Applying Statistical Machine Learning Methods to Analysis Differences in the Severity Level of COVID-19 among Countries

Wen Yin[1*], Chenchen Pan[2*], Nanyi Deng[3], Dong Ji[4]

[1] Department of Computer Science, Columbia University, NYC, NY, USA.
[2] Department of Management Science and Engineering, Stanford University, Palo Alto, CA, USA.
[3] Department of Applied Analytics, Columbia University, NYC, NY, USA.
[4] SuZhou Trust Co., SuZhou, Jiang Su, China.

* Corresponding author. Tel.: +1 646-469-8934; email: wy2337@columbia.edu.

Manuscript submitted March 10, 2021; accepted May 13, 2021.
doi: 10.17706/jsw.16.5.219-234

Abstract: The COVID-19 pandemic has caused a significant negative impact on countries around the world, and there appears to be an observable difference in severity among nations. This study aims to provide an insight into the roles many social and economic factors played in contributing to this variation. By investigating potential patterns through exploratory data analysis, followed by constructing models using several popular machine learning techniques, we examine the validity of the underlying assumptions and identifying any potential limitations. Total deaths per million population is used as dependent variable with log transformation to remove outliers. A set of factors such as life expectancy, unemployment rate and population are available in the dataset. After removing and transforming outliers, various machine learning methods with cross validation are implemented and the optimal model is determined by predefined metrics such as root-mean-squared-error (RMSE) and mean-squared-error (MAE). The results show that the Gradient Boost Machine (GBM) technique achieves the most optimal results in terms of minimum RMSE and MAE. The RMSE and MAE values indicate no over fitting issues and the GBM algorithm captures the most influential factors such as life expectancy, healthcare expense per Gross Domestic Product (GDP) and GDP per capita, which are clearly critical explanatory variables for predicting total deaths per million population.

Keywords: COVID-19, machine learning, social and economic factors.

1. Introduction

First identified at the end of 2019, COVID-19 has inflicted immense destruction upon the global population. According to data published by the World Health Organization (WHO), as of September 18th, 2020, there have been more than 30 million confirmed cases worldwide including more than 900 thousand deaths reported [1]. The nature of the COVID-19 virus makes it significantly more infectious and persistent than any other infectious diseases seen in recent history [2]. There have been repeated occurrences of pandemic outbreaks documented in human history, and existing research suggests that new pandemics will likely occur in roughly 10-50 years intervals [3]. With an expanding world population and increasing disturbances of the ecological systems through human activities, it is likely that global pandemics will occur again in the future. By developing a deep understanding of the factors that contribute to the spread of COVID-19, effective policies and preventative measures can be designed to attenuate not only the continuing crisis of COVID-19 itself, but also other negative effects of highly infectious diseases that may occur in the future.

Though the outbreak of COVID-19 started only months ago, there have been numerous articles published on COVID-19 related topics. For example, Matthew conducted a research to study the key epidemiologic parameter estimates for coronavirus disease identified in peer-reviewed publications, preprint articles, and online reports. The study findings are based on 101,927 cases and 3,486 deaths in 94 countries spanning 6 continents and reveal that the range estimates for incubation period are 1.8–6.9 days, serial interval 4.0–7.5 days, and doubling time 2.3–7.4 days. The research concludes that case burden and infection fatality ratios increase with patient age. Moreover, the implementation of combined interventions can reduce cases and delay epidemic peak up to 1 month [4].

Nussbaumer conducted a review on the effectiveness of quarantine during severe coronavirus outbreaks [5]. The review provides summary and analysis of 51 studies: 1) 4 observational studies, 2) 28 modelling studies on COVID-19, 3) 1 observational and one modelling study on MERS, 4) 3 observational and 11 modelling studies on SARS, and 5) 3 modelling studies on SARS and other infectious diseases. The paper reports a benefit of the simulated quarantine measures to avert 44% to 96% of incident cases and 31% to 76% of deaths compared to no measures based on different scenarios [5].

In addition, Liu conducted a systematic review and meta-analysis to study the children with COVID-19 symptoms. The dataset is collected and built from 20 studies with 4300 pediatric patients sourcing from PubMed, Google Scholar, and Web of Science. It concludes that, while children's COVID-19 symptoms are characteristic by mild presentation, they may become the main spreader in the pandemic, if government loose the strict managements like school closures [6].

Overall, research completed so far has mainly focused on transmission rates, death rates and effectiveness of quarantine policy, however, most studies ignore important long-term social and economic factors like, GDP per capita, hospital beds per thousands, life expectancy and society median age. Therefore, our research uses data from worldindata.org, tradingeconomics.com and the World Bank, seeking to provide an insight into the vital roles these social and economic factors can play in influencing the severity of the pandemic in a specific country.

After conducting data collection, clean and examination, we were left with data from around 120 countries. We chose to analyze data as of May 31, 2020, which specify the first wave of COVID-19. We began by conducting exploratory data analysis to gain a basic understanding of our data, noting any potential flaws along the way. Then, we performed clustering to identify potential subgroups within the country-level data. Afterwards, we ran a regression analysis to determine which predictors have significant effects on a country's coronavirus death rate. We also looked for potential non-linear relationships. Finally, we ran a decision tree on the analysis and compared its performance against our regression model. We conducted this analysis with the aim of understanding the differences in the severity level of COVID-19 between countries. Finally, we concluded that life expectancy, healthcare expense per GDP and GDP per capita are the most influential social and economic factors, which are critical in determining COVID-19 severity difference among countries.

2. Exploratory Data Preprocess and Analysis

Data used in this study was published by "Our World In Data" (https://ourworldindata.org) a part of their "research and data to make progress against the world's largest problems" initiative. Total death per million caused by COVID-19 in 210 geographical regions as of May 31th, 2020 and other COVID-19 related data were obtained from the "Our World in Data Covid-19 Dataset". The respective social economic factors, including stringency index, population, population age distribution, unemployment rate, balance of trade, hospital beds per thousand, life expectancy, health care expenditure as a percentage of GDP, diabetes prevalence and population density were also extracted from "Our World in Data" website.

2.1. Exploratory Data Preprocess

There exists a considerable portion of missing entries in the raw data as displayed in Fig. 1. This is particularly true for the variables related to new COVID-19 tests administered**.** The reasons behind this is that many countries have limited test capacities for detecting patients actively infected with the virus or chose not to report. Since there is no sensible way of filling in the missing data, variables with more than 30% of missing values, these are mainly variables related to new tests, were excluded from the analysis. Variables that exhibit high levels of multicollinearity with others, such as death rates per day and new deaths per million were also excluded. Geographical regions where data on various social and economic factors are scarce for reasons unrelated to the virus, such as Andorra, were removed in the data cleaning process. At last, we conclude the data of 124 geographical regions.

