Journal of Healthcare Informatics Research (2021) 5:181–200
https://doi.org/10.1007/s41666-021-00093-9

RESEARCH ARTICLE

Understanding Demographic Risk Factors for Adverse Outcomes in COVID-19 Patients: Explanation of a Deep Learning Model

Yijun Shao[1,2] · Ali Ahmed[1,2,3] · Angelike P. Liappis[1,2] · Charles Faselis[1,2] · Stuart J. Nelson[2] · Qing Zeng-Treitler[1,2]

Received: 20 July 2020 / Revised: 6 February 2021 / Accepted: 10 February 2021 /
Published online: 27 February 2021

Abstract

This study was to understand the impacts of three key demographic variables, age, gender, and race, on the adverse outcome of all-cause hospitalization or all-cause mortality in patients with COVID-19, using a deep neural network (DNN) analysis. We created a cohort of Veterans who were tested positive for COVID-19, extracted data on age, gender, and race, and clinical characteristics from their electronic health records, and trained a DNN model for predicting the adverse outcome. Then, we analyzed the association of the demographic variables with the risks of the adverse outcome using the impact scores and interaction scores for explaining DNN models. The results showed that, on average, older age and African American race were associated with higher risks while female gender was associated with lower risks. However, individual-level impact scores of age showed that age was a more impactful risk factor in younger patients and in older patients with fewer comorbidities. The individual-level impact scores of gender and race variables had a wide span covering both positive and negative values. The interaction scores between the demographic variables showed that the interaction effects were minimal compared to the impact scores associated with them. In conclusion, the DNN model is able to capture the non-linear relationship between the risk factors and the adverse outcome, and the impact scores and interaction scores can help explain the complicated non-linear effects between the demographic variables and the risk of the outcome.

Keywords Coronavirus disease · Artificial intelligence · Deep neural network · Explainable AI

✉ Yijun Shao
yshao@email.gwu.edu

Extended author information available on the last page of the article

1 Introduction

The coronavirus disease 2019 (COVID-19) caused by the novel severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) has infected about 8 million people globally causing over 400,000 deaths by June 2020. In the USA, there had been over 2 million cases and 100,000 deaths [1–3]. As of June 12, 2020, 16,765 Veterans in the US Veteran Affairs (VA) system have been diagnosed with COVID-19, of whom 1422 (8.5%) have died and over 20% hospitalized [4]. However, the risk factors for hospitalization and death in patients with COVID-19 are still being studied [5].

To date, there is no effective treatment to improve outcomes in patients with COVID-19. Although most patients with COVID-19 are asymptomatic, those who develop symptoms and are sicker contribute to most of the hospitalizations and deaths [5]. Thus, identifying risk factors for poor outcomes may help clinicians focus preventive efforts on high-risk subgroups and/or address modifiable risk factors. Some recent studies have examined the risk factors of adverse outcomes in COVID-19 patients, mostly using traditional statistical analysis [6–8].

Early descriptive data from China suggest that most of the deaths in patients with COVID-19 occurred among adults aged ≥60 years and among persons with serious underlying health conditions [9]. Early preliminary descriptive studies of outcomes in patients with COVID-19 in the USA suggested that case fatality was the highest in those aged 85 years and older, ranging from 10 to 27%, while for those aged 65–84 years it was 3 to 11%, 1 to 3% for persons aged 55–64 years, <1% among persons aged 20–54 years, and no fatalities among persons aged ≤19 years [10].

In one study from China, of 1590 patients hospitalized with COVID-19 in China, 50 patients died [11]. Significant predictors of mortality included age 75 years or older (hazard ratio (HR), 7.86; 95% confidence interval (CI), 2.44–25.35), age 65 to 74 years (HR, 3.43; 95% CI, 1.24–9.5), coronary heart disease (HR, 4.28; 95% CI, 1.14–16.13), cerebrovascular disease (HR, 3.1; 95% CI, 1.07 to 8.94), dyspnea (HR, 3.96; 95% CI, 1.42–11), procalcitonin level greater than 0.5 ng/mL (HR, 8.72; 95% CI, 3.42–22.28), and aspartate aminotransferase level greater than 40 U/L (HR, 2.2; 95% CI, 1.1–6.73).

In another study, also from China, 372 hospitalized patients with non-severe COVID-19 were followed for >15 days after admission [12]. Of these, 72 (19%) patients developed severe COVID-19. The authors trained a risk prediction model in a cohort of 189 patients and validated their findings in 2 independent cohorts of 165 and 18 patients. Predictors for transition to severe or critical COVID-19 were older age, higher serum lactate dehydrogenase, C-reactive protein, coefficient of variation of red blood cell distribution width, blood urea nitrogen, direct bilirubin, and lower albumin.

Most of these studies used traditional statistical analytic approaches such as logistic or Cox regression modeling. Such analyses yield odds or hazard ratios as a measure of the effect or association of a specific risk factor with outcomes. While these are considered to be robust and standard measurements, they represent estimates of risk for an entire group and do not account for individual differences and non-linear effects. For example, we cannot assume the relationship between age and COVID-19 outcomes to be linear. It may also be too simplistic to assume that for each individual of the same age, their chronological age poses the same risk.

The resurgence of artificial intelligence with deep learning as a key technology has led to many breakthroughs [13–18]. Some of deep learning's advantages are its ability to model complex relationships, accommodate a large number of variables, and take advantage of a large amount of data. The models resulted from deep learning thus offer an alternative to the traditional statistical models as we seek to understand the underlying relations between demographic risk factors and adverse outcomes in COVID-19 patients.

An important step in this study is the explanation of deep learning models. The interpretation or explanation of deep learning models is an active area of research [19–23]. We have developed and validated two measures which we refer to as impact score and interaction score [24–26]. They each have two versions: one at the individual level and one at the population level. In this study, we will restate and use them, and also introduce two new concepts—impact and interaction. Unlike the two score measures, impact and interaction are only defined at the individual level. They are introduced here because for individuals they provide more meaningful information than the score measures. In addition, we provide a visualization of the individual impacts and impact scores as part of the explanation, which also demonstrates the relationship and difference between the new and the old measures. We also include a comparison with logistic regression results and descriptive statistics. These measures are applied to a DNN model of multi-layer perceptron (MLP) type in this study. However, the measures are model-agnostic; hence, their use is not limited to DNNs of MLP type.

Identifying risk factors for poor outcomes may help clinicians stratify patients with COVID-19 by risk and develop and test interventions that may target modifiable risk factors, thus lowering the risk of deterioration resulting in hospitalization and/or death. It may also inform policymaking, e.g., tailored social distancing, frequency of testing, and priority of vaccination. The objective of the current study is to understand the impacts of the three key demographic variables of age, gender, and race on the adverse outcome of all-cause hospitalization or all-cause mortality in patients with COVID-19 using a deep neural network (DNN) analysis.

