Since January 2020 Elsevier has created a COVID-19 resource centre with free information in English and Mandarin on the novel coronavirus COVID-19. The COVID-19 resource centre is hosted on Elsevier Connect, the company's public news and information website.

Elsevier hereby grants permission to make all its COVID-19-related research that is available on the COVID-19 resource centre - including this research content - immediately available in PubMed Central and other publicly funded repositories, such as the WHO COVID database with rights for unrestricted research re-use and analyses in any form or by any means with acknowledgement of the original source. These permissions are granted for free by Elsevier for as long as the COVID-19 resource centre remains active.
Linear Regression Analysis to predict the number of deaths in India due to SARS-CoV-2 at 6 weeks from day 0 (100 cases - March 14th 2020)

Samit Ghosal a,*, Sumit Sengupta b, Milan Majumder c, Binayak Sinha d

a Consultant Endocrinologist. Nightingale Hospital, Kolkata, India
b Consultant Pulmonologist. AMRI Hospitals, Kolkata, India
c Milan Majumder, Independent Statistician, Pune, India
d Consultant Endocrinologist. AMRI Hospitals, Kolkata, India

ABSTRACT

Introduction: and Aims: No valid treatment or preventative strategy has evolved till date to counter the SARS CoV 2 (Novel Coronavirus) epidemic that originated in China in late 2019 and have since wrought havoc on millions across the world with illness, socioeconomic recession and death. This analysis was aimed at tracing a trend related to death counts expected at the 5th and 6th week of the COVID-19 in India.

Material and methods: Validated database was used to procure global and Indian data related to coronavirus and related outcomes. Multiple regression and linear regression analyses were used interchangeably. Since the week 6 death count data was not correlated significantly with any of the chosen inputs, an auto-regression technique was employed to improve the predictive ability of the regression model.

Results: A linear regression analysis predicted average week 5 death count to be 211 with a 95% CI: 1.31 – 2.60. Similarly, week 6 death count, in spite of a strong correlation with input variables, did not pass the test of statistical significance. Using auto-regression technique and using week 5 death count as input the linear regression model predicted week 6 death count in India to be 467, while keeping at the back of our mind the risk of over-estimation by most of the risk-based models.

Conclusion: According to our analysis, if situation continue in present state; projected death rate (n) is 211 and 467 at the end of the 5th and 6th week from now, respectively.

© 2020 Diabetes India. Published by Elsevier Ltd. All rights reserved.

1. Introduction

The pandemic of COVID-19 (Coronavirus disease 2019) caused by SARS-CoV-2 (severe acute respiratory syndrome coronavirus 2) has created a havoc on the human civilization. Since, its appearance in the city of Wuhan (Hebei district) in China, it has been a relentless march of new cases and deaths [1]. What makes it more scary is the novel strain of the virus and the unknowns associated with it [2]. The present strategy has been to prevent its spread by social isolation and a scientific overdrive to manufacture newer rapid diagnostic kits as well as medications [3–5]. Coronavirus belongs to a family of RNA viruses within the virus family Coronaviridae, order Nidovirales [6]. Coronaviruses are divided into three groups depending on the antigenic spikes produced by different protein structures of the virus (spike, membrane & nucleocapsid) [7]. The SARS coronavirus falls under group 2.

The ability of this family of viruses to readily undergo genetic recombination not only within same group, but also between group, makes them readily susceptible to natural selection and changing its nature of virulence [8]. The most striking feature however, is its ability to freely cross from one species to another. HCoV 229 E belongs to the group 1 of the coronaviruses family thought to be responsible for the epidemic of common cold [7]. Transmission from bats to humans is thought to be the initial transmission process for HCoV 229 E, which had happened within the last two centuries. However, the two dramatic events-
SARS-CoV (originated from bats and got transmitted from civet cats) & MERS-CoV (originated from bats and got transmitted from camels) in 2003 and 2012 respectively brough our focus back on the coronavirus family [8,9]. The present coronavirus pandemic is the result of changes in the receptor binding domain of the spike protein component via natural selection (?in human host/?in animal vector) resulting in its increased affinity for the ACE2 receptor site [10]. At present we have more than 450000 individuals affected with Cov-2 resulting in more than 12000 deaths worldwide [11]. In India, as the present day statistics holds, we have around 718 confirmed cases with 13 deaths [12]. Several countries including India have gone into a state of lockdown in order to prevent spread of this deadly virus. With new rapid-diagnostic kits coming in and trials with potentially helpful drugs underway, we need a better understanding of the disease process and what it holds for the near future. With all the Cov-2 related data available through reliable sources, we choose to assimilate the available data on total infection rates, total deaths, case fatality rates (CFR), recovery numbers from across the globe and create a predictive analysis on what we can expect in India in the coming weeks.

