Risk and Survival Analysis from COVID Outbreak Data : Lessons from India

Prasad Bankar PRASADBANKAR33@GMAIL.COM
Subhasis Panda SUBHASISPANDA94@GMAIL.COM
Vaibhav Anand VAIBHAV.HK.ANAND@GMAIL.COM
Vineet Kumar* VNTKUMAR8@GMAIL.COM
Indian Institute of Technology Kharagpur, WB, India, 721302

Abstract

The present analysis is an attempt to provide data-backed evidence around mortality due to COVID-19 in Indian context. We provide a description of the prevailing COVID-19 conditions in India by means of succinct visualisation via a dynamic dashboard and cluster analysis. Building upon this, we performed survival analysis on COVID-19 patients from the state of Karnataka, stratifying the data on the basis of age and gender. The findings of the same have been reported in this paper. To our knowledge, this is the *largest* retrospective cohort-based survival analysis in Indian context.

Keywords: Survival Analysis, Covid-19, Covid in India, Clustering Analysis

1. Introduction

Actionable insights to understand the spread of COVID-19 is the need of the hour. Modelling and Forecasting COVID-19 is very difficult given the unreasonable modelling assumption, lack of good quality data Ioannidis et al. (2020). Keeping this goal in mind, we developed a dashboard to enable policymakers to extract a variety of information from publicly available data on COVID-19. The dashboard provides risk summaries and enables users to look at data from different perspectives, arriving at actionable insights. The COVID–19 situation may be described through a number of characteristics such as the rate of occurrence of new cases, the breakup of confirmed cases by severity, rate of recovery, number of active cases, rate of testing, rate of detection of new cases through tests, and death rate.

When measured and presented appropriately, each of these characteristics communicates a different aspect of the overall scenario. Sometimes, a combination of a few characteristics may also be used. These characteristics or their combinations are called dimensions. We attempted to cover all the dimensions and a few others not described above. We also included some composite dimensions obtained by combining values of two or more directly observed dimensions. Although we intended to cover many dimensions, a few like the breakup of confirmed cases by severity could not be covered as data were not available. We observed that the daily recovery varies widely and is low in many states and cities. Daily recovery is even zero on several occasions and frequently varies by over five times or more on subsequent days. The process of recovery comprises of interdependent activities like sample collection, testing, retesting, and approval of final release. It is well-known that built-in

* All authors contributed equally, Names in alphabetical order

delays in such processes are often over 50% of the total time and standard techniques exist to reduce the cycle time drastically.

We did a cohort selection with age and gender as variates separately. Using Kaplan-Meier estimate and Log Rank Test, we tried to model the recovery and fatality rate. We also fitted a cox proportional hazard to quantify the effect of variates.

The paper is organized as follows — first, we describe our approach for understanding the holistic picture of COVID-19 spread, testing and risk in India. We performed the elementary descriptive analysis. We also provide the case based clustering of Indian states. Subsequently, we performed survival analysis by building our cohort of patients.

2. Covid Risk Analysis

When COVID-19 struck, the governments across the globe started enforcing the policy of lockdown. The Government of India followed suit Lancet (2020) and a strict lockdown was imposed on 25th March, 2020. However, the lockdown was just a measure to reduce the speed and extent of the spread of COVID-19. Recognising that the policymakers could benefit from a tool that facilitated visualisation along different dimensions, we built an interactive dashboard, using real-time data, using the Panel library in Python. We are thankful to COVID-19 India Org Data Operations Group (2020) for providing the public api for the data. The dashboard [1] aimed at supporting the "unlock" strategy design and facilitate intervention monitoring. It provided the following functionalities:

- Macro-level view: The dashboard allows the policymakers to have a birds-eye view of the country as well as individual States. This was achieved by means of a risk summary table and geographical heat-maps.
- Spread Assessment Metrics:
  - Average confirmed cases per day: Increasing trend of occurrence of new cases is likely to indicate that the virus is spreading.
  - Active cases: Increasing trend of active cases indicates the possibility of increased stress on resources at present or in the near future.
- Risk Assessment Metrics:
  - Avg. Confirmed Cases to Avg. Recoveries in Consecutive Weeks, Traffic Intensity (a term from queuing theory): It acted as a leading indicator of stress on resources, with values > 1 implying faster arrival of new cases compared to recovery.
  - Points plot of Tests in Each Week and % Positive in each week: Higher proportion of positive test (Fig: 1) results on a larger test population (possibly less targeting) are anomalous and need investigation. There are at least three possibilities: — (i) A rapid increase in the rate of infection (ii) Carrying out tests in new (hitherto untested) areas with a high but unknown rate of infection (iii) Improper testing leading to many incorrectly positive results

