Fourth International Conference on Electronics, Communication and Aerospace Technology (ICECA-2020)
IEEE Xplore Part Number: CFP20J88-ART; ISBN: 978-1-7281-6387-1

Personification and Safety during pandemic of COVID19 using Machine Learning

Prema Gawade
Research Scholar, Department of Computer Engineering
Pune Institute of Computer Technology
PUNE, INDIA

Prof. Sarang Joshi
Professor, Department of Computer Engineering
Pune Institute of Computer Technology
PUNE, INDIA

***Abstract*—COVID19 is a respiratory disease and World health organization [WHO] has classified this disease as pandemic because of its high mortality rates among people with poor medical history conditions. It is important to identify such individuals, who are not safe and to avoid severe complications if get exposed to COVID19. There is no system, which can alert the person based on his health condition. Proposing a machine learning approach would help to achieve safety, personification and mitigate the effects of COVID19 disease. In presence of scarcity of information about COVID19 for the purposes of model building is a key challenge. Proposed research used natural language embedded media data information maintained in hospitals about respiratory diseases and used this information to identify the individuals who can be not safe if exposed to COVID19 patients. The proposed approach will provide an intuitive way to understand the risk of being getting affected based on the immunization of respiratory system of an individual. The risk factor will provide a basis for personification and to take safety measures in this long lasting pandemic situation.**

***Index Terms*—COVID19, machine learning, personification, neural networks, privacy, symptoms, safety**

I. Introduction

Personification is an idea which provides better intents to the person based on his traits. Individual features that helps to distinguish. By using personification, it has been intended to ensure privatization to the user and hence ensure better security to protect his identity. In recent years studies have shown that respiratory diseases like asthma, allergy have increased their presence in the community and becoming one of the major factor for illness. Most of the respiratory diseases caused by infections and viruses [1], [2], [3]. The rate of increase of such respiratory diseases are caused by different reasons like socioeconomic factor, urbanization, population etc [4]. Today, people are under huge threat of one of the new respiratory virus, known as COVID19. In India approximately 78 lakhs active cases are there and 77 lakhs patients recovered and 1,17,748 deaths because of COVID19. It is a novel corona virus which cause illness with symptoms ranging from headache to difficulty in respiration. Corona virus gets transmitted between people and animals too. The reason in it is called as novel corona virus because it is causing severe acute respiratory syndrome which was never seen as symptom.

There are couple of signs which can help you to understand the infection. The symptoms are fever, shortness in breathing, continuous coughing, kidney failure, pneumonia and in some case death. Because of COVID19 there is a pandemic situation declared by WHO [5]. The total number of infected people across the globe are around 5 millions with 6 percent of mortality rates. WHO have provided the detailed information about the causes, symptoms, country wise number of infected people, measure to be taken, etc. As per the guidelines provided by different medical institutions and authorities, many countries imposed different restrictions like lock down, wearing of masks, traveling restriction etc. To mitigate the spread of novel corona virus. In addition, they have highlighted that this is a novel corona virus case for which there is not medicine as of now available.

Experts agree that social distancing and proper care through sanitization can help to reduce the growth of corona virus spread. This is very important step of personification to avoid collapse of healthcare infrastructure. It is witnessed that if proper care has not been taken to slowdown the spread of corona virus, healthcare infrastructures like US which is considered is one of the best healthcare system in world is also facing the challenges to cope up with novel corona virus [6]. The other long-term impacts like well being of people, economy of institutions are being under jeopardy. If any person is infected with COVID19, is isolated to give treatment for recovery. But based on the severity it can cause death and also people left with a higher level of depression. Almost all systems providing analysis about COVID19, are failed at providing apriory suggestion so that it can be prevented. There is no system which can alert the person based on his health condition. Proposed research used patient's information to identify the individuals who can be not safe.

II. Literature Survey

Zeng et.al. [7] developed an ML approach on SARS dataset and then applied the same model for anticipating novel corona virus samples. This approach show future expectation towards better integration of existing approaches towards finding better solutions using machine learning. Hassanien et al. [8] anticipated different approach of decisive model. This model performs analysis and inference of COVID19 for future predictions based on COVID19 data which is publicly available. Luccioni et al. [9] provided literature survey of different machine learning approaches applied on COVID19 analysis.

Most of the approaches discussed are useful in forecasting the spread of COVID19.