2 Methods

2.1 Explaining DNN Models with Impact and Interaction Measures

In this subsection, we will introduce four measures to explain DNN models: impact, impact score, interaction, and interaction score. The two score measures, i.e., impact score and interaction score, are each defined at two levels: the individual level and the population level. Although they have already been developed in [24–26], we will restate their definitions here for the sake of completeness. We will also illustrate mathematically that, when applied to simple logistic regression models, they reduce to the familiar log odds ratios and interaction coefficients. Unlike the score measures, impact and interaction are only defined at the individual level. For mathematical coherence, the content of this section will not be arranged in the order of old concepts first and new concepts next but rather in mixed order.

Consider a DNN with one input layer of n nodes, several hidden layers with a various number of nodes each, and one output layer with a single node outputting the risk scores. The activation function for the output layer is usually chosen as the sigmoid function $\sigma(x) = e^x/(1 + e^x)$ so that the output is always a value between 0 and 1. This type of DNN is commonly used for predicting dichotomized outcomes.

Let $p = F(x_1, \ldots, x_n)$ denote the final trained DNN model, where $x_1, \ldots, x_n$ are the n variables corresponding to the n nodes of the input layer, and p $(0 < p < 1)$ is the output of the model representing the risk of adverse outcome. Let y denote the outcome variable, then we expect to have $p = \text{Prob}(y = 1)$. This condition can be verified through the calibration curve [27]. If this condition is not satisfied, the predicted risk scores should be calibrated using methods such as Platt Scaling or Isotonic Regression [27]. The calibration process will be considered as part of the prediction of the DNN model; hence, we can still assume the condition $p = \text{Prob}(y = 1)$ is satisfied.

To define impact scores and interaction scores, we first define

$$f(x_1, \ldots, x_n) := \text{logit}(F(x_1, \ldots, x_n)) = \log F(x_1, \ldots, x_n) 1 - F(x_1, \ldots, x_n),$$

where logit is the inverse of the sigmoid function σ: $\text{logit}(p) = \log \frac{p}{1-p}$. The output of the logit function is also known as the log odds, which ranges from $-\infty$ to ∞. We use the logit function because it makes the impact scores calculated on a linear logistic regression model (a simple neural network with no hidden layers) be the same as the common log odds ratios, which will be demonstrated later in this section.

For each variable x_i $(i = 1, \ldots, n)$, we choose and fix a reference value x_i^r. The role of the reference values is to serve as the "background" situations for comparison purpose. Therefore, the general principle is to choose the most "common" (e.g., median, mean, and mode) value of the variable. For example, for a binary variable with values 1/0 representing the presence/absence of a diagnosis, usually 0 (absence) is the reference value because absence of a diagnosis is usually the most common situation. However, we do not impose any strict rules, so one is free to choose other values as the reference value based on the specific problem or study. For example, in a study of dementia patients with ages >=65 years, we can choose "65" as the reference value for age.

For an individual subject, we denote by x_i^c the value of x_i on this subject and call it the current value of x_i. We define the impact of the current value x_i^c relative to the reference value x_i^r as

$$\text{impact} = f\left(\cdots, x_i^c, \cdots\right) - f\left(\cdots, x_i^r, \cdots\right).$$

where "$\cdots$" represents the current values for all variables other than x_i. If $x_i^c \neq x_i^r$, we further define the impact score of x_i on this subject as

$$\text{impact score} = \frac{f\left(\cdots, x_i^c, \cdots\right) - f\left(\cdots, x_i^r, \cdots\right)}{x_i^c - x_i^r},$$

This defines individual-level impact score. The populational-level impact score of x_i is defined as the mean of all the individual-level impact scores of x_i.

The impact measures the change in risk (in terms of log odds) as x_i changes from the reference value x_i^r to the current value x_i^c while keeping all the other variables unchanged. The impact score measures the rate of change in risk relative to the change in x_i, or in other words, the change in risk for one unit change in x_i. If x_i is a binary variable taking values 0 and 1 with 0 being the reference value, the impact and impact score are equal numerically on any subject with $x_i^c = 1$. Moreover, the impact is a unitless quantity, while the impact score has units if x_i has units. If x_i is a continuous variable such as age and body mass index, the impact depends on both the current value and the reference value, while the dependence of the impact score on those values is diminished greatly (but not totally removed if the model is nonlinear in x_i).

As a simple illustration, we calculate the impact and impact score on a linear logistic regression model in n variables. The model can be written as

$$p = \sigma(\beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n),$$

where the coefficients β_i of x_i ($i = 1, \ldots, n$) are also known as the log odds ratios. Applying the logit function to both sides, we obtain

$$\mathrm{logit}(p) = \beta_0 + \beta_1 x_1 + \cdots + \beta_n x_n.$$

Then the impact of x_i^c on an individual subject is

$$\begin{aligned} f\left(\cdots, x_i^c, \cdots\right) - f\left(\cdots, x_i^r, \cdots\right) &= \left(\beta_0 + \beta_1 x_1^c + \cdots + \beta_i x_i^c + \cdots + \beta_n x_n^c\right) \\ &\quad - \left(\beta_0 + \beta_1 x_1^c + \cdots + \beta_i x_i^r + \cdots + \beta_n x_n^c\right) = \beta_i x_i^c - \beta_i x_i^r = \beta_i\left(x_i^c - x_i^r\right). \end{aligned}$$

And the impact score of x_i on this subject is

$$\frac{f\left(\cdots, x_i^c, \cdots\right) - f\left(\cdots, x_i^r, \cdots\right)}{x_i^c - x_i^r} = \frac{\beta_i\left(x_i^c - x_i^r\right)}{x_i^c - x_i^r} = \beta_i,$$

which is exactly the log odds ratio of x_i. Therefore, impact scores can be viewed as a generalization of the log odds ratios to DNN models. This also shows that the impact scores of x_i on all subjects are the same and equal to the coefficient of the variable x_i in the logistic regression model, which is known as the log odds ratio. In addition, the use of logit function in the calculation of impacts and impact scores is justified.

Next, for each individual subject and each pair of variables x_i and x_j, we define the interaction of the two current values x_i^c and x_j^c as

$$\text{interaction} = f\left(\cdots, x_i^c, \cdots, x_j^c, \cdots\right) - f\left(\cdots, x_i^c, \cdots, x_j^r, \cdots\right) - f\left(\cdots, x_i^r, \cdots, x_j^c, \cdots\right) + f\left(\cdots, x_i^r, \cdots, x_j^r, \cdots\right).$$

If $x_i^c \neq x_i^r$ and $x_j^c \neq x_j^r$, we further define the interaction score of x_i and x_j on this subject as

$$\text{interaction score} = \frac{f\left(\cdots, x_i^c, \cdots, x_j^c, \cdots\right) - f\left(\cdots, x_i^c, \cdots, x_j^r, \cdots\right) - f\left(\cdots, x_i^r, \cdots, x_j^c, \cdots\right) + f\left(\cdots, x_i^r, \cdots, x_j^r, \cdots\right)}{\left(x_i^c - x_i^r\right)\left(x_j^c - x_j^r\right)}.$$

This defines individual-level interaction score. The populational-level interaction score is defined as the mean of all the individual-level interaction scores.

