent accuracy, the deep neural network has slightly better overall performance, but not significantly higher. Neither models predict correctly for the disease cancer. For the

Table 3: Confusion matrix on test data using deep learning model estimates on the training data

Health Status	Normal	Cardiovas	Cancer	Other	Comorbid	Support	Percent	Precision	Recall	f1-score
Normal	0.647	0.118	0.000	0.018	0.217	1,205	31.65	0.406	0.647	0.499
Cardiovas	0.488	0.225	0.000	0.009	0.278	853	22.41	0.365	0.225	0.278
Cancer	0.689	0.067	0.000	0.000	0.244	45	1.18		0.000	
Other	0.498	0.090	0.000	0.046	0.366	675	17.73	0.392	0.046	0.082
Comorbi	0.349	0.124	0.000	0.017	0.509	1,029	27.03	0.409	0.509	0.454
Total	1,922	526	0	79	1,280					
Percent	50.49	13.82	0.00	2.08	33.62				0.40	

Notes: Rows correspond to actual health status and columns to model-predicted health status. The numbers in ’Percent’ column denote the percentage distribution of actual health status in the test data, and the numbers in the ’Percent’ row denote percentage distrubtion of the predicted race. The last number in the ’Recall’ column denotes the accuracy rate for all health statuses together

disease category Other, deep neural network model perform better about 7 percent than the multinomial logit model about 2 percent.

The accuracy of about 40 percent overall performance for a deep neural network model with so many neurons as compared to accuracy rates of more than 95 percent accuracy rates in handwritten character recognition problems may point to the fact that the effects of the left-out genetic factors and detailed measures of epigenetic factors throughout life are important determinants disease risks.

4.2 Influential factors in two models

4.2.1 Multinomial logit regression model

Maximum likelihood estimates and their standard errors for the multinomial logit model are reported in Table 4. Smaller the p-value, the stronger is the statistical significance of the parameter estimate. p-values below 0.001 are marked with ***, greater than or equal to 0.001 but less than 0.01 are marked with **, and greater than or equal to 0.01 but less than 0.05 are marked with *. Starred parameter estimates of input variables for a disease are taken to be the important factors for the disease.

These estimates show that whites have lower risk of cardiovascular diseases and higher risks of other disease category.

Females have lower risk of cardiovascular diseases but higher risks of all other diseases.

Individuals who had good health in childhood have lower risks of diseases in other category and two or more diseases.

High BMI individuals have significantly higher risk of all diseases and have no significant effect on the risk of cancer. They have the risk of cardiovascular diseases about $exp(0.702)$, i.e., 2.02 times higher than the risk of a low BMI individual.

Like high BMI , smoking significantly increases the risks of all diseases, except it does not have significant effect on the risk of cardiovascular diseases.

The stress measured by the variable CES-D has significant effect on a single disease in other category and for comorbidity.

The risk of cancers has not many significant predictors, except that females and the smokers have significantly higher risks.

4.2.2 Deep neural network model

As mentioned earlier, unlike the statistical models that produce parameter estimates and their p-values to get an idea about the importance, numerical effect and qualitative direction of the effect of a regressor on probabilities of outcomes, the machine learning models are unable to do these type of inference. The machine learning literature does not have a sound criteria for variable selection or its effect on outcomes. I adapt the permutation importance technique introduced by Breiman (2001a) for random forests models to the present deep neural network model of multinomial output to identify influential factors. The procedure I follow is described below:

For each disease, we have computed a performance measure, the accuracy rate reported in the diagonal elements in Table Table 3. Denote it by α. Take an input variable. If it has strong effect on the predictive power of the risk, when we reshuffle its values among the individuals, individual A gets B's value and B gets F's value and so on, then the predictive performance will deteriorate at least for many permutations. If the input is not a significant predictor of the risk, the accuracy will not decrease very much. For each input variable, I draw 5,000 random permutations. Let α_i be the accuracy rate in the shuffled input data induced by the i-th permutation. Let m be the mean of these α_i's and sd the standard deviation. If $\alpha > m + 2 * sd$, I define the input strongly influential, denoted with ***. If it is not strongly influential but $\alpha > m + sd$, I define it to be moderately influential, denoted with **. If it is not moderately influential but $\alpha > m$, I define it to be slightly influential, denoted with *. Otherwise the input has no significant effect on the prediction accuracy. These influential levels are shown in Table Table 5. Note that these *'s only mean that they have significant influence on predicting the outcome, but they do not tell us if the input is changed by a unit, how much will be the effect on the risk of the disease. This criterion for determining the influence of a variable may not work well if two input variables are highly correlated. This also applies to the p-value criterion used in the statistical model.