1. https://covid-isical.tech/

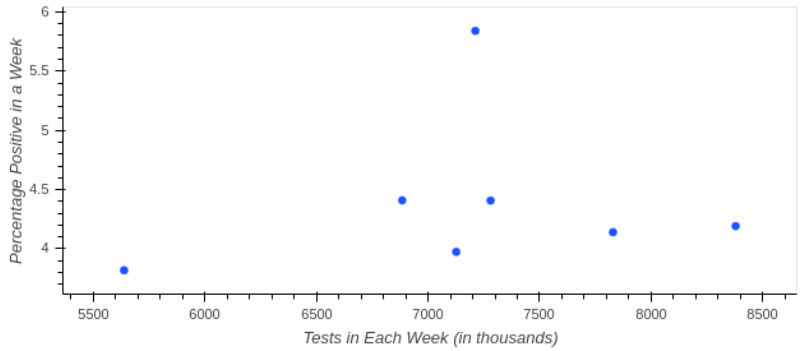

Figure 1: Assessing adequacy of testing in India

We also provided a tool for the comparative assessment of Indian states using colour-coded heatmaps showing the top States in India w.r.t confirmed, recovered, and deceased cases, respectively.

2.1. Cluster Analysis

In any nation, when COVID-19 outbreak occurred. there were some states/provinces/regions which were more affected as compared to other states/provinces/regions. The motivation behind performing cluster analysis is identifying such underlying structures within the COVID-19 cases. We performed euclidean measure based vanilla k-means clustering on the top 20 most affected states. We chose the value of k by iteratively experimenting and checking the standard silhouette score and elbow metric.

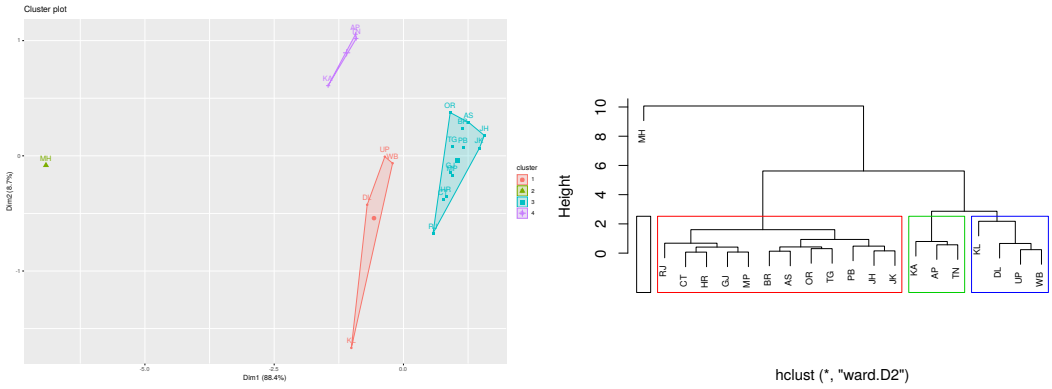

Figure 2: Cluster analysis (K-means and AGNES) based on Active Cases

We also performed the hierarchical clustering on the cases using Ward distance (chosen among other competing metric based on Agglomerative coefficient). Structures thus formed have physical interpretation and validates the ground reports. For instance, MH is the worst affected state, while {KL, DL, UP, WB} are among the second worst-hit states. Such

clustering analysis, performed across the last 7-8 months, gives a better picture of how a lesser affected state shifts from one cluster to another as cases increase and vice-versa.

With Mumbai, Pune, Thane, Nashik and Nagpur, Maharashtra has 5 of the top 10 districts with highest COVID19 burden, whereas other states like Delhi, Karnataka, UP and West Bengal have a single urban centre with very high disease burden. Inadequate healthcare infrastructure with high load and a good disease reporting mechanism also makes Maharashtra the worst hit state in our cluster analysis findings.

3. Covid Survival Analysis

In this section, we will discuss our approach of performing survival analysis Pocock et al. (2002) on the COVID-19 cases of India. There has been very minimal work on COVID-19 survival analysis data. One such work in the Indian context has been done by Mishra et al. (2020). However, their cohort size is very limited, with < 500 patients. Hence, findings are not very conclusive owing to the small sample size. In the present study, we illustrate survival analysis results on a reasonably large cohort of patients (26,714 patients) from the Indian State of Karnataka.