Hwang et al. [10] presented and discussed the applications of big data and machine learning for scalable analysis and forecasting of COVID19. This study provided a better insight to gauge the scalable demands of infrastructure and models to provide real time insights about COVID19 spread across regions. A. Hafeez et al. [11] introduced ResNet50 an deep learning based model which avoids the early overfitting and allows to understand better non-linearities for classifying the biological cells into normal, bacterial-pneumonia, COVID19, and viral pneumonia. This deep learning based approach report better results with 96.23% accuracy overall.

In, [11] the dataset is prepared with 1203 healthy patients, 68 COVID19 radiographs from 45 COVID19 patients, 660 patients with nonCOVID19 viral pneumonia, and 931 patients with a bacterial pneumonia. D. Bertolini et al. [12] proposed an approach of hierarchical analysis through CXR images for detection of COVID19 patterns. A public dataset used for the purpose of finding COVID19 related images and only 90 were related to COVID19 from 1144 X-ray images. From remaining one is normal (healthy) and five types of pneumonia. Aznarte JL et al. [13] provided insights about every day clinical confirmations in Madrid because of of dependence on biometeorological markers.

Ezzat and Ella [14] provided an interesting approach based on hybrid convolutional network which utilizes an optimization strategy to improve the the forecasting of COVID19 spread. Chatterjee et al. [15] developed a machine learning approach using linear regression useful in predicting the spread of COVID19. A deep neural network and vector auto regression model provided implemented on symptoms and pace of COVID19 cases in India. In another attempt, a family of EfficientNets, recently proposed in [16] have shown high performance on Imagenet dataset [17]. The adaptation of such pretrained model for COVID19 usecase is studied to understand the cost of adoption. This study proposed an insights about taxonomy and hierarchical compatibility of these architecture towards new problem statements.

For certain populations, the risk of corona virus is much higher than others. This list contains people who have chronic conditions, elder age etc. There are couple of study report published who suggests that if age is more then risk of death increases. In addition there are couple of observations from healthcare studies demonstrating people with heart disease, diabetics and blood clotting issues are more susceptible to the infection.

Though these observations and symptoms are easily analyzed still the death risk is quite difficult to understand [18], [19]. For healthy people the average death rate increases from 1 to 7 percent for people with diabetes and also at higher side, 15 percent for people above age 80 years. Evidences from small studies depicting that vulnerability is not straightforward to align with most important symptoms. In the presence of many number of symptoms and cases, simple rules can fail to understand the impact of complex factors such as blood group, accidents and weakness [20]. Which makes people more prone to corona virus infection.

III. Datasets

Data availability about symptoms is difficult. Dataset collected through embedded media applications contains patients historical data which helps us to understand the likelihood of corona virus infection given past medical history of individual. This dataset is of 2600 patients from different hospitals used for analysis purpose. Finalized and cleaned dataset contains total of 1000 records. Out of this final dataset is splitted into train and test sets with regular split ratio of 80%/20%, with 1000 people in the training and 200 in the test set. This is a standard ratio applied for ML training purpose. Looking the the size of dataset, research proposed is to ensure that there will be enough number of samples for testing and evaluation purpose. The labels for the classification task are annotated based on their corona virus test.

TABLE I: LIST OF FEATURES

Feature Name	**Description**
Pregnancy	Preganacy status and child birth.
Age	Current age in years, specified as an integer.
Gender	0 is female and 1 is male.
AdmissionDate	admission date into the hospital.
DischargeDate	discharge date from the hospital.
IcuState	0 for false and 1 for true.
Immunity	Immunity disorders.
Respiratory System	Diseases of the respiratory system.

List of features collected to apply machine learning models are Age, Gender, address, Race, Injury History, Discharge information and comments, Service Procedure data, Hospital Visits, ICU visits, Tobacco consumption, smoking evidence, BMI, Bloor Pressure, Diabetes PQI etc. Table I shows features selected and Feature engineering is applied to assign a binary encoding indicating the presence or absence of type of medical symptom.