If we rewrite the interaction as

$$\begin{aligned}\text{interaction} = &\left(f\left(\cdots,x_i^c,\cdots,x_j^c,\cdots\right)-f\left(\cdots,x_i^r,\cdots,x_j^r,\cdots\right)\right)\\ &-\left[\left(f\left(\cdots,x_i^r,\cdots,x_j^c,\cdots\right)-f\left(\cdots,x_i^r,\cdots,x_j^r,\cdots\right)\right)+\left(f\left(\cdots,x_i^c,\cdots,x_j^r,\cdots\right)-f\left(\cdots,x_i^r,\cdots,x_j^r,\cdots\right)\right)\right],\end{aligned}$$

then the interaction can be regarded as the difference between the “double” impact

$$f\left(\cdots,x_i^c,\cdots,x_j^c,\cdots\right)-f\left(\cdots,x_i^r,\cdots,x_j^r,\cdots\right)$$

and the sum of the two “single” impacts:

$$f\left(\cdots,x_i^c,\cdots,x_j^r,\cdots\right)-f\left(\cdots,x_i^r,\cdots,x_j^r,\cdots\right) \text{ and } f\left(\cdots,x_i^r,\cdots,x_j^c,\cdots\right)-f\left(\cdots,x_i^r,\cdots,x_j^r,\cdots\right).$$

Therefore, the interaction captures the impact of two variables that cannot be explained by the simple sum of impacts of each one. If the “double” impact is exactly the same as the sum, then both the interaction and the interaction score are zero, which means there is no interaction between the two variables.

Similar to the impact, the interaction is dependent on the current values x_i^c and x_j^c and the reference values x_i^r and x_j^r, and similar to the impact score, the dependence of the interaction score on those values is diminished greatly.

Again, as a simple illustration, we calculate the interaction and interaction score on a non-linear logistic regression model in 2 variables x_1 and x_2:

$$p = \sigma(a + bx_1 + cx_2 + dx_1x_2),$$

where dx_1x_2 is the interaction term and d is the interaction coefficient. Then, we find the interaction score between x_1 and x_2 to be

$$\begin{aligned}&\frac{\left(a+bx_1^c+cx_2^c+dx_1^cx_2^c\right)-\left(a+bx_1^r+cx_2^c+dx_1^rx_2^c\right)-\left(a+bx_1^c+cx_2^r+dx_1^cx_2^r\right)+\left(a+bx_1^r+cx_2^r+dx_1^rx_2^r\right)}{\left(x_1^c-x_1^r\right)\left(x_2^c-x_2^r\right)}\\ &=\frac{dx_1^cx_2^c-dx_1^rx_2^c-dx_1^cx_2^r+dx_1^rx_2^r}{\left(x_1^c-x_1^r\right)\left(x_2^c-x_2^r\right)}=\frac{d\left(x_1^c-x_1^r\right)\left(x_2^c-x_2^r\right)}{\left(x_1^c-x_1^r\right)\left(x_2^c-x_2^r\right)}=d.\end{aligned}$$

This also shows that for the logistic regression model, all the individual-level interaction scores are the same and equal to the population-level interaction score.

2.2 Study of the COVID-19 Patients

Data Source The data source was the VA’s Corporate Data Warehouse (CDW) administered by VINCI.

Cohort The cohort was defined as the Veterans who met both of the following criteria:

1) Tested positive for COVID-19 on or before May 1, 2020;

2) Not an inpatient at the time of the first positive test, or an inpatient at the time but was admitted no more than 24 h earlier than the first positive test.

To create the desired cohort, we first identified in the CDW all the patients who were tested positive for COVID-19 on or before May 1, 2020. Then we excluded all the non-Veteran patients (e.g., employees) from them because they could not use regular healthcare services in VA (but they could still be tested for COVID-19 in VA). Next, we excluded those Veterans who were already hospitalized for more than 24 h at the time of the positive test because those hospitalizations were very likely long-term stays such as stays in nursing home or psychiatric facilities, which were not caused by COVID-19.

Index Dates For each patient, the index date was defined as the date of the first positive test, if the patient was not hospitalized then. If the patient was already hospitalized at the time of the first positive test, then the index date was defined as the admission date of the hospitalization.

Adverse Outcome, Cases and Controls We defined the adverse outcome to be either all-cause hospitalization or all-cause mortality which occurred between the index date and May 15, 2020 (the date when the cohort was created). For the patients not hospitalized at the time of the first positive test, an adverse outcome must occur after the index date. For those who were hospitalized at the time of the first positive test, that hospitalization itself was the adverse outcome. The patients having the adverse outcome were called cases, and the remaining were called controls.

Predictors and Covariates The predictors were the demographic characteristics including age, gender, and race. Covariates were clinical characteristics, divided into two groups. The first group was the diagnosis data, which included all the ICD-10 codes occurring within 1 year before the index date for each patient. The second group was the medication data, which included all medications used within 2 weeks before the index date for each patient.

Variables and Values Predicting variables were defined based on the predictors and covariates. Age measured in years was a continuous variable and was named age. This variable was normalized to have zero mean and unit standard deviation before it was supplied to the input layer of the DNN model. However, in the calculation of impact score and interaction score, the variable was transformed back to the original scale so that it was still measured in years. For gender, the coding was male=0 and female=1. This variable was named gender_female(vs. male). We categorized races into 4 categories: White, African American (AA), Other, and Unknown, where "Other" included all races other than White or AA: Asian, American Indian, Alaska Native, Native Hawaiian, and other Pacific Islander. These 4 categories were coded as binary vectors of 3 dimensions: White = (0,0,0), AA = (1,0,0), Other = (0,1,0), and Unknown = (0,0,1). This was equivalent to defining 3 binary variables: race_aa(vs. white), race_other(vs. white), and race_unknown (vs. white), which corresponded to the 3 dimensions of the vectors, respectively.

For the diagnosis data and medication data, a variable was defined for each ICD-10 code and for each medication with a prevalence of ≥1% in the cohort. Each of the variables took binary values 1/0 representing the presence/absence of the corresponding code or medication.

The outcome variable was a binary variable as well, which took value 1/0 representing the presence/absence of the adverse outcome.

To calculate the impact/interaction scores, we chose the reference values for these variables as follows: for age, the only continuous variable, we chose the median as the reference value, and for gender and race variables, which were binary variables, we chose 0 as the reference value. This made them consistent with what their names suggested: male was the reference gender, and White was the reference race.