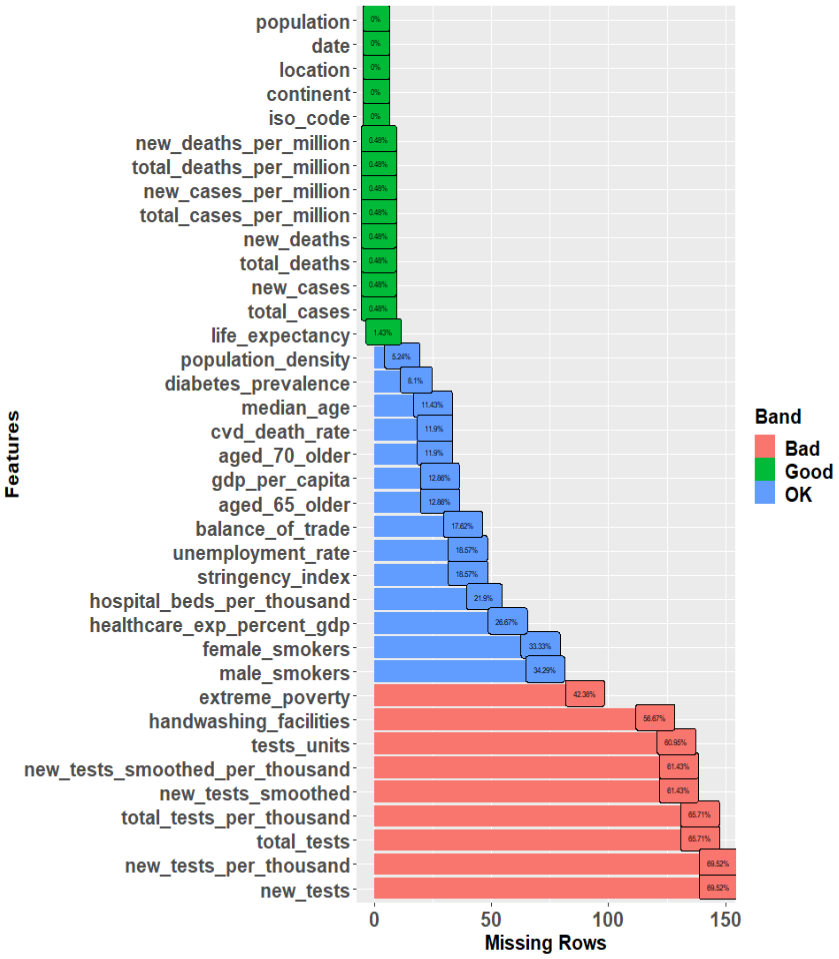


Fig. 1. Features against missing rows.

2.2. Data Transformation

Jack's research proposed that usually healthcare expenditure, unemployment rate, life expectancy, and stringency index are considered close to a normal distribution [2] and Fig. 2 below validates his view. Punn's research proposes that the population, population density, total COVID-19 cases, and total COVID-19 deaths pers million often show a strong right skewness [7], which can also be observed in our chart. Since it is not necessary that all predictors have a normal distribution, we choose to log transform the response variable, i.e. total deaths per million. Fig. 2 shows side-by-side histograms of each variable. Total deaths per million show a strong right skewness, indicating that a log transformation would be helpful.

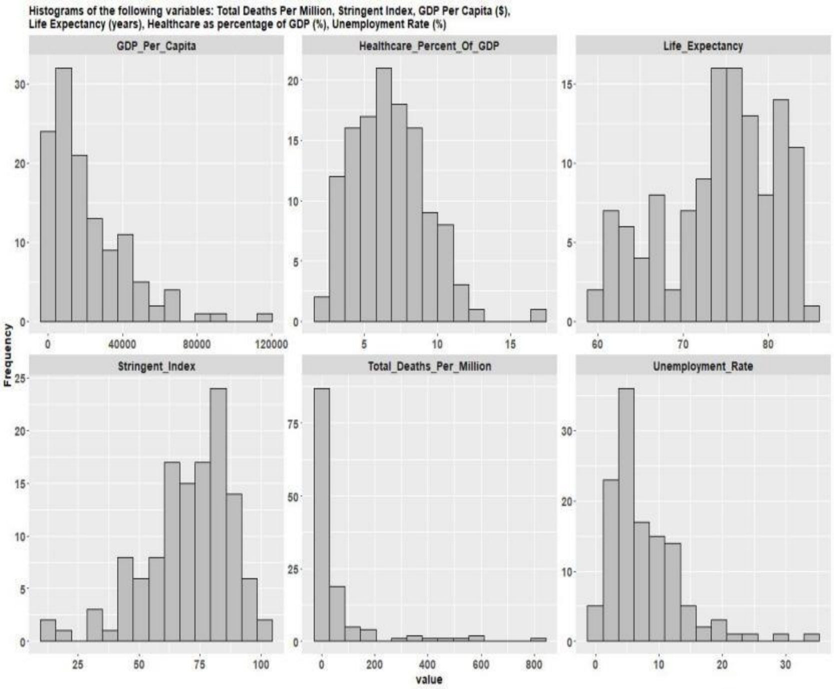


Fig. 2. Side-by-side histograms of each variable.

2.3. Pairwise Correlations

Next, we plot pairwise correlations and check for potential remaining multicollinearity between our predictors. Lauer's research proposes that 0.85 correlation coefficient could be viewed as a threshold for determining multicollinearity in COVID- 19 related civilian data [8]. From Fig. 3, it seems that the level of pairwise correlations is not an issue concerned in our data. No two predictors with a correlation greater than 0.85 were displayed. Therefore, we do not expect multicollinearity to be able to significantly reduce the precision of our estimates of the regression coefficients.

Fig. 3 is the Pairwise Correlation Plot. There does not seem to be evidence of significantly high multicollinearity in the data.