3.1. Dataset Description & Methods

Using the data of testing date and the date of discharge (either due to recovery or death) from hospitals for individual patients in Karnataka, the probabilities of a person being in infected condition as the days progress are calculated using Kaplan-Meier estimator of the survival function. Our particular work aims to study the difference in recovery time among gender and various age groups. **Data and Code** associated with our experiment is available at https://github.com/vntkumar8/covid-survival

Cohort Selection We downloaded the patient-specific data containing patient id, age, gender, admit date and recover/death date from Siva Athreya and Mishra (2020). These data were sourced from official government medical bulletins and recorded into a spreadsheet.

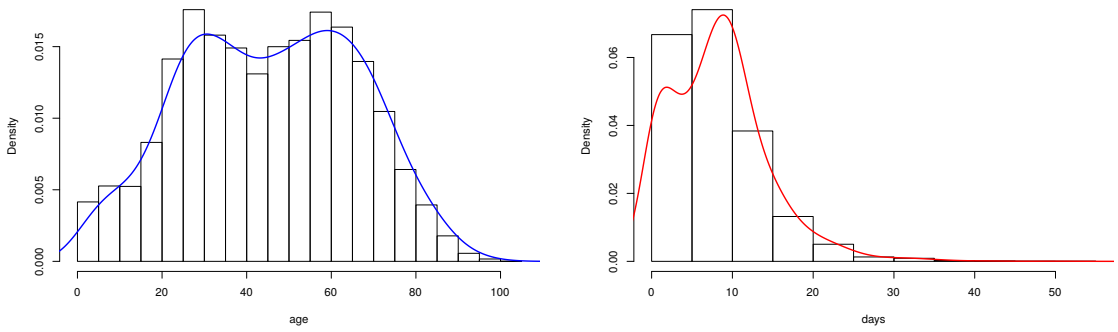

Figure 3: Stay & Age distribution of patients

After performing routine data cleaning like making all inputs lowercase, removing whitespaces etc., we calculated the number of days to recovery/death for each patient by subtracting the admit date with the discharge date.

We selected all such patients who were admitted in Karnataka as reported by government bulletins. We excluded four patients — one patient was a transgender, one patient had unusually high recovery time (84 days), and two patients did not have gender information. Our final cohort size was 26,741 patients. In our cohort, the mean age is 45 years (median 46) and the mean/median time to recovery is 8 days. Age and time to event distribution is shown in Fig: 3.

3.2. Gender & Age Stratified KM Estimate

First, to perform the Kaplan-Meier survival analysis Kaplan and Meier (1958), we stratified our cohort gender-wise – male and female. The estimator of the survival function $S(t)$ (the probability that life is longer than t) is given by:

$$\widehat{S}(t) = \prod_{i:\ t_i \leq t} \left(1 - \frac{d_i}{n_i}\right)$$

with t_i a time when at least one event happened, d_i the number of events (e.g., deaths) that happened at time t_i, and n_i the individuals known to have survived (have not yet had an event or been censored) up to time t_i.

The KM estimates for gender stratification are shown in Fig: 4(a). Median survival probability time for Male and Female is 14 and 21 days, respectively.

Also, to better understand the effect of age on the recovery time of patients, we stratified our entire patient cohort into three groups – young (age $<$ 18 years old), adult (age between 18 and 60 years) and old (age $>$60 years). Following the standard techniques, we computed the KM estimates for the three age strata. The same is illustrated in Fig: 4(b).

There is a large difference in median survival probability time for Adults and Old as 4 and 33 days, respectively.

Log Rank Test We performed the standard non-parametric Log Rank Test Bland and Altman (2004) to statistically compare the difference between the survival probability of various stratum. The null-hypothesis of test is – there is no difference between the two strata. The test accounts for the difference in the treatment factors between the two groups. We divide the data according to the levels of the significant prognostic/treatment factors and form a stratum for each level.

For gender-based KM curve Kaplan and Meier (1958) (Figure 4(a)) we found that $p <$ 0.0001, which is way less than our α significance level of 5% hence gender is statistically significant for the survival time of patients. Also, as evident from age stratifed KM curve (Figure: 4(b)), **log-rank test** also validates that age is significant predictor of mortality, which was anyways intuitive (p-value $<$ 0.0001).