An important variable and the degree of its importance is marked with *'s are almost identical to those in the statistical multinomial logit model above with the exception that two variables—childhood SES and College+—show significant influence on the risks of diseases in other category and in comorbid category.

Table 4: Estimates from a multinomial logistic regression model

	2-Cardiovas	3-Cancer	4-other	5-Comorbid
Intercept	0.234	-4.606 ***	-0.691 ***	0.852 ***
	(0.142)	(0.474)	(0.157)	(0.139)
White	-0.398 ***	0.514 *	0.401 ***	0.022
	(0.054)	(0.201)	(0.066)	(0.056)
Female	-0.114 *	0.929 ***	0.538 ***	0.517 ***
	(0.047)	(0.153)	(0.052)	(0.048)
Childhood SES	-0.102	-0.171	-0.117	-0.193 **
	(0.066)	(0.193)	(0.072)	(0.071)
Childhood Health	0.069	0.177	-0.285 ***	-0.372 ***
	(0.057)	(0.180)	(0.059)	(0.054)
College+	-0.060	0.107	-0.010	-0.104
	(0.058)	(0.165)	(0.063)	(0.062)
HIGH BMI	0.731 ***	-0.069	0.202 ***	0.869 ***
	(0.051)	(0.138)	(0.052)	(0.052)
CES-D	0.350 **	0.334	1.188 ***	1.667 ***
	(0.110)	(0.325)	(0.110)	(0.099)
Cognitive scores	-0.016 **	0.032	0.003	-0.032 ***
	(0.005)	(0.017)	(0.006)	(0.005)
Smoking	0.061	0.232	0.299 ***	0.387 ***
	(0.047)	(0.138)	(0.051)	(0.048)
Exercising	-0.298 ***	-0.274	-0.201 ***	-0.660 ***
	(0.054)	(0.160)	(0.059)	(0.052)
cohort 1948-53	-0.167 **	-0.209	-0.193 **	-0.451 ***
	(0.063)	(0.196)	(0.069)	(0.065)
cohort 1954-59	0.059	-0.180	-0.229 **	-0.420 ***
	(0.061)	(0.201)	(0.072)	(0.066)
AIC	41293.454	41293.454	41293.454	41293.454

*** $p < 0.001$; ** $p < 0.01$; * $p < 0.05$.

Table 5: Influential variables for prediction of mid-age diseases

X	normal	Cardiovas	Cancer	Other	Comorbid
White	*	***		***	
Female	***	*		***	*
Childhood SES	**	*		*	*
Childhood Health	**	*		*	*
College+	*			*	
High BMI	***	***		***	**
CES-D	***	*		*	***
Cognitive scores	**	*		***	*
Smoking	*	*		**	*
Drinking	***	*		*	***
cohort5_48_53	*	*		*	*
cohort6_54_59		**		*	*

Note: *** = strongly, ** = moderately, * = slighly influential based on the criterion of m - 3*s.d., m - 2*s.d. and m is greater than 0 respectively, where m is the mean and sd is the standard deviation calculated with 5000 random permutations for each variable.

5 Disease risk estimates for various social groups.

I consider three groups of individuals with characteristics given below and predict their disease risks using the multinomial logit model and the deep neural network model. The disease risks estimates for various groups can throw light on the effects of policies to close the gaps in the disease risks. Three groups are defined as follows.

Individuals with values of the childhood factors, Childhood SES = 0, Childhood Health = 0, College+ = 0 will be referred as disadvantaged, and as advantaged if these variables take value 1; individuals with values for biomarkers CES-D, Cognitive scores at their mean, and HIGH BMI = 1 (i.e., high HIGH BMI) as average biomarker, and health behaviors—behav_prev = 0, behav_smoke = 1, behav_drink = 1, behav_vigex = 0–as poor health practices, and as good health practices if these variables take the opposite values.