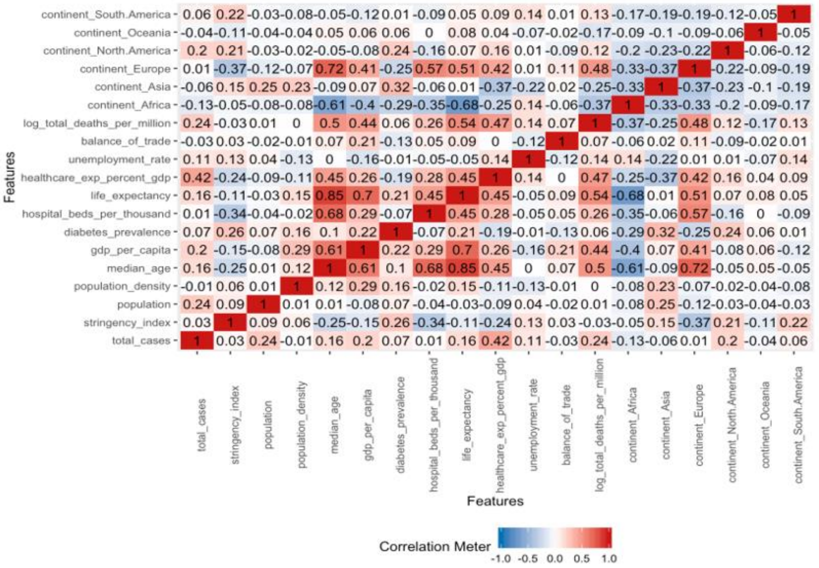


Fig. 3. Pairwise correlation plot.

In Fig. 4, we plot the scatterplots of total deaths per million with continent. We note that the continents of Africa, Asia, and Oceania have countries with relatively low death rates, while Europe has a number of countries with a death rate of over 400 per million.

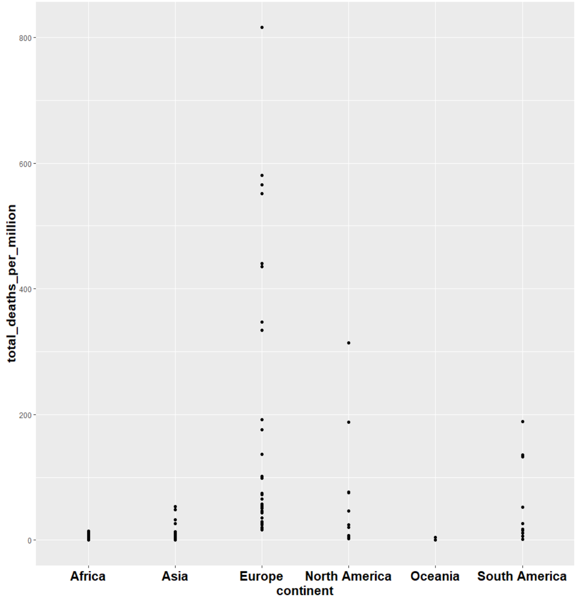


Fig. 4. Scatterplots of predictor variables against the response variable.

3. Methods

We apply both unsupervised and supervised learning methods to discover the association between total deaths per millions and the predictors and predict total deaths per million, respectively.

In unsupervised learning, we choose K-means clustering as it is a widely-known clustering method and well-implemented in R. Additionally, this algorithm runs faster than other clustering algorithms because only "k" distances are needed to compute per iteration. The purpose of implementing K-means clustering algorithm is to uncover association between dependent variable and predictors. For example, we want to see if high total deaths per million is associated with low GDP per capita.

For the supervised learning methods, we start with linear regression as it is the most well-known method and relatively easy to interpret. Due to relatively small sample size with 124 observations and 11 features with moderate signs of multicollinearity from the variation inflation index table, we decide to try forward / backward selection method and regularization method such as LASSO in order to eliminate model overfitting problem.

Linear regression is only limited to linear relationships between dependent variable and predictors. Therefore, we also implement ensemble (random forest) and boosting (gradient boosting) methods to uncover potential non-linear relationships between dependent variable and the predictors. Further, Cross validation technique is used to select the best set of hyper-parameters and eliminate model overfitting.

3.1. K-Means Clustering

K-means clustering partitions data into groups such that between groups variations are maximized (groups do not share many common similarities) and within group variations are minimized (samples within group are homogeneous). K-means clustering provides an important insight into the features that contribute to the variations in death rates among different countries. Here we use three clusters with the intention to mimic groups with high, medium and low levels of average death rates. The clustering was done based on scaled variables. As can be seen in Fig. 5 , these variables appear to have the ability to split total deaths into the desired groups, namely, high (shown in red), medium(shown in green) and low(shown in blue), with relatively few number of outliers.

It can also be observed from the scatter plots in Fig. 6.1, 6.2 that certain social and economic variables, such as GDP per capita, stringency index and life expectancy, exhibit considerable power in categorizing severity of the crisis into the three different groups. Higher values of the stringency index are correlated with a lower death rate. This correlation is expected as the stringency index measures the extent of the restrictions placed by governments in response to the pandemic; tougher measures, such as strict social distancing policies, are expected to have a positive impact.

Higher and lower income countries, as measured by GDP per capita, tend to experience lower death rates than middle income countries. One possible explanation for richer nations to have lower death rates is that these countries tend to have well established health systems that have a lower likelihood of collapsing and can function better under pressure. Studies have shown that older people are at much higher risk of suffering severe consequences from COVID-19 [9] and we can see from the scatter plot countries with the highest life expectancy tend to have higher death rates as well.

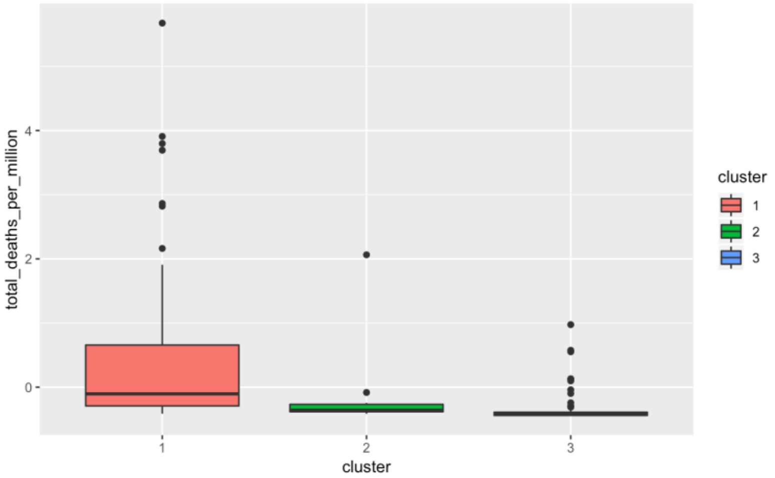


Fig. 5. Boxplot of total deaths per million by cluster.