IV. Proposed Approach

A. Machine Learning Approaches

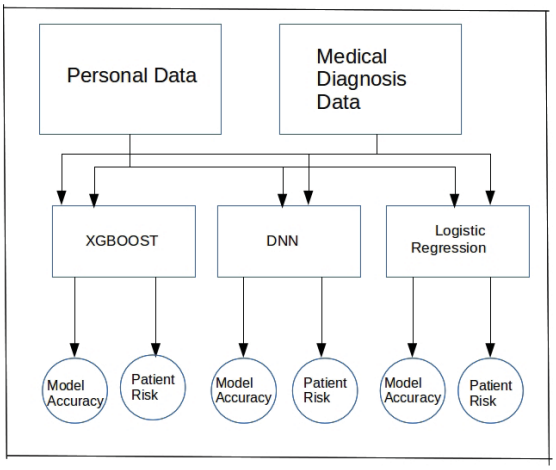

Fig. 1: Proposed Approach

Almost all systems are providing the analysis about COVID19, they are failed at providing apriory suggestion so that it can be prevented. There is no system which can alert the person based on his health condition. To identify the possibility of COVID19 infections and personification, research using 3 approaches to build models. These three approaches are implemented on the same dataset prepared as shown in Fig. 1. The metric for evaluating these three approaches are chosen as accuracy. This may not be the final verdict to select the model in addition interpretability and simplicity of the model can the criteria for pushing the model to production. Each model selection in based on the pros and cons of model property. Like in case of Gradient boosting tree the person specification and the medical history is used to analyze the impact.

1) Logistic Regression: Assume given dataset is,

$$\{\boldsymbol{x}_i, y_i\}_{i=1}^{n}, \boldsymbol{x}_i \in \mathbb{R}^d, \text{ and } y_i \in \{0, 1\} \tag{1}$$

Let the vector representation is,

$$\boldsymbol{x} = (x_1, x_2, \ldots, x_d) \tag{2}$$

and y is label for each vector in x, it is a ground truth generated through domain expert inputs. For Logistic Regression, class probability y = 1 can be modeled as follows, given x:

$$\begin{aligned} P(y = 1 \mid \boldsymbol{x}) &= p \\ &= \frac{1}{1 + \exp\left(-\boldsymbol{\beta}^\top \boldsymbol{x}\right)} \end{aligned} \tag{3}$$

where,

$$\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_d) \quad \in \mathbb{R}^d \tag{4}$$

the likelihood function can be written as,

$$\ell(\boldsymbol{y} \mid \boldsymbol{\beta}) = \prod_{i=1}^{n} p_i^{y_i} \left(1 - p_i\right)^{1 - y_i} \tag{5}$$

Algorithm 1: $SafetyFactorofCOVID19Patients$

Require: $PatientID$, $SymptomsDetails$
Ensure: $PersonifiedSafetyFactors.$
Preprocess the dataset:
$\Pr[f(D) \in O] \leq e^{\varepsilon} \cdot \Pr\left[f\left(D'\right) \in O\right]$
For each $XGBOOST$, DNN, LR
$y \leftarrow Train_Accuracy$
$loss \leftarrow L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i log\left(p_i\right) + (1 - y_i)\, log\left(1 - p_i\right)\}$
if $loss > threshould$ **then**
$loss \leftarrow$
$-\sum_{i=1}^{n} \{y_i log\left(p_i\right) + (1 - y_i)\, log\left(1 - p_i\right)\} + \lambda_\beta \|\boldsymbol{\beta}\|^2$
$Loss \leftarrow -loss$
$Y \leftarrow -y$
Return $Loss$, Y for each $XGBOOST$, DNN, LR
$SeedPoint \leftarrow 0$
$HighPoint \leftarrow 1$
$NoOfSafetyFactors \leftarrow 6$
$SafetyFactor$ = $(Math.abs(rangeLow) +$
$Math.abs(rangeHigh))/NoOfSafetyFactors$
end if
Return $SafetyFactor$

The Cross-entropy, logirithmic loss funtion is defined as,

$$L(\boldsymbol{\beta}) = \sum_{i=1}^{n} \{y_i \log\left(p_i\right) + (1 - y_i) \log\left(1 - p_i\right)\} \tag{6}$$

To improve the learning process, an L2 regularization parameter is substracted which basically provides better assessment of parameters so that it can be easy to reach to the global minima with minimum number of training epochs. So , the modified equation becomes,

$$L_\lambda(\boldsymbol{\beta}) = -\sum_{i=1}^{n} \{y_i \log\left(p_i\right) + (1 - y_i) \log\left(1 - p_i\right)\} + \lambda_\beta \|\boldsymbol{\beta}\|^2 \tag{7}$$

They identify risk features as age, specific disease history including individuals with diabetes, heart disease etc. are coming from the recommendations of domain experts through different channels, here looked for features that are important for model to guess the the seriousness and feed those feature vectors to the model. A dataset which contain the personal data and medical diagnosis of a person, used in training logistic regression model[10].