I consider three types of individuals, all having average values of the biomarkers and currently of age 51:

- Type 1: Disadvantaged individuals with poor health practices.
- Type 2: Advantaged individuals with poor health practices.
- Type 3: Advantaged individuals with good health practices.

Table 6: Disease risks estimates for various groups from the multinomial logit and deep neural network models

Group	Normal	Cardiovas	Cancer	Other	Comorbid
Multinomial Logit Model					
type 1	0.238	0.229	0.007	0.176	0.351
type 2	0.317	0.278	0.011	0.155	0.240
type 3	0.455	0.279	0.009	0.135	0.121
Deep Neural Network					
type 1	0.202	0.188	0.008	0.183	0.419
type 2	0.283	0.247	0.010	0.174	0.285
type 3	0.458	0.248	0.012	0.148	0.134

5.1 Determinants of childhood factors

Childhood health status (Childhood Health) is an important factor for later life health outcomes and educational attainments. Childhood SES influences the stressors of the cells environment and thus will affect Childhood Health. Apart from Childhood SES, other factors such as nutrition and pediatric health care are important factors.

Many childhood factors also determine College+ such as innate IQ, family background, preschool inputs, prenatal and postnatal stressors for brain development, the childhood health status, and mother's time input. See, Heckman (2008) and Raut (2018) for recent literature on the biology of brain development and the role of socioeconomic factors, and Heckman and Raut (2016) for a Logit model of college completion in which a IQ measure, family background measured with parents' education, preschool inputs and non-cognitive skills play important roles.

HRS does not have data on many of childhood variables. Based on the available data, Raut (2021) Table 6, reported the effect of childhood socioeconomic status and Childhood Health, together with a few other demographic variables as regressors in Logit models for Childhood Health and College+, which I reproduce in Table Table 7)

From the statistically significant parameter estimates of the variable Childhood SES in the models with Childhood SES as a regressor, we see that Childhood SES has positive effect on child health, and on the probability of college completion and the probability better (i.e., normal as compared to diseased) Init.HLTH.

A better childhood health leads to a higher probability of college completion and a higher probability of being in normal health in one's early 50s. An education level of at least a college degree also has a significant positive effect on the probability of being in normal health in one's early 50s.

Furthermore, the estimates show that White has higher probabilities for better childhood health, attaining college degree and more, and better initial health outcomes in one's early 50s and Female has lower probability of completing college and lower probability of remaining in good health in their early 50s.

It is possible that after controlling for Childhood SES, the White might have better health-related behaviors. I cannot control for health behaviors in the models of this subsection as the HRS data does not have data on health behaviors prior to the survey years. I will

Table 7: Effects of childhood factors, race and sex on childhood health and college education

Variables	Childhood Health	College+	aHLTH
Intercept	0.859 *** (0.031)	-2.081 *** (0.049)	-0.672 *** (0.036)
White	0.150 *** (0.031)	0.286 *** (0.038)	0.070 * (0.029)
Female	-0.099 *** (0.029)	-0.308 *** (0.033)	-0.212 *** (0.026)
Childhood SES	0.666 *** (0.046)	1.630 *** (0.038)	0.140 *** (0.038)
Childhood Health		0.549 *** (0.041)	0.280 *** (0.031)
College+			0.242 *** (0.033)
N	24342	24342	24342
R squared	0.010	0.094	0.009
Loglik	-14011.024	-11689.239	-16134.502

*** p < 0.001; ** p < 0.01; * p < 0.05. Standard errors are in parentheses.
Source: Raut(2021), Table 6: Effects of childhood factors, race and sex on childhood health, college education and initial health in early 50s.

examine this with the second measure of health outcomes, i.e., probabilities of following different health trajectories starting at age 51 in the next two subsections.

6 Discussions

The above is a static model of health outcomes at one point of time in life-cycle, i.e. in early 50's. Health progression over time and the feedback effect of earlier health status and health behaviors on later health outcomes are also important to analyze from the healthcare and public policy perspectives. For analysis of pathways through health states, the statistical literature generally uses a multi-state time to event framework with Cox regression models capturing the effect of time varying covariates. In Raut (2017) and Raut

(2021), I have used such models for prediction of disease risks over time. A multi-state statistical model predicts the probability of disease incidence over many periods. These models assume a Markovian structure and assumes that covariates influence the probabilities of health outcomes through a linear aggregator $X_t'\beta$ at time t in a Cox proportionality fashion, similar to the multinomial logit model in Eq. (1). For estimation a method like maximum likelihood or partial likelihood and to derive sampling theory of the parameter estimates, these statistical models assume that the true data-generating process is one of the member of the parametric or semi-parametric family. Like the question we addressed about the above static model, a similar question arises in this longitudinal case: Could a deep neural network model by relaxing the above restrictions on functional form and the data-generating process perform better than a statistical multi-state model?