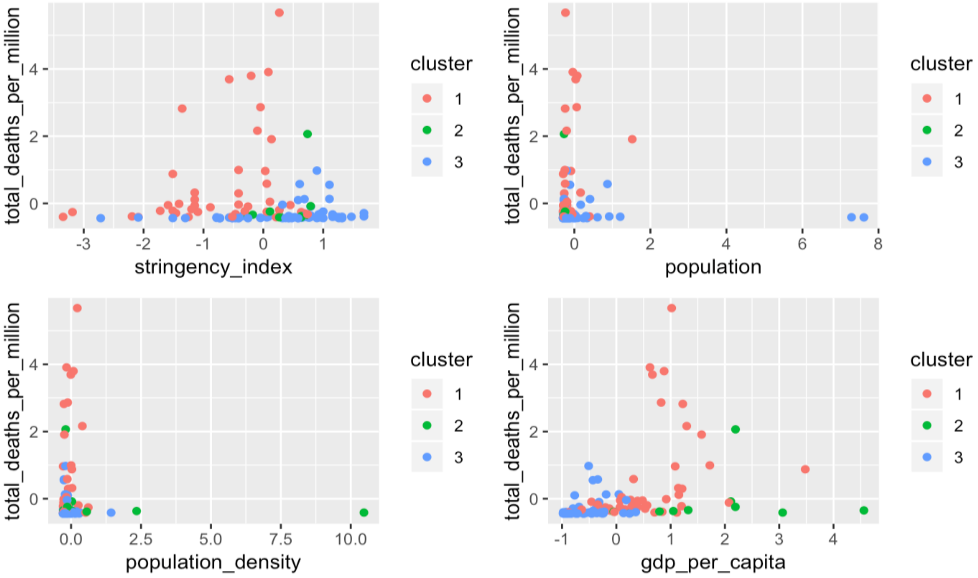


Fig. 6.1 Scatter plot of explanatory variables and deaths per million.

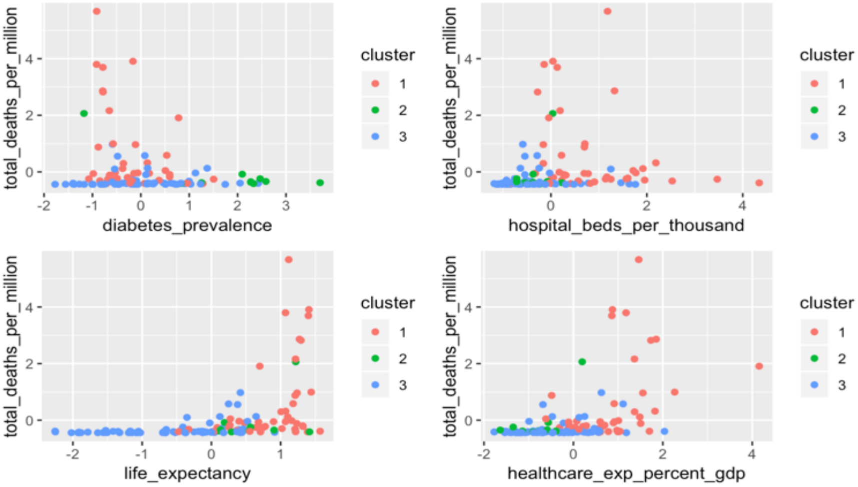


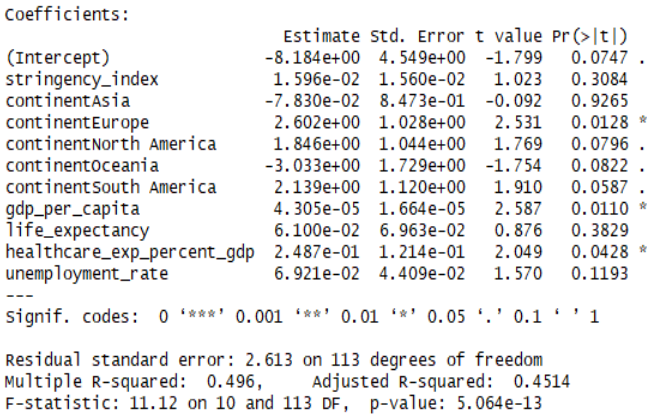
Fig. 6.2. Scatter plot of explanatory variables and deaths per million.

3.2. Regression Analysis

While the k-means method gives us an intuitive illustration of the relationship between factors and death rate, it cannot provide quantitative results. To get more detailed results, we move on to explain the severity of COVID-19's impact on a country by means of a regression model.

3.2.1. Multivariate linear regression

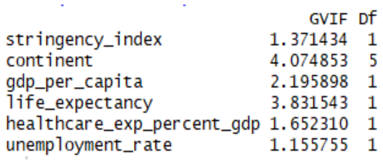
The predictors of linear regression model are continent (multi-level), stringency index, GDP per capita, health care as percentage of GDP, life expectancy and unemployment rate and they are chosen based on common sense and expert judgement. The result of the linear regression model shows that adjusted R-squared is 45.14% but important variables such as healthcare expense of GDP and unemployment rate were not significant, which implies that multicollinearity might be present. Fig 7.2 shows the output of variance inflation index and the result indicates some moderate multicollinearity and model overfitting become a concern given the small number of observations in the dataset. In addition, Fig. 7.3 shows that the model residuals are not normally distributed and there are only 124 data points in the dataset. Therefore, the p-value and confidence interval derived from the model might have issues. Hence, even though the linear regression model does not exhibit serial correlation (Ljung-Box test) and heteroskedasticity (Breusch-Pagan test), we decide to conduct automatic feature selection and use methods such as Lasso regression in the subsequent sections of the regression analysis for the next step.



```
Coefficients:
                             Estimate Std. Error t value Pr(>|t|)
(Intercept)                -8.184e+00  4.549e+00  -1.799   0.0747 .
stringency_index            1.596e-02  1.560e-02   1.023   0.3084
continentAsia              -7.830e-02  8.473e-01  -0.092   0.9265
continentEurope             2.602e+00  1.028e+00   2.531   0.0128 *
continentNorth America      1.846e+00  1.044e+00   1.769   0.0796 .
continentOceania           -3.033e+00  1.729e+00  -1.754   0.0822 .
continentSouth America      2.139e+00  1.120e+00   1.910   0.0587 .
gdp_per_capita              4.305e-05  1.664e-05   2.587   0.0110 *
life_expectancy             6.100e-02  6.963e-02   0.876   0.3829
healthcare_exp_percent_gdp  2.487e-01  1.214e-01   2.049   0.0428 *
unemployment_rate           6.921e-02  4.409e-02   1.570   0.1193
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 2.613 on 113 degrees of freedom
Multiple R-squared:  0.496,     Adjusted R-squared:  0.4514
F-statistic: 11.12 on 10 and 113 DF,  p-value: 5.064e-13
```

Fig. 7.1 Results from multivariate linear regression.



```
                              GVIF  Df
stringency_index          1.371434  1
continent                 4.074853  5
gdp_per_capita            2.195898  1
life_expectancy           3.831543  1
healthcare_exp_percent_gdp 1.652310  1
unemployment_rate         1.155755  1
```

Fig. 7.2 Results from model variance inflation index.