DNN Model We designed a DNN as follows: it had one input layer with the number of nodes equal to the number of predicting variables; 4 hidden layers with 50, 20, 20, and 10 nodes, respectively; and one output layer with only one node. The hidden layers and the output layer were all fully connected to their previous layer with a non-linear activation function. The 4 hidden layers used the rectified linear unit (ReLU) function [18] as the activation function. The output layer used the sigmoid function σ as the activation function so that the output value was always between 0 and 1. The output values were called risk scores, as higher values corresponded to higher risks of the adverse outcome. For model training, we chose the binary-cross-entropy function as the loss function, which measures the error between the prediction and the actual result. This function was to be minimized in the training process. Following the convention, the area under ROC curve (AUC) was used as the main metric for measuring the prediction performance. The additional metrics included accuracy, sensitivity, and specificity, which all depended on a choice of threshold on the predicted risk scores. We chose the threshold that maximized the accuracy.

Training, Validation and Testing We randomly partitioned the cohort into 3 subsets: (1) training (80%), (2) validation (10%), and (3) testing (10%). Before the training, the weights of the DNN model were initialized as small random numbers. Then, the weights were iteratively updated using the mini-batched Nesterov [28] gradient descent method to decrease the value of the loss function. Each mini-batch consisted of 50 patients.

One pass over the whole training set was called an epoch. After each epoch, an AUC was computed on the validation set based on the model prediction. To prevent overfitting to the training set, we took an early stopping strategy: the training should stop at the end of the number of epochs such that the validation AUC was higher than all the validation AUCs of the previous epochs and also higher than the validation AUCs of the next 10 epochs. The model obtained at the early-stop point was the final model. Then, the AUC computed on the testing set was reported as the performance of the final model prediction.

Logistic Regression Model We fitted a logistic regression (LR) model as the baseline. The LR model was fitted to the training set, and the AUC computed on the testing set was reported as the performance measure.

3 Results and Analyses

We developed a final cohort of 5407 patients, of whom 2355 (43.6%) had an adverse outcome (including 566 (10.5%) who had a mortality outcome) and 3052 (56.4%) had a favorable outcome. The basic demographic characteristics are summarized in Table 1, and some prevalent conditions are summarized in Table 2.

The overall trend was that the proportion of cases grows with age (Fig. 1). However, as age passes 80 years, the proportion starts plateauing and then dropping for ages >90 years. The dropping after 90 is of particular interest to us because a higher age is generally considered as a key harmful factor. This will be investigated further in a later part of this section.

The calibration curve (Fig. 2) of the final DNN model shows the predicted risk scores well represented the probabilities of having the adverse outcomes, as indicated by its closeness to the diagonal line. Hence, the risk scores can be interpreted as the probabilities of having the adverse outcome without any additional calibration as described in Section 2.1.

The performance on the test set showed that the DNN model was better than the LR model (Table 3), which was expected as the DNN model can model non-linear relationships between the predictors and outcome. The ROC curve of the DNN model is shown in Fig. 3.

Based on the final DNN model, both individual-level and population-level impact scores of the demographic variables were calculated. Since age was a continuous variable, the impact of each individual's age on the risk was also calculated. The reference value for age was chosen to be the median age of the cohort, which was 63 years.

The single-numbered population-level impact score provides not only a succinct but also simplistic summarization of the relationship between the predicting variable and outcome (risk). It is comparable to the log odds ratios obtained from the LR model for the same variable. We can see (Table 4) that they were very close on age, but on other variables, the log odds ratios generally had much larger magnitude. The impact scores

Table 1 Summary of the demographic characteristics

Characteristics	Cases (N = 2355)	Controls (N = 3052)	Overall (N = 5407)
Age			
Mean±SD	68.8±13.6	56.6±15.8	61.9±16.1
Median (Q1, Q3)	70.5 (60.6, 76.3)	57.5 (44.3, 67.7)	63.0 (51.1, 72.8)
Gender			
Female	122 (5.2%)	382 (12.5%)	504 (9.3%)
Male	2233 (94.8%)	2670 (87.5%)	4903 (90.7%)
Race			
AA	1236 (52.5%)	1342 (44.0%)	2578 (47.7%)
White	930 (39.5%)	1377 (45.1%)	2307 (42.7%)
Other	48 (2.0%)	85 (2.8%)	133 (2.4%)
Unknown	141 (6.0%)	248 (8.1%)	389 (7.2%)

Table 2 Examples of some prevalent diagnoses and their prevalences in the cohort

ICD code	Description	Cases (*N* = 2355)	Controls (*N* = 3052)	Overall (*N* = 5407)
I10.	Essential (primary) hypertension	1624 (69%)	1338 (43.8%)	2962 (54.8%)
E78.5	Hyperlipidemia, unspecified	1078 (45.8%)	966 (31.7%)	2044 (37.8%)
E11.9	Type 2 diabetes mellitus without complications	892 (37.9%)	670 (22%)	1562 (28.9%)
M54.5	Low back pain	605 (25.7%)	839 (27.5%)	1444 (26.7%)
G47.33	Obstructive sleep apnea	568 (24.1%)	572 (18.7%)	1140 (21.1%)
K21.9	Gastro-esophageal reflux disease without esophagitis	543 (23.1%)	500 (16.4%)	1043 (19.3%)
E66.9	Obesity, unspecified	393 (16.7%)	525 (17.2%)	918 (17%)
F43.12	Post-traumatic stress disorder, chronic	328 (13.9%)	522 (17.1%)	850 (15.7%)

and log odds ratios agreed on signs/directions on age, gender_female(vs. male), and race_aa(vs. white), which means both models regarded the female gender as being associated with decreased risk while older ages and the AA race being associated with increased risk. However, they disagreed on race_other(vs. white) and race_unknown(vs. white). The impact scores of these two variables show that the other and unknown races had very small impacts on the risk although they were associated with decreased risks. In contrast, the log odds ratios show that these two race categories had big impacts on the risk and were associated with increased risks.

To deepen the understanding of the impact and impact score of age, we graphically present in Fig. 4 the impacts and impact scores at various levels: the individual level, the average by age, and the 7-year moving average by age. We see that the individual impacts/impact scores of age were very different even in patients of similar ages. This shows that the DNN model captured the non-linear relationship between the predictors and the outcome. For comparison, we present in Fig. 5 the impacts and impact scores calculated based on the LR model to show what they look like for a linear model.