A few papers (Faraggi and Simon, 1995; Biganzoli et al., 1998; Katzman et al., 2018; Lee et al., 2018; Ranganath et al., 2016) used feed-forward 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. (2018) who studied competing risks, by breaking death into various causes of death. Katzman et al. (2018) introduced a more general non-linearity of the covariate effects, but still kept the Cox proportionality assumption. Ranganath et al. (2016) 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 are 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 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 involves 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).

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 experiments, 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 Raut (2020), I formulated an LSTM-RNN model of multi-state time-to-event model, implemented in Keras module of Tensorflow 2.0 for Python, and compared its predictive performance with that of a multi-state statistical model. Similar to the accuracy rate as the criterion to evaluate performance of the models in multi-class classification given in previous section, I used the c-index criterion to discriminate the performance of various models in predicting time-to-event probabilities. I found that a LSTM-RNN type neural network model did better job in predicting time-to-event probabilities than a multi-state statistical model.

7 Conclusion

This paper addressed two issues: First, to predict risks of diseases and to detect influential variables for these risks, how does a statistical multinomial logit model compares with a deep neural network model? Second, using the Health and Retirement Surveys data, the paper evaluates these two types of models and examines how the early childhood factors, and health behaviors affect the risks of diseases for adults in their early 50s. The paper considers four categories of chronic diseases: Cardiovascular, Cancer, single other disease, and comorbidity. The input variables used are dummy variables—White, Female, Childhood SES, Childhood Health, College+, High HIGH BMI , Smoking and Exercising—and continuous variables—CES-D and Cognitive scores. The variable CES-D measures stress.

The paper explained that under strong assumptions on the true data-generating process, the statistical models use maximum likelihood or similar optimizing methods to estimate parameters of the model, derive sampling properties of the estimates which provide criteria such as p-values and other hypothesis testing procedures to pick important variables. The estimated parameters in most models give numerical estimates of the effects of variables on the risk of diseases. For a deep neural network model, which uses more general assumptions on the data-generating process, does not have good methods to identify important variables and to estimate the numerical effects of variables on the risks of various diseases. This paper adapts the permutation importance technique, introduced by Breiman (2001b) for random forests models, to the present deep neural network framework involving multinomial outcomes to identify influential variables for the prediction of risks.

For both models, the paper used a common test dataset to fit models and computed the predictive accuracy on a test dataset to compare the predictive performance of the two models. The paper finds that the statistical multinomial logit model has an accuracy rate of 38.9 and the deep neural network model has a slightly higher accuracy rate of 39.4.

Both models pick the same variables as influential in prediction of risks, with the exception that the childhood SES and College+ variables were not significant, that is not influential, in the statistical multinomial logit model, but they are influential in the deep neural network model for the prediction of risks of the single disease in the other category and for the risks of comorbidity. For the quantitative effects on risks of diseases, as mentioned above this can be computed for the statistical multinomial logit model but not for the deep neural network models, the paper finds that health status in childhood, HIGH BMI and smoking habits have strong effects on the risks diseases. The paper found for instance that a high HIGH BMI individual has risk of cardiovascular disease 1.95 times higher than the risk of a low HIGH BMI individual and an individual who smokes has the risk of cancer 1.57 times higher than the risk of an individual who does not smoke.

While the performance of the deep neural network model in this paper did not show significant improvement in performance over the statistical multinomial logit model, Raut (2020) formulated a deep LSTM type recurrent neural network (RNN) model of probability of diseases over time and compared its performance with the analogous statistical multi-state model using the Health and Retirement dataset. The paper found that a deep

LSTM type RNN model attained substantially higher performance over the analogous statistical multi-state model. It is possible that other type of deep neural network architecture such as convolutional neural network (CNN) architecture may provide significantly better performance compared to the feed-forward deep neural network considered in this paper.