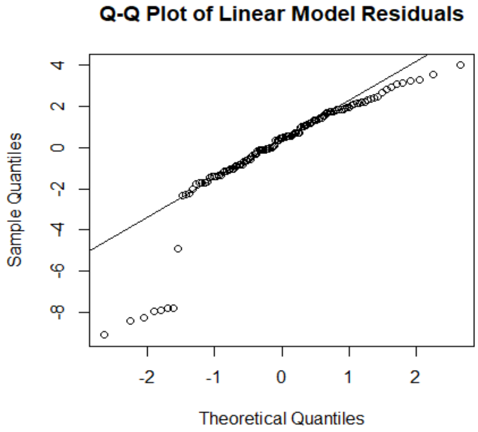


Fig. 7.3 Results from residual Q-Q plot.

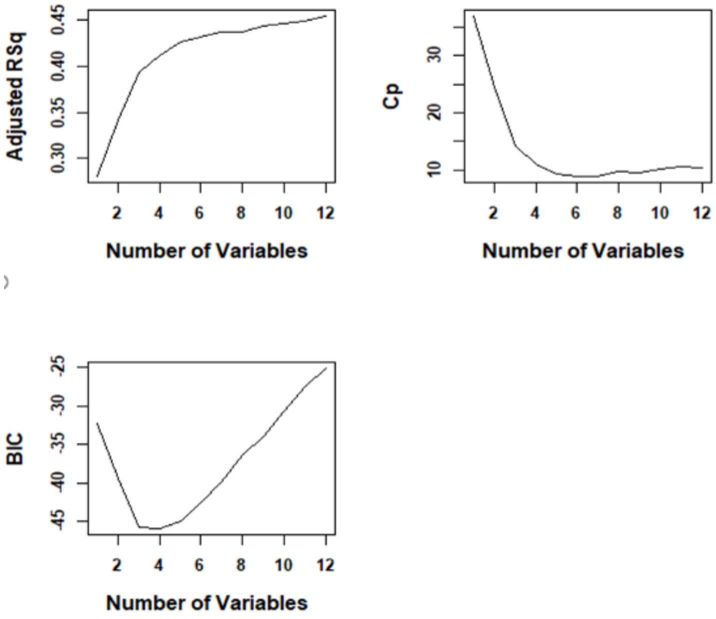


Fig. 8. Forward stepwise regression.

3.2.2. Forward stepwise selection

Due to the fact that our cleaned and transformed dataset only has 124 observations and there are 11 features, model overfitting and multicollinearity are possible and should be avoided. We use automatic model selection techniques such as forward and backward stepwise regression. The chosen forward stepwise regression model specification is the one which maximizes adjusted R-squared (45.42%) or minimizes Mallow's Cp (8.84). We hope to arrive at a regression model that does an adequate job of fitting a country's total death rate.

We begin with forwards and backwards stepwise regression to perform variable selection and select a 6-variable model as a reasonable choice in terms of Mallow's Cp, BIC, and Adjusted R-squared.

Table 1. Results from Forward Stepwise Regression

	Coefficients
(Intercept)	-14.0467
Continent Asia	-1.5573
Continent Oceania	-5.1537
GDP_per_capita	**0.0000**
Life_expectancy	0.1774
Healthcare_exp_percent_GDP	0.2622
Unemployement_rate	0.0663

3.2.3. Backward stepwise regression

We repeat the above process, this time using backward selection. The final model specification has 8 predictors, which performed the best in both Mallow's Cp (4.81) and Adjusted R-squared (46.12%). As one can see, there is a slight difference in terms of predictors selected between backward and forward regression.

Table 2. Results from Backward Stepwise Regression

	Coefficients
(Intercept)	-3.1234
Continent Europe	2.919
Continent North America	2.407
Continent Oceania	-2.786
Continent South America	2.7630
Population	1.9776E-09
GDP_per_capita	5.48009E-05
Healthcare_exp_percent_GDP	0.2428
Unemplyment_rate	0.0739

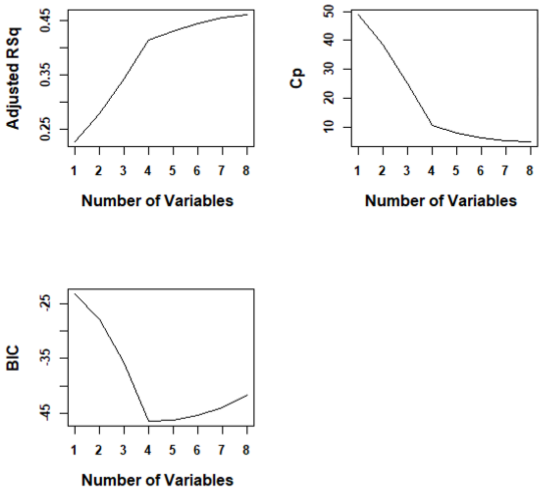


Fig. 9. Results from backward stepwise regression.

3.2.4. Lasso selection

Variable selection can also be made via LASSO regression, through selecting a shrinkage penalty parameter and then determining which variables to exclude by identifying which coefficients are coerced to zero [10]. Fig. 10 shows how the shrinkage penalty parameter lambda is selected using cross-validation against a grid of values based on minimum Mean-Squared Error (MSE). As the lambda increases, the coefficients of some predictors approach exactly zero. The LASSO identified a 6-predictors model specification, in which three of the predictors are dummy variables for the continent and another three predictors are GDP per capita, life expectancy and health care as percentage of GDP. The chosen model specification is a little bit different from that of forward and backward stepwise regression because different criteria is used to select the most optimal model (MSE in LASSO and adjusted R-squared and Mallow's Cp in the stepwise regression).

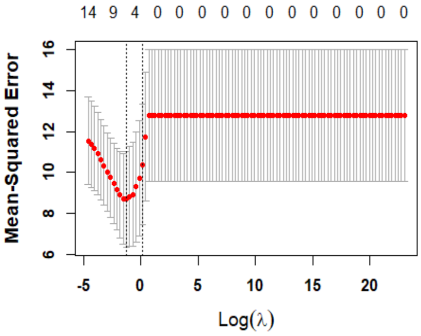


Fig. 10. Lambda selected by cross-validation.