4. Discussion & Conclusion

Karnataka is one of the badly hit Indian states in terms of the number of COVID-19 cases. Further, the gender and age data was available for deceased, recovered, and active patients

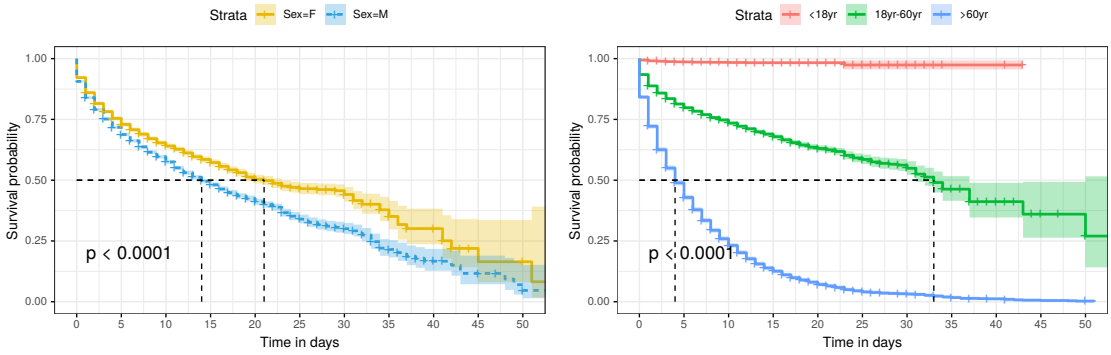

Figure 4: KM estimates for (a) Gender & (b) Age stratification separately

via detailed state bulletins. Our cohort characteristics and fundamental KM analysis results are shown in Table: 1. Our cohort had 26,741 patients in total. 65% of them were male. Among all patients, 57% of them recovered.

Among the recovered patients, 62% were male. The median survival time was 10 days, and the inter-quartile range (IQR) was 8 to 13 days. In the young cohort (< 18years old) of 2295 patients, 98% of them recovered. Among the old cohort (>60 years old) of 8066 patients, mere 14% recovered. We also fitted a KM estimate on Age groups and Gender combined and same is shown in Fig: 5. It is evident from the figure that for older group (> 60 yr) there is not much difference in median survival time across gender but for adult cohort (18 – 60 yr) median survival is varying by 4 days across gender startum respectively.

Cox Proportional Hazard We also fitted a semiparametric cox proportional hazard (cox-ph) model to *quantify* the effect of age on mortality. Cox-ph model gives Hazard Ratio, which is interpreted as ratio of hazards between two groups (Male & Female) at any particular point in time. HR estimate turns out to be 1.26 with $[1.21, 1.31]$ as 95% Confidence Intervals. (p-value < 0.001). This implies **Male** gender is associated with 1.26 **times increased risk** or decreased survival. Male population in our cohort is more vulnerable, reason can be traced to men having higher burden of comorbidity, exposure to virus due to relatively higher contact with riskier environment (workplace, commute, shopping etc.). The same finding has been reported by Wei et al. (2020). Also evident from analysis, with increasing age the probability of survival from COVID-19 decreases.

	Age Category	Infected	Dead	% Dead	*rmean	*se(rmean)	median	0.95LCL	0.95UCL
Sex=F	<18 yr	1,072	14	1.31%	52.26	0.20	-	-	-
Sex=F	18 yr – 60 yr	5,665	1,430	25.24%	33.24	1.23	37.00	32.00	-
Sex=F	>60 yr	2,553	2,087	81.75%	8.00	0.27	5.00	4.00	5.00
Sex=M	<18 yr	1,223	23	1.88%	51.54	0.55	-	-	-
Sex=M	18 yr – 60 yr	10,715	3,141	29.31%	28.93	0.96	33.00	28.00	37.00
Sex=M	>60 yr	5,513	4,815	87.34%	6.73	0.12	4.00	4.00	4.00
	All	26,741	11,510	43.04%	20.87	0.38	16.00	16.00	17.00