8 Appendix

Analytical Details of MLP learning algorithm

Keeping the example depicted in Figure 1 in the background, I will describe the algorithm for the general case. For a function $\sigma(z)$, the first derivative of the function is denoted as $\sigma'(z)$ and for a vector function the notation will denote the vector of the derivatives. The sets of neurons at layer $\ell = 0, 1, 2$ are $N^0 = \{1, 2, 3\}$ corresponding the three inputs, $N^1 = \{1, 2, 3, 4\}$, i.e. 4 hidden units, and $N^L = \{1, 2\}$ at the output layer corresponding to 2 outputs. Let $n^k = \#N^k$, i.e., the number of neurons in the k-th layer of the network, for $k = 0, 1, ..., L$.

1. First compute all neuron outputs a_i^ℓ, $i \in N^\ell$ using the feed-forward iteration: $a^\ell = \sigma^\ell(W^\ell a^{l-1} + b^\ell)$, $\ell = 1, ..., L$, with initial condition $a^0 \equiv x$ (input vector).

2. To calculate $\frac{\partial \mathcal{L}}{\partial w_{ij}^\ell}$, and $\frac{\partial \mathcal{L}}{\partial b_i^j}$, denote by $z^\ell \equiv W^\ell a^{l-1} + b^\ell$ and $\delta_i^\ell \equiv \frac{\partial \mathcal{L}}{\partial z_i^\ell} \cdot \sigma^{\ell\prime}(z_i^\ell)$, and calculate first all δ_i^ℓ, $i \in N^\ell, \ell = 1, 2, ..., L$, using the backpropagation iteration: $\delta^\ell = W^{\ell+1}\delta^{l+1} \odot \sigma^{\ell\prime}(z^\ell)$, $\ell = L{-}1, L{-}2, ..., 1$, with the initial condition $\delta^L = \frac{\partial \mathcal{L}}{\partial a^L} \odot \sigma^{L\prime}(z^L)$, where $\odot$ denotes the Haddamard product, i.e., element wise product.

3. Now compute $\frac{\partial \mathcal{L}}{\partial b_i^\ell} = \delta_i^\ell$ and $\frac{\partial \mathcal{L}}{\partial w_{ij}^\ell} = \delta_i^\ell a_j^{\ell-1}$, $i = 1, 2, ..., n^\ell, j = 1, 2, ..., , n^{\ell-1}, \ell = 1, 2, ..., L$. In matrix notation, $\frac{\partial \mathcal{L}}{\partial b^\ell} = \delta^\ell$ and $\frac{\partial \mathcal{L}}{\partial W^\ell} = \delta^\ell a^{\ell-1\prime}, \ell = 1, 2, ..., L$.

References

[1] Ahmad, M. A. et al. “Interpretable Machine Learning in Healthcare”, IEEE Intelligent Informatics Bulletin. Vol. 19. 1. 2018. DOI: 10.1145/3233547.3233667 (cit. on p. 3).

[2] Altae-Tran, H. RNNs for Model Predictive Control in Unknown Dynamical Systems with Low Sampling Rates, working paper, Stanford University (2016) (cit. on p. 29).

[3] Altae-Tran, H. et al. Low Data Drug Discovery with One-Shot Learning, ACS Central Science, 3, no. 4 (2017), 283–293. DOI: 10.1021/acscentsci.6b00367 (cit. on p. 29).

[4] An, X. and Yung, Y.-F. Item response theory: what it is and how you can use the IRT procedure to apply it, SAS Institute Inc. SAS364-2014 (2014) (cit. on p. 16).

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

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

[7] Biganzoli, E. et al. Feed forward neural networks for the analysis of censored survival data: a partial logistic regression approach, Statistics in Medicine, 17, no. 10 (1998), 1169–1186. DOI: 10.1002/(SICI)1097-0258(19980530)17:10<1169::AID-SIM796>3.0.CO;2-D (cit. on p. 28).

[8] Breiman, L. Random forests, Machine Learning, 45, no. 1 (2001), 5–32. DOI: 10.1023/a:1010933404324 (cit. on pp. 12, 21).