2) Gradient Boosting Decision Trees: GBDT algorithm creates many number of decision trees one after the other, where each tree tries to fit the residual of the previous trees. With ongoing efforts in refining the opensource libraries, it helps to achieve better results on difficult machine learning tasks like learning to rank (Burges 2010), prediciton of clicks(Ragno 2007) and many more.

Formally, GBDT is an ensemble model which trains a number of decision trees and perform sequential refinement on each tree successively. Given a convex loss function l and a dataset with n examples and d number of features,

$\{(\mathbf{x}_i, y_i)\}_{i=1}^{n} \left(\mathbf{x}_i \in \mathbb{R}^d, y_i \in \mathbb{R}\right)$. The Loss function minimization is achieved through,

$$\tilde{\mathcal{L}}^{(t)} = \sum_{i=1}^{n} \left[g_i f_t\left(\mathbf{x}_i\right) + \frac{1}{2} f_t^2\left(\mathbf{x}_i\right) \right] + \Omega\left(f_t\right) \tag{8}$$

In case of COVID19 case, it is important to preserve the privacy of the data. This privacy preservation is achieved through differential privacy theory which is proposed by Dwork et. al. Let f be a function and epsilon is an positive integer number which provides differential privacy for two different datasets D and D‘ is defined as,

$$\Pr[f(D) \in O] \leq e^{\varepsilon} \cdot \Pr\left[f\left(D'\right) \in O\right] \tag{9}$$

Above models result in higher accuracy. Drawback with these models is increase in complexity. To learn features that are eccentricities of the training data, but do not extend well to future data, gradient boosting trees are robust. Simpler XGBoost model allows diagnosis with full histories.

V. Result and Discussion

Patients Symptoms Dataset is used to demonstrate and evaluate the performance of different machine learning algorithms. The assessment of performance is made based of metrics such as ROC, AUC Curve, F1 score, Accuracy, precision and recall. All machine learning models are implemented in Python 3.7.1 and executed on Jupyter Notebook IDE. It is most popular IDE used to model development and supported machine and deep learning libraries and packages in scikit-learn, TensorFlow and Keras. This Data science work bench is installed on a Ubuntu system with an Intel Core i7 CPU at 2.4GHz, 12 GB of RAM.

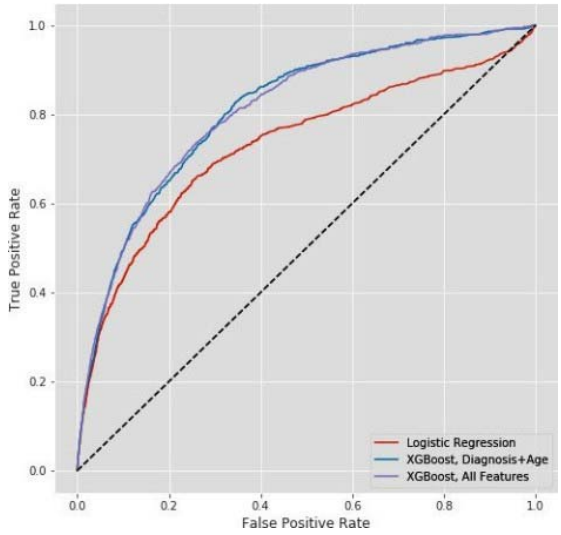

Fig. 2: ROC, AUC CURVE

As shown in Fig. 2, standard performance metrics like ROC and AUC curve used to visualize the results of research analysis. This above mentioned graph helps to understand the performance of classifier towards more number of true positive and least numbers of false positive. This helps in life critical fields like health care. The analysis shows that performance of both models is almost same. It is evident from the graph that COVID19 population and the threshold of decision boundary are linear to each other.

