Diabetes & Metabolic Syndrome: Clinical Research & Reviews 14 (2020) 311–315

Contents lists available at ScienceDirect

Diabetes & Metabolic Syndrome: Clinical Research & Reviews

journal homepage: www.elsevier.com/locate/dsx

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

ARTICLE INFO

Article history:
Received 26 March 2020
Received in revised form
27 March 2020
Accepted 27 March 2020

Keywords:
India
Coronavirus
Death rates
Correlation
Regression

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 and467 at the end of the 5th and 6th week from now, respectively.

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-

* Corresponding author.
E-mail address: ramdasghosal@gmail.com (S. Ghosal).

https://doi.org/10.1016/j.dsx.2020.03.017
1871-4021/

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 lock-down 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.

The aim was to identify the top 15 countries i.e. those most heavily affected and hence could contribute to a substantial quantity of robust data, and compute a predictive model for India. We thought this was of paramount importance, since it would help understanding as well as planning for the future course of action.

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 [Recovery cases | USA] = b0 + b1* [Total cases | USA], with b0 = -781.05 and b1 = 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).

Table 1
Raw data including all coronavirus-related variables for week 1 and the total death outputs for week 5 through 9, including the imputed value.

Countries	Total cases	Active cases	Recovery cases	Week 4 deaths	CFR	Week 5 deaths
China	74185	57805	65112	2004	2.701	2715
Italy	21157	17750	12207	1441	6.811	4825
Spain	5232	4906	3097	133	2.542	1093
Iran	11364	7321	9919	514	4.523	1433
France	3661	3570	482	79	2.158	450
UK	798	769	495	11	1.378	177
Netherlands	804	792	134	10	1.244	106
Germany	3675	3621	3130	8	0.218	68
Belgium	559	555	139	3	0.537	37
Switzerland	1139	1124	303	11	0.966	56
South Korea	7979	7198	7294.42	67	0.840	94
Austria	504	497	431	1	0.198	6
Brazil	151	150	151	0	0.000	11
Indonesia	69	60	38	4	5.797	32
USA	2183	2126	1117	48	2.199	255
India	606	554	42	10	1.650	

Table 2
Correlation analysis determining the relationship between week 5 deaths and all the input variables.

Total cases	*Active cases*	*Recovery cases*	*Week 4 deaths*	*CFR*	*Week 5 deaths*	
Total cases	1					
Active cases	0.99904861	1				
Recovery cases	0.994753954	0.991471532	1			
Week 4 deaths	0.922623423	0.924523558	0.883996909	1		
CFR	0.268208625	0.266050424	0.209369645	0.511501668	1	
Week 5 deaths	0.635636081	0.644536597	0.561097633	0.876211223	0.696402315	1

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	0.990327848							
R Square	0.980749246							
Adjusted R Square	0.970054383							
Standard Error	234.1358914							
Observations	15							
ANOVA								
	df	*SS*	*MS*	*F*	*Significance F*			
Regression	5	25135569.86	5027113.972	91.70283143	1.92537E-07			
Residual	9	493376.5407	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	115.0074695	0.734083868	0.481583332	−175.7398428	344.590099	−175.7398428	344.590099
Total cases	−0.069994422	0.218157313	−0.320843804	0.755653571	−0.56350055	0.423511705	−0.56350055	0.423511705
Active cases	0.121557776	0.155384517	0.782303013	0.454125958	−0.229946423	0.473061974	−0.229946423	0.473061974
Recovery cases	−0.095715087	0.109664289	−0.872800868	0.405455473	−0.343792945	0.152362771	−0.343792945	0.152362771
Week 4 deaths	3.49748606	0.70391798	4.968598845	0.000771382	1.905112961	5.08985916	1.905112961	5.08985916
CFR	33.51344079	46.33770995	0.723243355	0.487899902	−71.30974168	138.3366233	−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	

- 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.

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	0.955366444					
R Square	0.912725042					
Adjusted R Square	0.864238954					
Standard Error	687.4807679					
Observations	15					
ANOVA						
	Df	*SS*	*MS*	*F*	*Significance F*	
Regression	5	44485034.68	8897006.936	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%*
Intercept	115.6866622	337.6903173	0.342582112	0.739777934	−648.2219079	879.5952322
Total cases	0.174235894	0.640563717	0.272004002	0.791755917	−1.274819906	1.623291695
Active cases	0.159410883	0.456247295	0.349395788	0.734828115	−0.872692204	1.19151397
Recovery cases	−0.392375137	0.322001422	−1.218550947	0.253991509	−1.12079296	0.336042685
Week 4 deaths	3.061382182	2.066876933	1.481163263	0.172701393	−1.614218276	7.736982641
CFR	101.8408025	136.0589538	0.748504966	0.4732618	−205.9459343	409.6275393

Table 6
Prediction for 6th week death count in India based on the auto-regression analysis technique.

In 95% Confidence Interval	Intercept and Co-efficient		6th Week predicted death
Mean point of estimation	b0	184.33	467
	b1	1.34	
Lower point of estimation	b0	−119.77	120
	b1	1.14	
Upper point of estimation	b0	488.44	813
	b1	1.54	

- Step 2: Multiple regression analysis was done to ascertain the strength of association between the input variables and the output, including ruling out issues related to multi-collinearity. Once again the model created was very robust with r = 0.95, R^2 = 0.91 & adjusted R^2 = 0.86 (Table 5). However, the P-value for significance was not evident for any of the input variables.
- Step 3: **Auto-Regression technique-** The 5th week death count data was incorporated as the input variable, in view of the fact that this end-point was significantly associated with the week 4 death count. A separate correlation analysis was performed between week 5 and week 6 death counts and a very robust association (r = 0.97) was found justifying its inclusion as the input variable.
- Step 4: Using week 5 death count as input a simple linear regression was performed to create a model predicting the week 6 outcomes. The model was robust (r = 0.96, R2 = 0.94, adjusted R2 = 0.94) and statistically significant (P-value <0.001, 95% CI: 1.13–1.54).
- Step 5: From the regression model formula the minimum, maximum, and average death count was estimated (Table 6). The average predicted death count for India was estimated to be 467.

4. Discussion

India is in the 4th week of the coronavirus pandemic. What lies ahead for India is the crucial stage, week 5–6 where effective preventive measures can prevent a potential catastrophe, which countries like China, Italy and the United States of America are experiencing, with an exponential growth of both infection as well as deaths. Exponential progression in the number of infected cases have occurred from the 4th week onwards, in the above mentioned countries [11].

At the point of going to the press, there are approximately 4,89,853 confirmed cases of coronavirus infection worldwide including 22,152 deaths [11]. Luckily, at 4th week, India has escaped the brunt of the disease with figures hovering around 693 confirmed infections and 13 deaths. It is the next couple of weeks which holds the key to the direction the virus takes or doesn't take if we take adequate preventive steps.

India has already taken strong measures including complete lockdown of both its internal and external borders as well as social isolation.

4.1. How does this analysis help?

Assessing the trends of the top 15 most infected countries a predictive model was created for India assuming that the same trend would follow. In other words can we justify the drastic 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=S2468-2667%2820%2930061-X. Accessed on: 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 syndromerelated 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] reportCoronavirus disease (COVID-2019) situation reports. World Health Organization. [Online] Available at:: https://www.who.int/emergencies/diseases/novel-coronavirus-2019/situation-reports/[Accessed on: 26nd March 2020].
[12] INDIA COVID-19 TRACKER. 2020 [Online] Available at: https://www.covid19india.org/. Accessed on: 26nd March 2020.