India has entered week 4. This analysis was aimed at tracing a trend related to death counts expected at the 5th and 6th week of the COVID-19 in India.

2. Materials and methods

Global data was collected from the WHO COVID-19 situation report and the Indian data was updated from the website covid19india.org. Data was collected in a CSV file and uploaded in Jupyter notebook and analysed with the Python 3.8.2 software. As a re-validation process and for simplicity of understanding the data was also analysed using excel with XL-STAT statistical software.

Inputs: Total number of infected cases, active cases, recovery numbers.

Outputs: Total deaths and case fatality rates (CFR)

In order to get a good predictive value data was analysed for the top 15 infected countries with India the 16th country.

2.1. Pre-analysis phase

There was one missing data (NA) in the dataset, which was the recovery numbers from the US. In view of the heterogeneity of data and significant outliers data imputation with mean was ruled out.

As a recovery strategy a correlation analysis was conducted (leaving out the US data) using python and a strong $r = 0.99$ ($P < 0.001$) was found between total number of infected cases and recovery. Utilising this robust association and the formula generated from linear regression ($Y \text{ [Recovery cases | USA]} = b_0 + b_1 \times \text{ [Total cases | USA]}$, with $b_0 = -781.05$ and $b_1 = 0.869$), the missing value (1117) was derived. The analysis was conducted thereafter (Table 1).

3. Results

Analysis for week 5 death number prediction:

- Step1: A correlation analysis was performed to ascertain the presence of and thereafter the strength of association between the output (week 5 death count) and the inputs from week 4. There was a strong correlation between week 5 deaths and all the input variables (Table 2).
Step 2: A multivariate regression analysis ascertained the most important input parameters which would be used to build the model for the 5th week death prediction in India. The model came out to have a very strong predictive capacity ($r = 0.99$, $R^2 = 0.98$, adjusted $R^2 = 0.97$). However, the P-value was significant only for the 4th week death input parameter (Table 3).

Step 3: A simple regression analysis was subsequently done to predict the death counts from the strongest input variable (Table 4). The model was robust with $r = 0.87$, $R^2 = 0.77$ & adjusted $R^2 = 0.75$, $P < 0.001$, 95% CI: 1.31–2.60. Based on the upper limit (maximum) & the lower limit (minimum) of the confidence intervals, the minimum, maximum and average death counts for week 5 was computed-211 (Table 4). Hence the week 5 death counts for India was predicted based on the available data from the top 15 infected countries.

### 3.1. Death number prediction for week 6

Step 1: Correlation study was conducted to ascertain the relationship between the output (week 6 death counts) and the input variables from week 4. A good correlation was observed with all the input variables.

Step 3: A simple regression analysis was subsequently done to predict the death counts from the strongest input variable (Table 4). The model was robust with $r = 0.87$, $R^2 = 0.77$ & adjusted $R^2 = 0.75$, $P < 0.001$, 95% CI: 1.31–2.60. Based on the upper limit (maximum) & the lower limit (minimum) of the confidence intervals, the minimum, maximum and average death counts for week 5 was computed-211 (Table 4). Hence the week 5 death counts for India was predicted based on the available data from the top 15 infected countries.

### Table 3

Results from the multiple regression analysis conducted with 5th week death count as output and all the 4th week parameters as input. * Goodness of fit (Adjusted R Square) shows the high predictive power of the model in this multivariate linear regression. However, most of predictors fail to show their significance of contribution in model except Week 4 death.