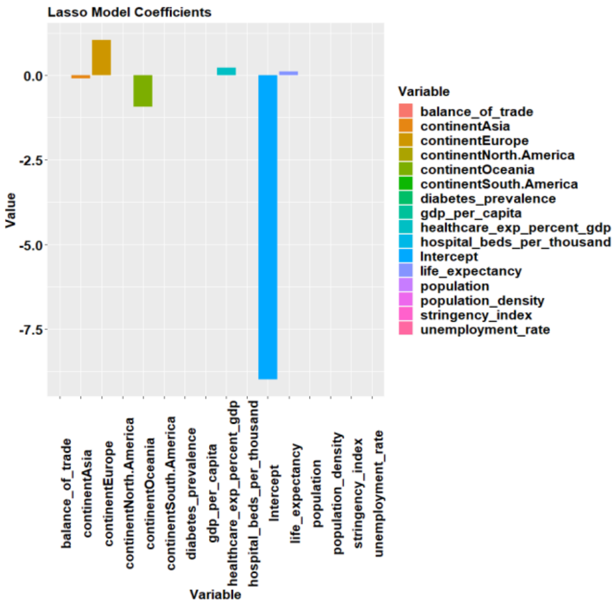


Fig. 11. Predictors selected by LASSO regression.

Fig. 11 shows the predictors selected by LASSO regression. These are the variables selected from forward stepwise regression (considering continent as one predictor). Variables such as continent, GDP per capita, life expectancy and health care expense as percentage of GDP.

3.2.5. Final predictors for linear regression model

After examining forward and backward stepwise regression, in addition to the LASSO regression, we arrive at a 6-predictor linear model. Based on Table 4, the predictors are: stringency index, continent, GDP per capita, diabetes prevalence, hospital beds per thousand, and all the funds spent on healthcare of GDP. The adjusted

R-squared for the model is around 0.44. The variables continent of Europe, Oceania, GDP per capita, as well as healthcare expenditure, are statistically significant, holding the other predictors fixed. We then test the performance of the 6-predictor model by using K-fold cross-validation with 10 folds to estimate the mean squared prediction error. We finally gain a prediction error of around 8.85.

Table 3. Summary of final Linear Regression Model

Residual standard error: 2.638 on 113 degrees of freedom	
Multiple R-squared: 0.4773	Adjusted R-squared: 0.4409
F-statistic: 13.12 on 8 and 113 DF	p-value: 5.06e-13

Table 4. Results from Linear Regression Model

	Estimate	Std.Error	t value	Pr (> | t |)
(Intercept)	-7.2828	4.4689	-1.630	0.1059
Continent Asia	-0.2591	0.8351	-0.310	0.7569
Continent Europe	2.3201	1.0282	2.257	0.0259
Continent North America	1.873	1.0191	1.838	0.0687
Continent Oceania	-3.527	1.7286	-2.040	0.0436
Continent South America	2.3548	1.1027	2.135	0.0348
GDP_per_capita	0.0000	0.0000	2.318	0.0222
Life_expectancy	0.0743	0.0670	1.063	0.2902
Healthcare_exp_per_GDP	0.2446	0.1190	2.056	0.0420

3.3. Random Forest

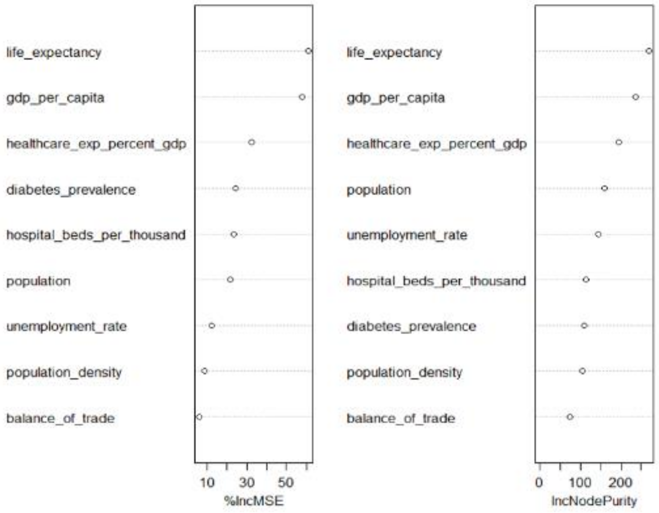


Fig. 12. Variable importance plots for a random forest model fitted to the full dataset.

A random forest model is fitted using stringency index, population, population density, median age, GDP per capita, diabetes prevalence, hospital beds per thousand, life expectancy, healthcare expenditures, unemployment, and balance of trade to predict total deaths per million from COVID-19. Five-fold cross-validation on 500 random sets of folds is used to optimize model parameters. For the final model, the number of trees to construct is 10,000, and we set the number of variables to try at each split to 2.

Five-fold cross-validation using optimized parameters results in a five-fold cross-validation RMSE of 3.551, which is lower than that we've found using the linear regression model. The most important variables, as shown in Fig. 12, are health-related variables (life expectancy, hospital beds per thousand, and healthcare expenditures) and one economic variable (GDP per capita). Life expectancy and GDP per capita increased the error the most when removed from the model, and these three also cause the cleanest node splits.

Overall, training and test accuracy highly vary across models. As shown in Fig. 13, the full dataset shows many countries with few deaths and a few countries with many deaths. Therefore, training accuracy for countries with few deaths is higher than training accuracy for countries with many deaths. A similar pattern can be seen in one case study of prediction accuracy (shown in Fig. 14) – prediction accuracy is much higher for countries with few deaths compared to countries with many deaths.

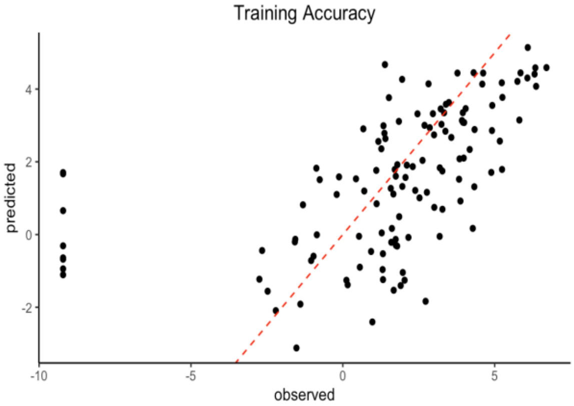


Fig. 13. Random forest model train accuracy for the full dataset (RMSE= 2.833). The red line is the identity, x=y, and represents perfect agreement between observed and predicted values.

Due to the small size of the data set and heterogeneity of number of deaths, prediction RMSE tends to be highly dependent on the exact cases used for fitting and prediction (shown in Figure. 15). Five-fold cross-validation RMSE values show variability across different fold sets, and within fold sets, RMSE varies widely. For some fold sets (Sets 6, 8, and 9) the largest and smallest RMSE values differ significantly. Other fold sets (Sets 2 and 3) show much smaller ranges. The variability in model fits and predictive capabilities emphasizes that a larger dataset with more balanced representation of cases across the full range of total death values would improve fit and prediction for the random forest model.