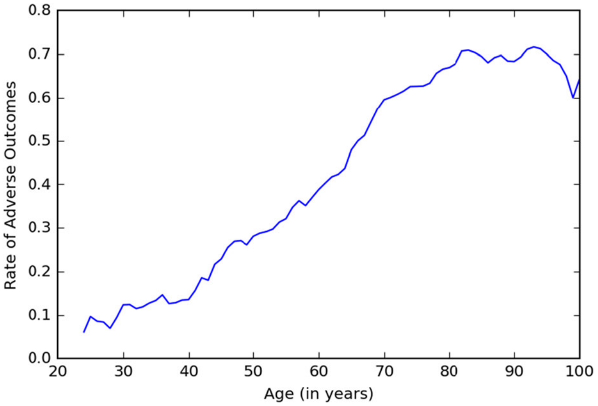

Fig. 1 Proportion (7-year moving average) of cases by age

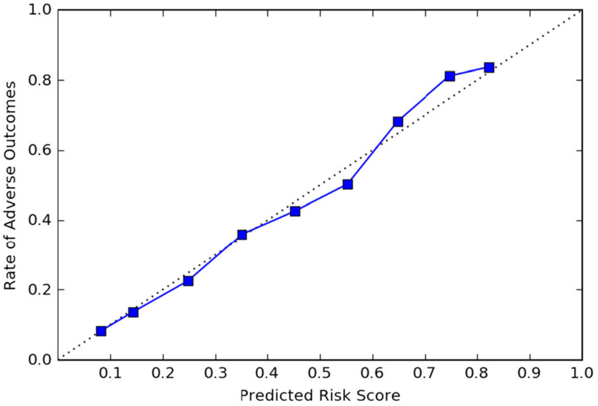

Fig. 2 The calibration curve of the predicted risk scores of the DNN model

The impact score of age should be interpreted as the change in risk per unit (1 year) change in age. Since the impact is the total change in risk, Fig. 4a and Fig. 4b are geometrically related as follows: the slope of the line connecting from the point (63,0) to any point in Fig. 4a corresponding to a patient is equal to the *y*-value of the point in Fig. 4b corresponding to the same patient. In this sense, Fig. 4b does not introduce any new information than Fig. 4a but provides a different view in which the special role of the reference age is not pronounced. The same relationship holds for the LR model: in Fig. 5a, all the dots are on a line passing through the point (63,0), and in Fig. 5b, all the dots are on a horizontal line at a level equal to the slope of the line in Fig. 5a. Note that in Fig. 5a the reference age still plays a special role, but in Fig. 5b, the dependence on the reference age is completely removed.

Specifically, Fig. 4a shows that the impacts of age were positive on all patients aged >63 years and negative for all patients aged <63 years, which shows that age was associated with an increased risk for all patients. On the other hand, the average by age curves show that the growth of risk was faster for younger patients (age <70 years) but slower for older patients (age >70 years). This corresponds to what we see in Fig. 4b: lower ages had higher impact scores while higher ages had lower impact scores.

This may be due to the fact that older patients tend to have more comorbidities, and the model had learned to attribute the higher risks of the older patients more to their comorbidities than to their ages, which effectively reduced the impact score at higher ages. If this were true, then we would observe a negative correlation between the impact scores and the number of comorbidities among the elderly patients. Moreover, if this were true, then we could also explain the decreased risk for ages >90 years as shown in Fig. 1: the patients aged >90 years had fewer comorbidities than those aged 80–90 years, and this might be due to the survivor effect [29], i.e., the patients with

Table 3 Predictive performance of the DNN and LR models

Model	AUC	Accuracy	Sensitivity	Specificity
DNN	0.762	0.720	0.683	0.743
LR	0.732	0.669	0.686	0.659

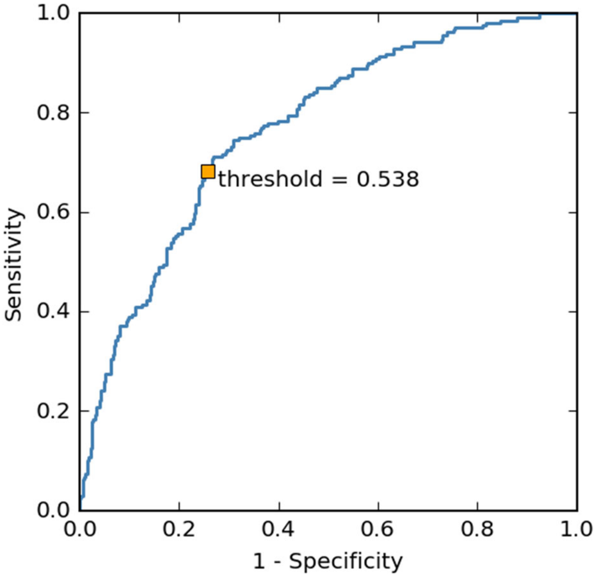

Fig. 3 The ROC curve of the DNN model on the test set. The square dot corresponds to the threshold which maximizes the accuracy

ages >90 years survived to an age of >90 years because they had fewer comorbidities than those with ages 80–90 years.

Therefore, we investigated the data further to find evidence for the above hypotheses, i.e., (1) the model attributed the higher risks of the older patients more to their comorbidities than to their ages, and (2) the patients aged >90 years had fewer comorbidities than those aged 80–90 years.

We first calculated the mean ICD count (number of distinct ICD codes) per patient by impact score (Fig. 6) on patients aged ≥60 years. We can see that as the impact score increases from 0 to 0.08, the average number of ICD counts per patient decreases from 48.6 to 6.4. This shows a negative correlation between the impact scores and the number of comorbidities among the older patients.

We next calculated the mean ICD count per patient by age (Fig. 7) and the prevalence of 6 common comorbidities (Fig. 8)—hypertension, hyperlipidemia, type 2 diabetes, obstructive sleep apnea, obesity, and chronic obstructive pulmonary disease—on all patients. All these show a common trend: as age increases starting from 20 years, the number/prevalence of comorbidities increases first and then

Table 4 Impact scores and log odds ratios of the demographic variables

Variable	Impact score	Log odds ratio
age	0.045	0.046
gender_female(vs. male)	−0.108	−0.370
race_aa(vs. white)	0.178	0.497
race_other(vs. white)	−0.037	0.535
race_unknown(vs. white)	−0.002	0.469

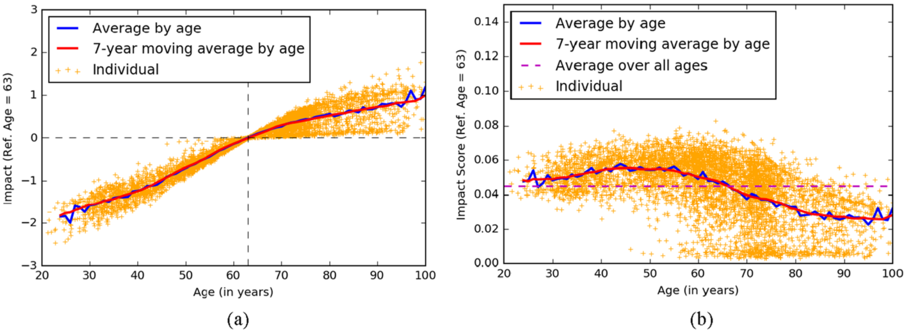

Fig. 4 Impacts and impact scores of age through the DNN model. The horizontal dashed line in (b) represents the population-level impact score of age (=0.045) as shown in Table 4

decreases. In particular, all show the same pattern that the patients aged >90 years had fewer comorbidities than those aged 80–90 years.

Similar to age, the variations of impact scores of the other variables were not revealed by the population-level impact scores shown in Table 4. The distributions of the impact scores (Fig. 9) demonstrate that, for all of the gender and race variables, there were both positive and negative impact scores, which means that these factors were associated with increased risks on some patients and decreased risks on the other patients. In fact, the “span” of the individual impact scores was much larger than the magnitude of the corresponding population-level impact scores for all the variables.