### SUMMARY OUTPUT

| Regression Statistics | Multiple R | R Square | Adjusted R Square | Standard Error | Observations |
|-----------------------|------------|----------|-------------------|----------------|--------------|
| 0.990327848           | 0.980749246 | 0.970054383 | 234.1358914      | 15             |

| ANOVA                  | df         | SS         | MS      | F        | Significance F |
|------------------------|------------|------------|---------|----------|----------------|
| Regression             | 5          | 25135569.86 | 5027113.97 | 91.70283143 | 1.92537E-07    |
| Residual               | 9          | 491376.5407 | 54819.61564 |           |                |
| Total                  | 14         | 25628946.4  |          |          |                |

| Coefficients          | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
|-----------------------|----------------|--------|---------|-----------|-----------|-------------|-------------|
| Intercept             | 84.42512812    | 0.734083868 | 0.48158332 | −175.739842 | 344.590099 | −175.739842 | 344.590099   |
| Total cases           | −0.069994422   | −0.320843804 | 0.755651571 | −0.56350055 | 0.42351705 | −0.56350055 | 0.42351705   |
| Active cases          | 0.121357776    | 0.782303013 | 0.454129589 | −0.22994623 | 0.473061974 | −0.22994623 | 0.473061974  |
| Recovery cases        | −0.095715087   | −0.8728080868 | 0.40545473 | −0.343792945 | 0.152362771 | −0.343792945 | 0.152362771  |
| Week 4 deaths         | 3.49748606     | 4.568598845 | 0.56350055 | 1.905112961 | 0.80895916 | 1.905112961 | 0.80895916   |
| CFR                   | 33.51344079    | 46.33770995 | 0.723243355 | 0.48789902 | −71.30974168 | −71.30974168 | 138.3366233  |

### Table 4

The maximum, minimum and average predicted death counts for week 6 based on the equation of the linear regression model.

| In 95% Confidence Interval | Intercept and Co-efficient | 5th Week predicted death |
|----------------------------|-----------------------------|----------------------------|
| Mean point of estimation   | b0 | 191.644 | 211 |
|                           | b1 | 1.957 | |
| Lower point of estimation  | b0 | −229.314 | −216 |
|                           | b1 | 1.312 | |
| Upper point of estimation  | b0 | 612.602 | 639 |
|                           | b1 | 2.602 | |

### Table 5

Multiple regression analysis with week 6 death counts as input and all the 4th week variables as input. * Goodness of fit (Adjusted R Square) shows the high predictive power of the model in this multivariate linear regression. However, all the predictors fail to show their significance of contribution in model.

### SUMMARY OUTPUT

| Regression Statistics | Multiple R | R Square | Adjusted R Square | Standard Error | Observations |
|-----------------------|------------|----------|-------------------|----------------|--------------|
| 0.955366444           | 0.912725042 | 0.864238954 | 687.4807679      | 15             |

| ANOVA                  | df         | SS         | MS      | F        | Significance F |
|------------------------|------------|------------|---------|----------|----------------|
| Regression             | 5          | 4485030.68 | 8857006.936 | 18.82447281 | 0.000158714    |
| Residual               | 9          | 4253668.256 | 472629.8062 |         |                |
| Total                  | 14         | 48738702.93 |          |          |                |

| Coefficients          | Standard Error | t Stat | P-value | Lower 95% | Upper 95% | Lower 95.0% | Upper 95.0% |
|-----------------------|----------------|--------|---------|-----------|-----------|-------------|-------------|
| Intercept             | 115.6866622    | 337.6903173 | 0.342582112 | 0.39777934 | −648.2219079 | 879.5952322 |
| Total cases           | 0.174235894    | 0.640563717 | 0.27004002 | 0.791755917 | −1.274819906 | 1.623291695 |
| Active cases          | 0.159410883    | 0.456247295 | 0.349395788 | 0.73482115 | −0.872692204 | 1.19151397 |
| Recovery cases        | −0.392375137   | −1.218550947 | 0.253991509 | −1.120759296 | 0.336042685 | |
| Week 4 deaths         | 3.061382182    | 2.066876933 | 1.481163263 | 0.172701393 | −1.614218276 | 7.736982641 |
| CFR                   | 101.840825     | 136.0589538 | 0.748504966 | 0.4732618 | −205.9459343 | 409.6275393 |
measures being taken? What can we expect, if we allow the present trend to continue and mimic the exponential growth experienced by China and our western counterparts? Our analysis predicts a jump from approximately 10 deaths at week 4—211 at week 5 and then 467 by week 6.