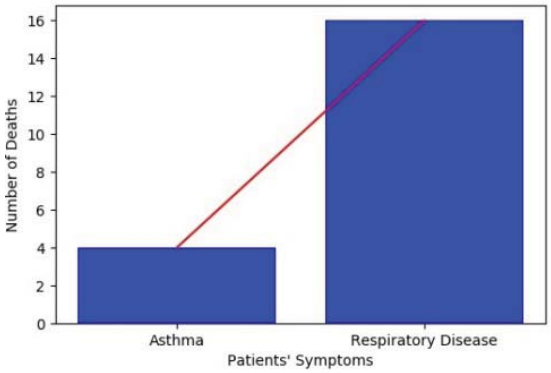

Fig. 3: Analysis of Symptomwise Death Rate

Logistic regression provides lesser sensitivity for alert rate which can be considered as future improvements. The second part of assessment aims at providing important features from the dataset which are the most useful features introducing linear non-linearity for model learning. The list of such important features is extracted using correlation. These are the features with highest effect on predication accuracy. As shown in the Fig. 3 one of the important factors causing deaths are Asthma and Diseases related to respiratory system. This analysis can help in proposing better features selection strategy for COVID19 prediction. It can allow in restricting the search space of the data for effective data collection.

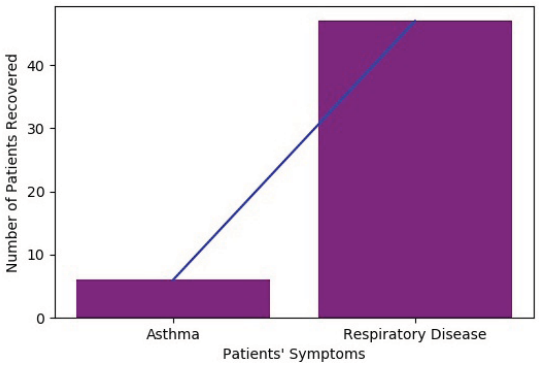

Fig. 4: Analysis of Symptomwise Recovery Rate

Continuation with the above analysis, Fig. 4 shows demonstration with recovery rate and featurewise impact on the cure of COVID19 infection. To some extent, it follows the same trend as above. The reason might be less number of samples to demonstrate the impact of these symptoms towards cure.

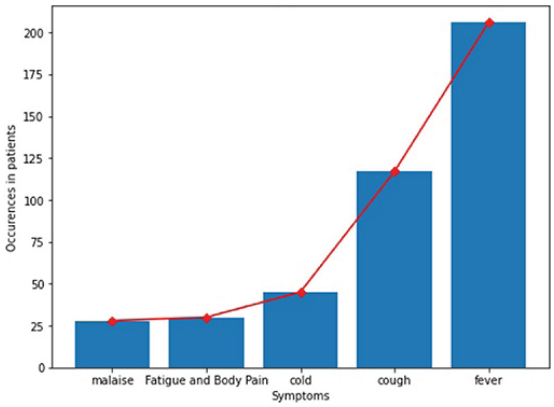

Fig. 5: Importance of Symptoms in COVID19

As shown in Fig. 5, it is evident that the common symptoms in COVID19 infections are in line with the important features listed by most of the machine learning models.

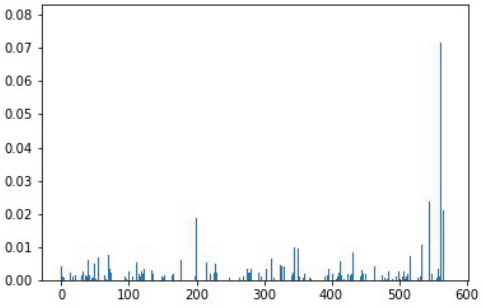

Fig. 6: Featurewise importance for Random Forest Classifier

Above Fig. 6 shows impacts of different symptoms when the database is trained on Random Forest Classifier. RFC is one of the classifier used in experiments. For demonstration purpose, sample is provided here. Analyses of probability scores are done and not the predictions so that it can help in personification and understanding the risk level of a person being infected by COVID19. Six levels are used ranging from -3 to 3 exclusing 0, where -3 category explains highest possibility of COVID19 infection and 3 is the one with least possibility of COVID19 infections. To allow such categorization, probabilities of each sample predictions are feeded to softmax function and then again packetized into six categories. Based on past health history, this is one the first approach which demonstrates the linkage of past medical history with the possibility of COVID19 infection.

VI. Conclusion

As health care facilities are limited in number in developing countries as well as the ecosystem is not well prepared to defend against in developed countries, the spread of the pandemic can claim thousands of lives. The way to address the situation proactive measure to understand the health of an individual based on his medical status, history etc. From this research, it is evident that machine learning models are well learnable with the knowledge which is provided to them through embedded media dataset and it supports the claims of subject matter experts in personification. Given the data about list of symptoms as expected by this research, pretrained models can provide the safety score with more than 95 percent accuracy. Based on this data, it is evident from research that one can get a safety measure. These risk factors will raise an alarm to prepare him based on preconditions and to achieve personification. This can help to observe and identify the disease well before to reduce the mortality rate even further than today.