Lastly, we calculated the interaction score for each pair of demographic variables (Table 5). We can see that the interactions between age and the other variables were generally very small compared to the impact score of age. Ignoring the magnitudes, we find that the female gender, the AA race, and the other races were associated with slightly decreased impact score of age. Furthermore, the female gender was associated with decreased impact score of the AA race and the unknown race but with increased impact score of the other races.

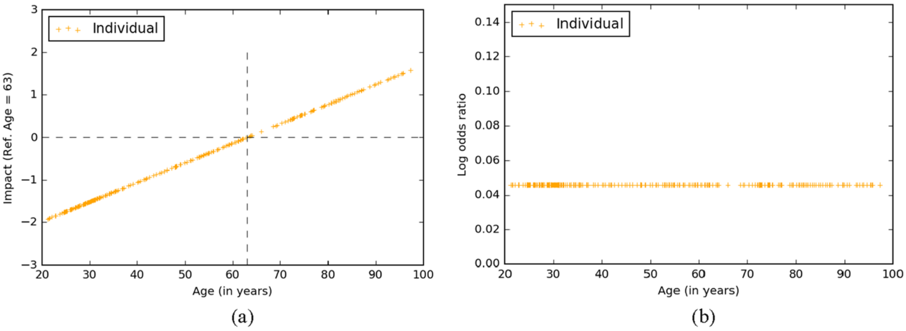

Fig. 5 Impacts and impact scores (=log odds ratios) of age through the LR model. Only the 300 randomly selected individuals are used for illustration

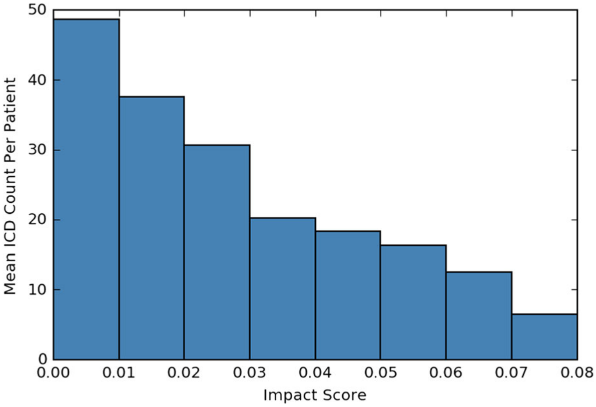

Fig. 6 Mean ICD count per patient by impact score of age on patients aged ≥60 years

4 Discussions and Conclusions

Based on our analysis, higher age and African American race are associated with increased risks while female gender is associated with decreased risks, which is consistent with the literature and empirical observations [30–34]. Age, in particular, is the most important factor. The values of impact score and odds ratio of age in Table 4 do not appear to be as large as those of gender and African American race. However, for age, impact score and log odds ratio are calculated as per unit (year) change. For a given patient who could easily be 5 or 10 years younger or older than the reference age of 63, his/her age can play a big role in the risk of adverse outcomes. We also found that age, gender, and race had very small interactions with each other: the female patients of a race other than AA and White are at a slightly higher risk than the combined (added) risks of being female and of other race, and the female African American patients are at slightly lower risk than the combined risks of being female and being African American. Other interactions are comparatively trivial.

For every variable, we observed a range of individual impacts/impact scores. This range should not be confused with the confidence interval. The overall (population-

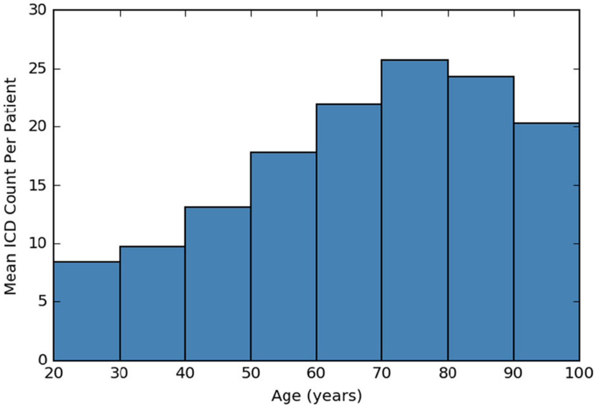

Fig. 7 Mean ICD count per patient by age

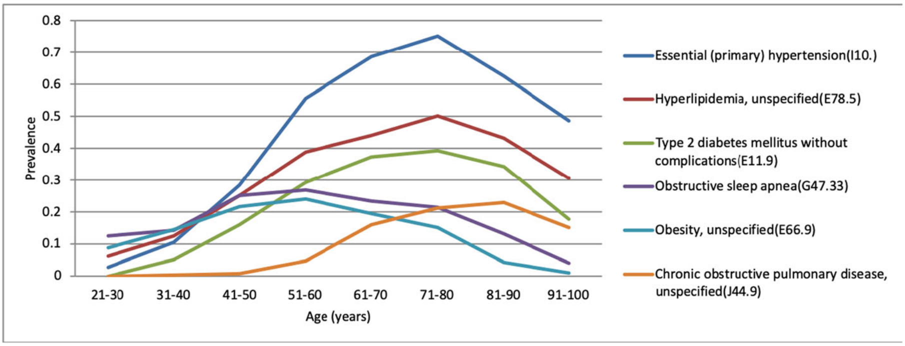

Fig. 8 Prevalence of six common comorbidities by age

level) impact scores are calculated directly as the mean of the individual impact scores. As shown in Fig. 4b, in many cases, the individual scores can be quite different for a given age. Especially in some patients aged >60 years, the impact score of age is very small (near zero). This result may occur because of the presence or absence of comorbid conditions. Some patients have more than 3 chronic conditions while some have none. In the DNN model, their risk of adverse outcome was mainly explained by their comorbid conditions.

While the effects of demographic variables are notable, even when combined with comorbidities, the risk prediction performance has room for improvement. The individual differences are also very large. Broad guidelines such as allowing individuals below 60 years old to continue to work may not be sufficiently precise.

Based on Fig. 4, the effect of age can be observed to be non-linear, nor should we expect it to be. While it is common for scientific literature to report the increase in risk per one unit of a measure (e.g. year), it is also well known that relationships between certain variables and outcomes are not linear. For example, the relationship between weight and mortality is considered to be a "U" shape with both over and underweight associated with high mortality [35]. Based on the impact and impact score analysis, age is a bigger risk factor in the younger patients (aged <50–60 years) than the older patients.