[9] Breiman, L. Statistical Modeling: The Two Cultures (with comments and a rejoinder by the author), Statistical Science, 16, no. 3 (Aug. 2001), 199â€“231. DOI: 10.1214/ss/1009213726 (cit. on pp. 12, 30).

[10] Bugliari, D. et al. RAND HRS Data Documentation, Version P. Tech. rep. 2016 (cit. on p. 14).

[11] Cai, L. et al. Item Response Theory, Annual Review of Statistics and Its Application, 3, no. 1 (June 2016), 297–321. DOI: 10.1146/annurev-statistics-041715-033702 (cit. on p. 16).

[12] Chen, M. et al. Disease Prediction by Machine Learning Over Big Data From Healthcare Communities, IEEE Access, 5 (2017), 8869–8879. DOI: 10.1109/access.2017.2694446 (cit. on p. 4).

[13] Chollet, F. et al. Keras. 2015 (cit. on p. 12).

[14] Cybenko, G. Approximation by superpositions of a sigmoidal function, Mathematics of Control, Signals, and Systems, 2, no. 4 (Dec. 1989), 303â€“314. DOI: 10.1007/bf02551274 (cit. on p. 4).

[15] Efron, B. and Hastie, T. Computer Age Statistical Inference. Institute of Mathematical Statistics Monographs. Cambridge University Press, 2016 (cit. on p. 12).

[16] Embretson, S. and Reise, S. Item Response Theory. Item Response Theory for Psychologists. Taylor & Francis, 2000 (cit. on p. 16).

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

[18] Fisher, G. G. and Ryan, L. H. Overview of the Health and Retirement Study and Introduction to the Special Issue, Work, Aging and Retirement, 4, no. 1 (Dec. 2017). Ed. by M. Wang, 1–9. DOI: 10.1093/workar/wax032 (cit. on pp. 13, 14).

[19] Gluckman, P. D. et al. Effect of In Utero and Early-Life Conditions on Adult Health and Disease, New England Journal of Medicine, 359, no. 1 (July 2008), 61–73. DOI: 10.1056/nejmra0708473 (cit. on p. 4).

[20] Goodfellow, I. et al. Deep Learning. MIT Press, 2016 (cit. on p. 9).

[21] Graves, A. Supervised Sequence Labelling with Recurrent Neural Networks. Springer Berlin Heidelberg, 2012. DOI: 10.1007/978-3-642-24797-2 (cit. on pp. 9, 29).

[22] Graves, A. et al. Neural Turing Machines, CoRR, abs/1410.5401 (2014) (cit. on p. 28).

[23] Hair, J. F. et al. A Primer on Partial Least Squares Structural Equation Modeling (PLS-SEM) [2nd ed.] Sage, 2017 (cit. on p. 5).

[24] Hammer, B. On the approximation capability of recurrent neural networks, Neurocomputing, 31, no. 1 (2000), 107–123. DOI: https://doi.org/10.1016/S0925-2312(99)00174-5 (cit. on p. 28).

[25] Hastie, T. and Tibshirani, R. Generalized Additive Models. GLM. Aug. 2006. DOI: 10.1002/0471667196.ess0297.pub2 (cit. on p. 12).

[26] Hastie, T., Tibshirani, R., and Friedman, J. The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Second Edition (Springer Series in Statistics). Springer, 2009 (cit. on p. 12).

[27] Heckman, J. J. Schools, Skills and Synapses, Economic Inquiry, 46, no. 3 (July 2008), 289–324. DOI: 10.1111/j.1465-7295.2008.00163.x (cit. on p. 26).

[28] Heckman, J. J. and Raut, L. K. Intergenerational long-term effects of preschool-structural estimates from a discrete dynamic programming model, Journal of Econometrics, 191, no. 1 (2016), 164–175. DOI: 10.1016/j.jeconom.2015.10.001 (cit. on pp. 15, 26).

[29] Hochreiter, S. and Schmidhuber, J. Long Short-Term Memory, Neural Computation, 9, no. 8 (1997), 1735–1780. DOI: 10.1162/neco.1997.9.8.1735 (cit. on p. 29).

[30] Hornik, K. et al. Multilayer feedforward networks are universal approximators, Neural Networks, 2, no. 5 (Jan. 1989), 359â€“366. DOI: 10.1016/0893-6080(89)90020-8 (cit. on pp. 4, 10).