References

[1] Wong GW, Leung TF, Ko FW. "Changing prevalence of allergic diseases in the Asia-Pacific region", Allergy Asthma Immunol Res.2013,5:251–7.
[2] Braman SS. "The global burden of asthma.Chest", 2006, 130(1 Suppl):4S–12S.
[3] Pawankar R, Bunnag C, et.al. "Allergic rhinitis and its impact on asthma in Asia Pacific and the ARIA update 2008", World Allergy Organ J., 2012, 5(Suppl 3):S212–7.
[4] World Health Organization. Global Surveillance, "Prevention and Control of Chronic Respiratory Diseases: A Comprehensive Report", Geneva: WHO, 2007.
[5] World Health Organization, WHO Director-General's Opening Remarks at the Media Briefing on COVID19",11th March 2020.Who.Int, Accessed 15 Mar. 2020.
[6] Specht, Liz.,"Simple Math Offers Alarming Answers about COVID19", Health Care. STAT, 10 Mar. 2020, www.statnews.com/2020/03/10/simple-math-alarming-answers-COVID19/. Accessed 15 Mar. 2020.
[7] Yang Z, Zeng Z, et.al.,"Modified SEIR and AI prediction of the epidemics trend of COVID19 in China under public health interventions", J Thorac Dis 2020, 12(3):165.
[8] Rizk-Allah RM, Hassanien AE,"COVID19 forecasting based on an improved interior search algorithm and multi-layer feed forward neural network", 2020, arXiv preprint arXiv:2004.05960.
[9] J. Bullock, A. Luccioni, et.al.,"Mapping the landscape of artificial intelligence applications against COVID19",2020,arXiv:2003.11336.
[10] Q. Pham, D. C. Nguyen, et.al.,"Artificial intelligence and big data for corona virus (COVID19) pandemic: A survey on the state-of-the-arts", 2020, doi:10.20944/preprints202004.0383.v1.
[11] M. Farooq, A. Hafeez,"Covid-resnet: A deep learning framework for screening of COVID19 from radiographs", 2020, arXiv preprint arXiv:2003.14395.
[12] R. M. Pereira, D. Bertolini, L. O. Teixeira, C. N. Silla Jr, Y. M. Costa, "COVID19 identification in chest x-ray images on flat and hierarchical classification scenarios, Computer Methods and Programs in Biomedicine", 2020,105532.
[13] Navares R, Díaz J, et.al.,"Comparing ARIMA and computational intelligence methods to forecast daily hospital admissions due to circulatory and respiratory causes in Madrid", 2018, Stoch Env Res Risk Assess 32(10):2849–2859.
[14] Ezzat D, Ella HA,"GSA-DenseNet121-COVID19: a hybrid deep learning architecture for the diagnosis of COVID19 disease based on gravitational search optimization algorithm", 2020, arXiv preprint arXiv:2004.05084.
[15] Sujatha R, Chatterjee J,"A machine learning methodology for forecasting of the COVID19 cases in India", 2020.
[16] M. Tan, Q. V. Le, "Efficientnet: Rethinking model scaling for convolutional neural networks", 2020, arXiv preprint arXiv:1905.11946.
[17] J. Deng, W. Dong, et.al.,"Imagenet: A largescale hierarchical image database", IEEE Conference on Computer Vision and Pattern Recognition, 2009, pp. 248–255.
[18] Page, Michael Le. "Why Is It so Hard to Calculate How Many People Will Die from COVID19? new Scientist", 11 Mar. 2020, Accessed 15 Mar. 2020.

[19] Zhou F, Yu T, et al.,"Clinical course and risk factors for mortality of adult inpatients with COVID19 in Wuhan", China: a retrospective cohort study [published online ahead of print, 2020 Mar 11] [published correction appears in Lancet. 2020 Mar 12;:]. 2020, doi:10.1016/S0140-6736(20)30566-3.
[20] Hubbard, Ruth E, et al.,"Frailty Status at Admission to Hospital Predicts Multiple Adverse Outcomes. Age and Aging",vol.46,no.5,2017,pp.801-806.