To assess the "impact" of a variable on the outcome, we have two measures: the impact and the impact score. The former is new in this paper whereas the latter has been introduced before. The two measures are closely related, but they serve different purposes and are useful in different situations. For the individuals, the impact is more useful than impact score, because the impact measures the absolute change in risk while

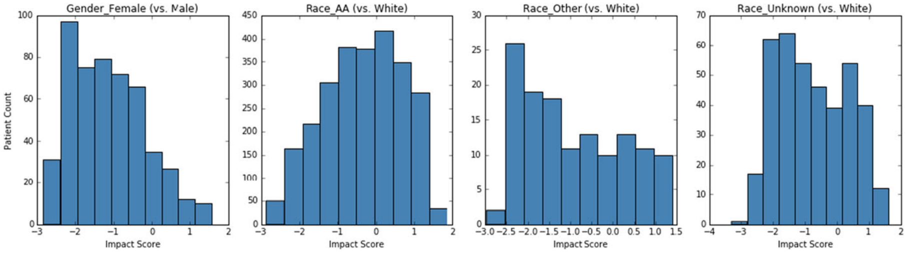

Fig. 9 Frequency distributions of the individual-level impact scores of the gender and race variables

Table 5 Interaction scores between the demographic variables

Variable #1	Variable #2	Interaction score
age	gender_female(vs. male)	−0.00075
age	race_aa(vs. white)	−0.00063
age	race_other(vs. white)	−0.00081
age	race_unknown(vs. white)	0.00094
gender_female(vs. male)	race_aa(vs. white)	−0.00703
gender_female(vs. male)	race_other(vs. white)	0.01042
gender_female(vs. male)	race_unknown(vs. white)	−0.00154

the impact score measures the rate of change in risk per unit change in the variable. It is possible for an individual that the impact score is small, but the total impact is large; thus, impact is more informative than impact score in this case. On the other hand, to understand the overall impact of a variable on the outcome, we must move to the population-level, and impact score will be necessary because only the average of the (individual-level) impact scores over a population is meaningful. For the interaction and interaction score, the same arguments apply as well.

One concern about the impact score and interaction score as explanations may be their dependence on the reference values. This is a valid concern because for linear models, with which we are most familiar, the explanations (e.g., log odds ratios) are independent of any reference values. However, such experience does not transfer directly to non-linear models, and we choose to make the reference values as part of the explanations to deal with the complexity of those models. Actually, the idea of reference values is not new, and we have already used it in linear regression models, that is, when we deal with multivalued categorical variables (e.g., race). To incorporate a categorical variable with *n* values into a regression model, the standard procedure is to convert the categorical variable into *n*-1 binary variables, with one value chosen as the reference. It is obvious that the final explanation of categorical variable would depend on what the reference value. Therefore, it is not too surprising to see our explanation method depends on reference values. However, we still strive to reduce the variability of the explanations, by making as a general guideline to choose the most “common” value as the reference value. This is intuitively plausible, since we usually compare a particular individual to an “average” person. Although there may still be different choices (e.g., mean, median, and mode) for the most “common” value, the difference is usually very small, and the explanations should be consistent. Nevertheless, we emphasize that the reference values are part of the explanations. We also note that reference values have also been used by other machine learning explaining methods such as LIME [21] and Integrated Gradients [36].

Another concern may be the different values and even the opposite signs between the impact scores and the log odds ratios for the same variables (Table 4), and more generally, different impact scores that may be derived from different DNN models for the same variables. On one hand, the difference in the values comes from the difference in the models, which further comes from the different assumptions underlying the models. On the other hand, we should acknowledge that every model is an approximation to the "ground truth," which underlies the data but is unknown. Therefore, it is expected to see that some parts of the explanations of different models are similar (e.g., age in Table 4), and the other parts are far apart (e.g., other race in Table 4).

The explanation method in this study is designed for DNNs whose output layer is a single node with a sigmoid activation function. However, this method is not limited to such type of output layer and can be easily modified to adapt to other types. For example, for DNNs making multinomial classification or prediction, assuming there are *n* classes, there will be *n*-1 impact scores for each predictor variable, and *n*-1 interaction scores for each pair of predictor variables. Actually, this is similar to the generalization of the odds ratios for binary logistic regression models to multinomial regression models. Moreover, the impact score and interaction score for multinomial regression models, the impact scores and interaction scores should coincide with the generalized log odds ratios and interaction coefficients for such models.

In summary, this study showed that the DNN model was able to capture the complicated non-linear relationship between the risk factors and the adverse outcome, and the explanation method we developed provided a tool to find the complicated non-linear effects of the demographic variables.

5 Limitations and Future Work

The sample size of about 5000 COVID-19 patients is not big data in the traditional sense, but it is larger than most published COVID-19 studies. As we accumulate more data, we plan to repeat these analyses.

We used hospitalization and mortality as a combined outcome. This is a common approach in outcome dichotomization (adverse vs. favorable) as both reflect a negative clinical connotation and the latter is a competing risk factor of the former. We did so also because the number of mortality cases is relatively low, which would mean a low accuracy in prediction for this outcome. However, we understand that hospitalization and mortality are different outcomes, and there are interests to study their own respective risk factors. Moreover, a small number of COVID hospitalizations were used in order to protect the patients with unstable housing or other social situations because this way a contagious person would be physically isolated from close contacts and/or the community. Therefore, in future analyses, especially when we have a much larger cohort, we plan to separate the two outcomes.

In this study, the deep learning model performed better than logistic regression in terms of AUC, but only by 3 percentage points. Since the number of Veterans with positive COVID-19 tests has doubled [4] in the 6–7 weeks after May 1, 2019 (the cutoff date used by this cohort creation), we can now have a much larger cohort, with which we anticipate the deep learning model's performance will improve. We limited the number of hidden layers in our deep learning model to 4 because of the large number of hyperparameters. With a larger sample, we can include more hidden layers and experiment with more complex architectures to optimize the model performance. Arguably, a better-fitted model can lead to a better understanding of the relationship between predictors such as demographics and outcomes.

Funding This study was supported by Award Number UL1TR001876 from the NIH National Center for Advancing Translational Sciences.

Declarations

Ethics approval This study was approved by the Institution Review Board for the Washington DC VA Medical Center.

Conflict of interest The authors declare no competing interests.

Disclaimer Its contents are solely the responsibility of the authors and do not necessarily represent the official views of the National Center for Advancing Translational Sciences or the National Institutes of Health.