[31] Juster, F. T. and Suzman, R. An Overview of the Health and Retirement Study, The Journal of Human Resources, 30 (1995), S7. DOI: 10.2307/146277 (cit. on p. 13).

[32] Katzman, J. L. et al. DeepSurv: personalized treatment recommender system using a Cox proportional hazards deep neural network, BMC Medical Research Methodology, 18, no. 1 (Feb. 2018), 24. DOI: 10.1186/s12874-018-0482-1 (cit. on p. 28).

[33] Kourou, K. et al. Machine learning applications in cancer prognosis and prediction, Computational and Structural Biotechnology Journal, 13 (2015), 8–17. DOI: https://doi.org/10.1016/j.csbj.2014.11.005 (cit. on p. 4).

[34] Lee, C. et al. "Deephit: A deep learning approach to survival analysis with competing risks", Thirty-Second AAAI Conference on Artificial Intelligence. 2018 (cit. on p. 28).

[35] Luo, Y. and Waite, L. J. The Impact of Childhood and Adult SES on Physical, Mental, and Cognitive Well-Being in Later Life, The Journals of Gerontology: Series B, 60, no. 2 (Mar. 2005), S93–S101. DOI: 10.1093/geronb/60.2.s93 (cit. on p. 15).

[36] Pearl, J. Causal inference in statistics: An overview, Statistics Surveys, 3 (2009), 96–146. DOI: 10.1214/09-SS057 (cit. on p. 5).

[37] Ranganath, R. et al. "Deep Survival Analysis", Proceedings of the 1st Machine Learning for Healthcare Conference. Vol. 56. PMLR. 2016, 101–114 (cit. on p. 28).

[38] Raut, L. K. Exits from Disability: Estimates from a Competing Risk Model, Social Security Bulletin, 77, no. 3 (2017), 15–38 (cit. on p. 27).

[39] Raut, L. K. Health Outcomes in Mid-Ages: Multistate time to event Statistical Models versus Long Short Term Memory (LSTM) Recurrent Neural Network (RNN) Models, 2020 ASSA Meetings, San Diego, January 3 - 5, 2020 (2020) (cit. on pp. 29, 30).

[40] Raut, L. K. Long-term Effects of Preschool on School Performance, Earnings and Social Mobility, Studies in Microeconomics, 6, no. 1-2 (June 2018), 24–49. DOI: 10.1177/2321022218802023 (cit. on p. 26).

[41] Raut, L. K. Pathways to Disability and Death Before Disability in middle ages: Estimates from the Health and Retirement Study Data, Working Paper (2021) (cit. on pp. 26, 27).

[42] Ren, K. et al. "Deep recurrent survival analysis", Proc. AAAI. 2019, 1–8 (cit. on p. 28).

[43] Shmueli, G. To Explain or to Predict?, Statistical Science, 25, no. 3 (Aug. 2010). DOI: 10.1214/10-sts330 (cit. on p. 5).

[44] Siegelmann, H. T. and Sontag, E. D. "On the computational power of neural nets", Proceedings of the fifth annual workshop on Computational learning theory - COLT '92. ACM Press, 1992. DOI: 10.1145/130385.130432 (cit. on p. 28).

[45] Simons, R. L. et al. Economic hardship and biological weathering: The epigenetics of aging in a U.S. sample of black women, Social Science & Medicine, 150 (2016), 192–200. DOI: 10.1016/j.socscimed.2015.12.001 (cit. on p. 4).

[46] Sonnega, A. et al. Cohort Profile: the Health and Retirement Study (HRS), International Journal of Epidemiology, 43, no. 2 (Mar. 2014), 576–585. DOI: 10.1093/ije/dyu067 (cit. on p. 13).

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

[48] TensorFlow Developers. TensorFlow. 2021. DOI: 10.5281/ZENODO.4724125 (cit. on p. 12).

[49] Vable, A. M. et al. Validation of a theoretically motivated approach to measuring childhood socioeconomic circumstances in the Health and Retirement Study, PLOS ONE, 12, no. 10 (Oct. 2017). Ed. by A. Fraser, e0185898. DOI: 10.1371/journal.pone.0185898 (cit. on p. 15).

[50] Venables, W. and Ripley, B. Modern Applied Statistics with S. Fourth. New York: Springer, 2002 (cit. on p. 12).