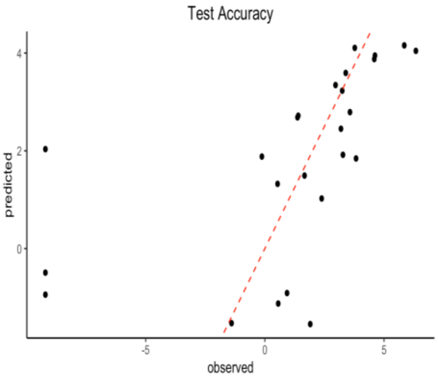


Fig. 14. Random forest model test accuracy for one random test set (RMSE= 3.551). Red line is the identity, x=y, and represents perfect agreement between observed and predicted values.

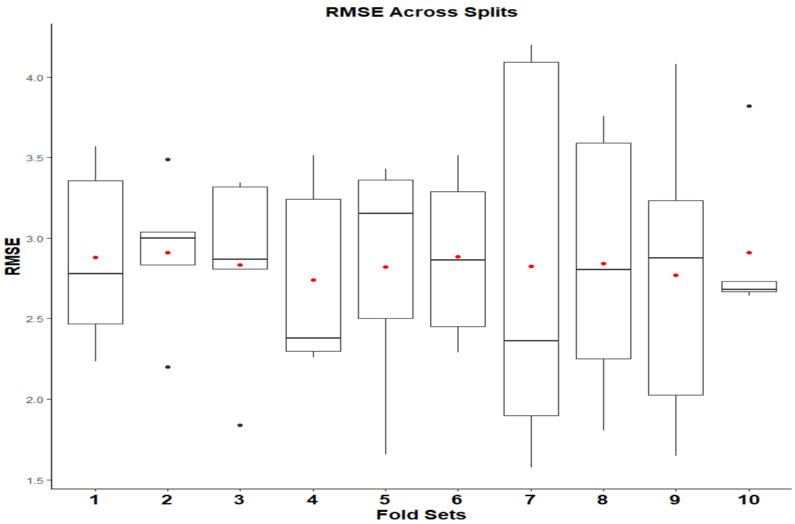


Fig. 15. Consistency in test error across different sets of random train/test folds. Red points are 5-fold cross-validation RMSE values for each fold set, and boxes show RMSE values for individual folds within a set.

3.4. Gradient Boost

A gradient boost model is constructed to predict "log of total deaths per million" using all predictors except for "total deaths per million". Gradient boost is a powerful method in which a sequence of regression trees is built based on the residuals of previous trees and predictions gradually move toward the true values driven by optimized loss function in each tree. Using the caret package in R and setting the method to "GBM", we obtain the variance importance of gradient boost method (see Fig. 16 below).

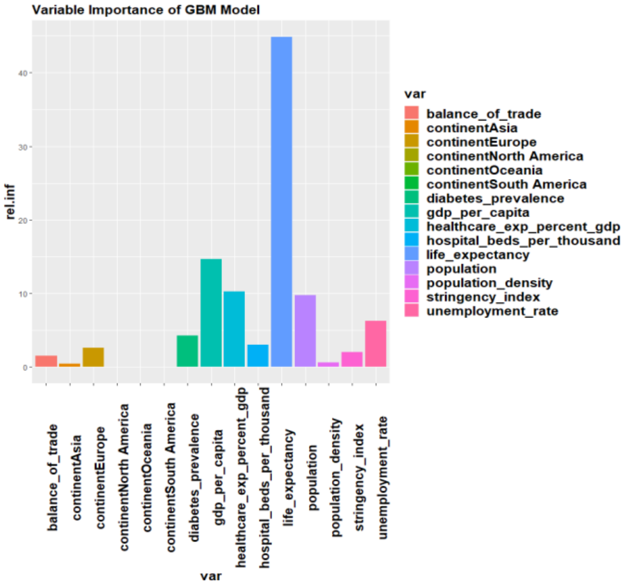


Fig. 16. Variance importance chart of gradient boost model

Fig. 16 showed that "life expectancy", "Healthcare expense as percentage of GDP" and "GDP per capita" are the top 3 most influential factors in determining "log of total deaths per millions".

Boosting algorithm in general is known to reduce model bias by adjusting sample weights based on the residuals from previous steps. Although model overfitting could be a potential issue, 10-fold cross validation with 10 repeats is implemented to address model overfitting. Table 5 shows the performance metrics of 10-

fold cross validations using different combinations of hyper-parameters. The most optimal model based on minimized error metrics have an RMSE and MAE of 2.8174 and 2.11, respectively. Based on the optimal model specification selected in Table. 5, prediction on test set and its error metrics are also calculated, as shown in Table 6. Table 6 shows RMSE and MAE results are similar between training and test set, indicating that the gradient boost model has no model overfitting and the error rates are better than those of other models.

Table 5. Performance Metrics with Combination of Hyper-Parameters

Shrinkage	Interaction.depth	N.Minobsinnode	N.Trees	RMSE	MAE
0.1	1	10	50	2.8174	2.1100
0.1	2	10	50	2.8346	2.1279
0.1	3	10	50	2.8638	2.1553
0.1	1	10	100	2.9265	2.2316
0.1	2	10	100	3.0029	2.2853
0.1	3	10	100	2.9887	2.2840
0.1	1	10	150	3.0137	2.3100
0.1	2	10	150	3.1014	2.3659
0.1	3	10	150	3.0737	2.3508

Table 6. Train vs. Test

Dataset	RMSE	MAE
Training	2.8174	2.1100
Testing	2.8885	1.9453

4. Conclusion

In the current study, statistical learning methods were applied to understand differences in the severity level of COVID-19 among countries. We selected the coronavirus death rate (deaths per one-million people) as the measurement variable of severity and ran a statistical analysis against several variables of interest. Data of 120 countries were collected. The aim of the study is to understand differences in the severity level of COVID-19 between countries using statistical methods. Our findings show that the healthcare expenditure, unemployment rate, life expectancy, and stringency index are somewhat close to a normal distribution. We note that the continents of Africa, Asia, and Oceania have countries with relatively low death rates caused by COVID-19, while Europe has a number of countries with a death rate of over 400 per million. We also note that predictions are relatively unreliable for those countries with a total death rate of zero. To alleviate this, it would be beneficial for future analysis to take into account the reliability of the death rate estimates, as some countries may deny to expose the true data to the public. Some of these countries may also have a death rate of zero due to unique circumstances, such as having little connection with other countries.