References

1. WHO. Coronavirus disease (COVID-2019) situation report 64. https://www.who.int/docs/default-source/coronaviruse/situation-reports/20200324-sitrep-64-covid-19.pdf?sfvrsn=703b2c40_2. 2020].
2. Wu Z, McGoogan J (2020. Published online February) Characteristics of and important lessons from the coronavirus disease 2019 (COVID-19) outbreak in China. Summary of a report of 72 314 cases from the Chinese Center for Disease Control and Prevention. JAMA 24:2020. https://doi.org/10.1001/jama.2020.2648
3. CDC. Coronavirus Disease 2019 (COVID-19): Cases in U.S. https://www.cdc.gov/coronavirus/2019-ncov/cases-updates/cases-in-us.html. 2020].
4. VA Office of Public Health. Novel coronavirus disease (COVID-19): VA COVID-19 cases. https://www.publichealth.va.gov/n-coronavirus/. 2020 March 26, 2020].
5. Price-Haywood EG, Burton J, Fort D, Seoane L (2020) Hospitalization and mortality among black patients and white patients with Covid-19. N Engl J Med 382(26):2534–2543
6. Weiss P, Murdoch DR (2020) Clinical course and mortality risk of severe COVID-19. Lancet 395(10229):1014–1015
7. Zheng Z, Peng F, Xu B, Zhao J, Liu H, Peng J, Li Q, Jiang C, Zhou Y, Liu S, Ye C, Zhang P, Xing Y, Guo H, Tang W (2020) Risk factors of critical & mortal COVID-19 cases: a systematic literature review and meta-analysis. J Inf Secur 81:e16–e25
8. Jordan RE, Adab P, Cheng KK (2020) Covid-19: risk factors for severe disease and death. BMJ 368:m1198
9. Epidemiology Working Group for Ncip Epidemic Response, C.C.f.D.C. and Prevention (2020) The epidemiological characteristics of an outbreak of 2019 novel coronavirus diseases (COVID-19) in China. Zhonghua Liu Xing Bing Xue Za Zhi 41(2):145–151
10. Team CC-R (2020) Severe outcomes among patients with coronavirus disease 2019 (COVID-19) - United States, February 12-March 16, 2020. MMWR Morb Mortal Wkly Rep 69(12):343–346
11. Chen R et al (2020) Risk factors of fatal outcome in hospitalized subjects with coronavirus disease 2019 from a nationwide analysis in China. Chest
12. Gong J et al (2020) A tool to early predict severe corona virus disease 2019 (COVID-19) : a multicenter study using the risk nomogram in Wuhan and Guangdong, China. Clin Infect Dis
13. Pastur-Romay LA et al (2016) Deep artificial neural networks and neuromorphic chips for big data analysis: pharmaceutical and bioinformatics applications. Int J Mol Sci 17(8)
14. Munir K et al (2019) Cancer diagnosis using deep learning: a bibliographic review. Cancers (Basel) 11(9)
15. Mumtaz W, Qayyum A (2019) A deep learning framework for automatic diagnosis of unipolar depression. Int J Med Inform 132:103983
16. Miotto R et al (2017) Deep learning for healthcare: review, opportunities and challenges. Brief Bioinform
17. Lee JG, Jun S, Cho YW, Lee H, Kim GB, Seo JB, Kim N (2017) Deep learning in medical imaging: general overview. Korean J Radiol 18(4):570–584
18. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. Nature 521(7553):436–444
19. Gunning D, Aha DW (2019) DARPA's explainable artificial intelligence program. AI Mag 40(2):44–58
20. Bahdanau D, Cho K, Bengio YJapa (2014) Neural machine translation by jointly learning to align and translate. arXiv

21. Ribeiro MT, Singh S, Guestrin C Why should i trust you?: explaining the predictions of any classifier. In: Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining. 2016. ACM, New York
22. Binder A et al (2016) Layer-wise relevance propagation for neural networks with local renormalization layers. In: International Conference on Artificial Neural Networks. Springer, Berlin
23. Chakraborty S et al (2017) Interpretability of deep learning models: a survey of results. In: IEEE SmartWorld, Ubiquitous Intelligence & Computing, Advanced & Trusted Computed, Scalable Computing & Communications, Cloud & Big Data Computing, Internet of People and Smart City Innovation, vol 1-6. IEEE, San Francisco
24. Zeng-Treitler QSY, Redd D, Goulet J, Brandt C, Bray B (2019) Explaining AI models for clinical research: validation through model comparison and data simulation. In: IADIS International Conference e-Health 2019 (part of MCCSIS 2019)
25. Redd D et al (2020) Using explainable deep learning and logistic regression to evaluate complementary and integrative health treatments in patients with musculoskeletal disorders. In: Hawaii International Conference on System Sciences (HICSS) Proceedings (Accepted)
26. Shao Y, Ahmed A, and Zeng Q (2019) Detection of covariate interactions by deep neural network models, in KDD Workshop on Applied Data Science for Healthcare. Bridging the Gap between Data and Knowledge, https://dshealthkdd.github.io/dshealth-2019/assets/DSHealth_2019_paper_11.pdf: Anchorage, AK, USA.
27. Niculescu-Mizil A, Caruana R (2005) Predicting good probabilities with supervised learning, in ICML ‘05. In: Proceedings of the 22nd international conference on Machine learning, pp 625–632
28. Sutskever I et al (2013) On the importance of initialization and momentum in deep learning. Proceedings of the 30th International Conference on Machine Learning. PMLR 28(3):1139–1147
29. Newman AB, Murabito JM (2013) The epidemiology of longevity and exceptional survival. Epidemiol Rev 35:181–197
30. Parohan M, Yaghoubi S, Seraji A, Javanbakht MH, Sarraf P, Djalali M (2020) Risk factors for mortality in patients with Coronavirus disease 2019 (COVID-19) infection: a systematic review and meta-analysis of observational studies. Aging Male:1–9
31. Lee JY, Kim HA, Huh K, Hyun M, Rhee JY, Jang S, Kim JY, Peck KR, Chang HH (2020) Risk factors for mortality and respiratory support in elderly patients hospitalized with COVID-19 in Korea. J Korean Med Sci 35(23):e223
32. Gubatan J, Levitte S, Patel A, Balabanis T, Sharma A, Jones E, Lee B, Manohar M, Swaminathan G, Park W, Habtezion A (2020) Prevalence, risk factors and clinical outcomes of COVID-19 in patients with a history of pancreatitis in Northern California. Gut gutjnl-2020-321772
33. Chen L, Yu J, He W, Chen L, Yuan G, Dong F, Chen W, Cao Y, Yang J, Cai L, Wu D, Ran Q, Li L, Liu Q, Ren W, Gao F, Wang H, Chen Z, Gale RP, Li Q, Hu Y (2020) Risk factors for death in 1859 subjects with COVID-19. Leukemia 34:2173–2183
34. Chen F et al (2020) Clinical characteristics and risk factors for mortality among inpatients with COVID-19 in Wuhan, China. Clin Transl Med
35. Troiano RP, Frongillo EA Jr, Sobal J, Levitsky DA (1996) The relationship between body weight and mortality: a quantitative analysis of combined information from existing studies. Int J Obes Relat Metab Disord 20(1):63–75
36. Sundararajan M, Taly A, Yan Q (2017) Axiomatic attribution for deep networks. In: Proceedings of the 34th International Conference on Machine Learning

Affiliations

Yijun Shao[1,2] • Ali Ahmed[1,2,3] • Angelike P. Liappis[1,2] • Charles Faselis[1,2] • Stuart J. Nelson[2] • Qing Zeng-Treitler[1,2]

Ali Ahmed
ali.ahmed@va.gov

Angelike P. Liappis
angelike.liappis@va.gov

Charles Faselis
charles.faselis@va.gov

Stuart J. Nelson
stunelson@email.gwu.edu

Qing Zeng-Treitler
zengq@email.gwu.edu

[1] Washington DC VA Medical Center, Washington, DC, USA

[2] George Washington University, Washington, DC, USA

[3] Georgetown University, Washington, DC, USA