We speculate the need for urgent interventions (which are being taken as of now), to prevent this drastic and sharp rise in death rates which indirectly also indicates an increase in infection rate.

4.2. Limitations of this analysis

The main limitation of this analysis was that it takes most input data into consideration without taking into account the logistic actions being taken or not taken during the process. However, the end of weeks results are highly indicative of both the virus-related natural trajectory as well as the local government’s reactions.

Secondly, limiting our analysis to the top 15 most infected countries could lead to an over-estimation of the outcomes. However, faced with a catastrophe of such magnitude, it is worth over-estimating rather than under-estimating.

4.3. Strength of the study

In spite of all the limitations the biggest strength of this study was very high adjusted $R^2$ found in all the predictive models. In addition there was cross-validation with two different software practically ruling out any error creeping in from one mode of analysis.

5. Conclusion

According to our analysis, if situation continue in present state; projected death rate ($n$) is 211 and 467 at the end of the 5th and 6th week from now, respectively. Keeping these projected mortality data in mind, current measured for containment of COVID-19 must be strengthened or supplemented.

Funding

None.

Declaration of competing interest

None to declare.

References

[1] Remuzzi A, Remuzzi G. COVID-19 and Italy: what next? The Lancet. 2020 [Online] Available at: https://www.thelancet.com/action/showPdf?pii=S0140-6736%2820%2930627-9. Accessed at: 26th March 2020.

[2] Cascella M, Rajnik M, Cuomo A, Dulebohn SC, Napoli RD. Features, evaluation and treatment coronavirus (COVID-19). 2020 [Online] Available on, https://www.ncbi.nlm.nih.gov/books/NBK554776/. Accessed at: 26th March 2020.

[3] Armitage R, Nellums LB. COVID-19 and the consequences of isolating the elderly. The Lancet 2020 [Online] Available at: https://www.thelancet.com/action/showPdf?pii=S2215-2995%2820%2930601-7. Accessed at: 26th March 2020.

[4] Coronavirus disease (COVID-19) technical guidance: laboratory testing for 2019-nCoV in humans. World Heath Organisation; 2020 [Online] Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/technical-guidance/laboratory-guidance. Accessed on: 26th March 2020.

[5] Information for clinicians on therapeutic options for COVID-19 patients. Centres for disease control and prevention. 2020 [Online] Available at: https://www.cdc.gov/coronavirus/2019-ncov/hcp/therapeutic-options.html. Accessed on: 26th March 2020.

[6] The species Severe acute respiratory syndrome-related coronavirus: classifying...
2019-nCoV and naming it SARS-CoV-2. Coronaviridae Study Group of the International Committee on Taxonomy of Viruses. Nature Microbiology 2020;5:536–44.

[7] Peiris JSM. Coronaviruses. In: Medical microbiology, eighteenth ed. 2012 Elsevier Ltd; 2012 [Chapter 35].

[8] Brian DA, Baric RS. Coronavirus genome structure and replication. Curr Top Microbiol Immunol 2005;287:1–30.

[9] Burrell JC, Howard CR, Murphy FA. Coronaviruses. In: Fenner and white’s medical virology, fifth ed. © 2016 Elsevier Inc; 2017.

[10] Anderson KG, Rambaut A, Lipkin WI, Holmes EC, Gary RF. The proximal origin of SARS-CoV-2 [Online] Available at: https://www.nature.com/articles/s41591-020-0820-9.pdf. Accessed on: 22nd March 2020.

[11] Coronavirus disease (COVID-19) situation reports. World Health Organization. [Online] Available at: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports [Accessed on: 26th March 2020].

[12] INDIA COVID-19 TRACKER. 2020 [Online] Available at: https://www.covid19india.org/. Accessed on: 26th March 2020.