After modeling the cleaned and transforming data using various modeling techniques with cross validation, we reach a conclusion that Gradient Boost Machine (GBM) is the most optimal model in terms of minimum RMSE and MAE; also, the GBM model does not exhibit high model variance given the comparable RMSE and MAE between training and testing datasets. In addition, the algorithm selected the most influential predictors such as life expectancy (41.26%), healthcare expense as a percentage of GDP (18.06%) and GDP_per_capita (13.58%), which made intuitive sense in predicting total deaths per million people. Lastly, we believe that countries should aim to increase the health care spending as percentage of government budget and take a long-term approach on economic development to gradually increase GDP per capita and boost citizens' life expectancy in addition to implementing short-term healthcare measures such as social distancing, washing hands and mask wearing in order to flatten the death rate of potential future pandemics.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

Wen Yin conduct the research, ChenChen Pan collect and analyze the data, Wen Yin wrote the paper. Nanyi Deng and Dong Ji edit the paper; all authors had approved the final version

References

[1] Carrillo-Larco, R. M., & Castillo-Cara, M. (2020). Using country-level variables to classify countries according to the number of confirmed COVID-19 cases: An unsupervised machine learning approach. *Welcome Open Research*, *5(56)*, 56.

[2] Michael, O. C. (2020). COVID-19: A visual data science analysis and review. Retrieved July 21, 2020, from the TIBCO Blog https://www.tibco.com/blog/2020/03/18/covid-19-a-visual-data-science-analysis-and-revie/

[3] Cohen, J. P., Morrison, P., Dao, L., Roth, K., Duong, T. Q., & Ghassemi, M. (2020). COVID-19 image data collection: Prospective predictions are the future. Retrieved August 21,2020, from https://arxiv.org/abs/2006.11988

[4] Matthew, B., Benjamin, J. C., Zulma, M. C., Linh, D., & Neil, M. (2020). Early insights from statistical and mathematical modeling of key epidemiologic parameters of COVID-19. Retrieved Sep. 20, 2020, from the WHO COVID-19 Modelling Parameters Group https://wwwnc.cdc.gov/eid/article/26/11/20-1074_article

[5] Nussbaumer-Streit, B., Mayr, V., Dobrescu, A., Chapman, A., Persad, E., Klerings, I., Wagner, G., Siebert, U., Ledinger, D., Zachariah, C., & Gartlehner, G. (2020). Quarantine alone or in combination with other public health measures to control COVID-19: A rapid review. *Cochrane Database of Systematic.*

[6] Liu, C., He, Y., Liu, L., Li, F., & Shi, Y. (2020). Children with COVID-19 behaving milder may challenge the public policies: A systematic review and meta-analysis. *BMC Pediatrics, 20(1), 410.*

[7] Punn, N. S., Sonbhadra, S. K., & Agarwal, S. (2020). COVID-19 epidemic analysis using machine learning and deep learning algorithms. Retrieved August 10, 2020, from https://www.medrxiv.org/content/10.1101/2020.04.08.20057679v2

[8] Lauer, S. A., Grantz, K. H., Bi, Q., Jones, F. K., Zheng, Q., Meredith, H. R., & Lessler, J., *et. al*. (2020). The incubation period of coronavirus disease 2019 (COVID-19) from publicly reported confirmed cases: Estimation and application. *Annals of Internal Medicine, 172(9), 577-582.*

[9] Daniel, E. L. P. (2020). A geroscience perspective on COVID-19 mortality. *The Journals of Gerontology: Series A, 75(9).*

[10] Boulos, M. N. K., & Geraghty, E. M. (2020). Geographical tracking and mapping of coronavirus disease COVID-19/severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2) epidemic and associated events around the world: How 21st century GIS technologies are supporting the global fight against outbreaks and epidemics.

Wen Yin was born in ZhenJiang, China on June 14, 1989. He received a BA in bioengineering from East China University of Science and Technology in Shanghai, China in 2011. In 2020, he earned a MA in statistics from Columbia University in New York. He is current pursing a MS in computer science from Columbia University in New York. He has worked as a graduate student research at Columbia University since 2019. Prior this role, he worked as a data analyst intern at Microsoft and a Trader as T3 trading group. He has a publication of "Predictive Modelling of U.S Housing Prices Reveals Key Indicators of Real Estate Prices and Economic Health" at International Conference on Computing and Data Science. His primary research interest is data management and data visualization. Mr. Yin has received the Beta Gamma Sigma Honor Society in business.

Chenchen Pan was born in ZhenJiang, China on February 18, 1992. She received a BA in economics from Peking University in Beijing, China in 2013. In 2015, she earned an MS in management science from Stanford University in Stanford, CA. She is current pursing a PhD in innovation and entrepreneurship at Stanford University. She has worked as a graduate student researcher at Stanford University since 2015. Prior to this role, she worked as a research intern for Microsoft, an investment banking intern at CICC, and an assessment assistant at China Development Bank. Ms. Pan has received the Teaching Excellence Award and the William Linville Memorial Fellowship from Stanford University and the President's Undergraduate Research Fellowship from the Peking University.

Nanyi Deng was born in Wenzhou, China on June 8, 1994. She is currently engaging in a master's degree in applied analytics at Columbia University in New York City, New York. Prior to this degree, she earned a bachelor's degree in psychology from University of Kansas in Lawrence, Kansas, in 2016.

She is currently serving as a graduate research assistant at Columbia Medical Center in New York City, New York. Previously, she served as a research assistant in Dr. Monica Biernat's Stereotype and Prejudice Social Psychology lab and in Dr. Kelsie Forbush's Center for the Advancement of Research on Eating Behaviors. Her primary research interest is applied clinical mental health.

Ms. Deng is a member of Psi Chi, the International Honor Society in Psychology. She won Women of Distinction from University of Kansas Emily Taylor Center for Women and Gender Equity as well as Clark Coan International Leadership Award from University of Kansas International Student Service.

Dong Ji was born in ZhenJiang, China on Nov. 10, 1989. He earned a master's degree in economics and strategy for business from Imperial College London (Business School) in the United Kingdom in 2017. Prior to this degree, he also achieved a bachelor's degree in finance from Durham University (Business School) in the United Kingdom and graduated with honors in 2015.

He currently works as a research analyst for Suzhou Trust Co. in Suzhou. Previously, he has worked at Blueprint Capital as an IPO Analyst and as a Credit Analyst Intern at United Overseas Bank. His primary research interest is analytics for applied economics.

Mr. Dong is a candidate of the Chartered Financial Analyst Institute.