Table 1: Cohort Characteristics and Mean/Median Survival Estimate

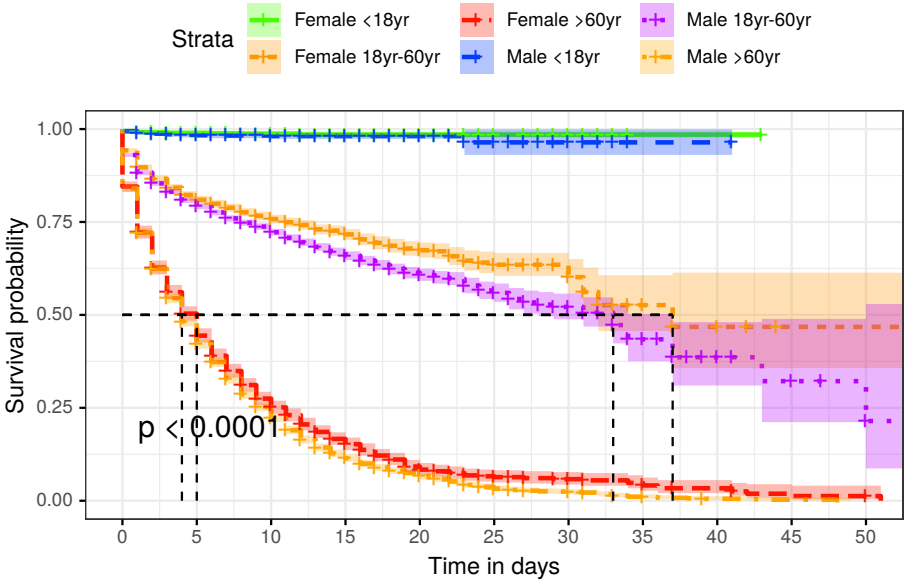

Figure 5: Gender & Age Stratified KM estimate

The actionable insights/policy decisions were based on the metrics covered in the dashboard and lie in the domain of the end-users (policymakers in our case). The analysis was communicated via visual aids since interpretability and ease of understanding were paramount. It was beyond a mere exploratory data analysis since the graphs were based on metrics that would be useful at policy-level. For instance, if the traffic intensity was found to rise sharply in a State, it was a sign that resources might get overwhelmed soon. Similarly, graphing the proportion of people tested positive with the number of total tests in a State allowed the users to get a measure of the testing efficacy.

In our knowledge this is largest retrospective-cohort based survival analysis Clark et al. (2003) study on COVID-19 outbreak in Indian context Tapnikar et al. (2020). Our Risk analysis & reporting (dashboarding) is also unique. We worked with limited resources which prevented us from delving deeper. Since very limited variables about the patient could be obtained from the government bulletins, scope of predictive modeling is limited. Comorbidity data was unavailable for recovered patients. If more information about symptoms, comorbidities, and other subtle differences can be found, detailed analysis and better insights can be carried out. Despite these limitations, this study attempts to obtain meaningful information from publicly available data for use in public health actions.

5. Acknowledgement

The authors acknowledge the hard work of COVID-19 India Org Data Operations Group (2020) for collating, tracking & reporting the COVID-19 data and making it available for public use. The authors would also like to thank the anonymous reviewers for their constructive criticism and suggestions, which helped in substantially improving the technical and editorial quality of the paper.

References

J Martin Bland and Douglas G Altman. The logrank test. *Bmj*, 328(7447):1073, 2004.

Taane G Clark, Michael J Bradburn, Sharon B Love, and Douglas G Altman. Survival analysis part i: basic concepts and first analyses. *British journal of cancer*, 89(2):232–238, 2003.

COVID-19 India Org Data Operations Group. Covid-19 india tracker. Accessed on 2020-25-11 from https://api.covid19india.org/, 2020.

John PA Ioannidis, Sally Cripps, and Martin A Tanner. Forecasting for covid-19 has failed. *International journal of forecasting*, 2020.

Edward L Kaplan and Paul Meier. Nonparametric estimation from incomplete observations. *Journal of the American statistical association*, 53(282):457–481, 1958.

The Lancet. India under covid-19 lockdown. *Lancet (London, England)*, 395(10233):1315, 2020.

Vinayak Mishra, Ajit Deo Burma, Sumit Kumar Das, Mohana Balan Parivallal, Senthil Amudhan, Girish N Rao, et al. Covid-19-hospitalized patients in karnataka: Survival and stay characteristics. *Indian Journal of Public Health*, 64(6):221, 2020.

Stuart J Pocock, Tim C Clayton, and Douglas G Altman. Survival plots of time-to-event outcomes in clinical trials: good practice and pitfalls. *The Lancet*, 359(9318):1686–1689, 2002.

Nitya Gadhiwala Siva Athreya and Abhiti Mishra. Covid-19 india-timeline an understanding across states and union territories., 2020. Ongoing Study at http://www.isibang.ac.in/~athreya/incovid19.

Lata Tapnikar, Sneha Patil, and Jaydeep Nyse. Interpreting kaplan meier's survival curve in covid-19 patients: a systematic review. *International Journal Of Community Medicine And Public Health*, 8(1):424–433, 2020. ISSN 2394-6040. doi: 10.18203/2394-6040.ijcmph20205733. URL https://www.ijcmph.com/index.php/ijcmph/article/view/7379.

Xiyi Wei, Yu-Tian Xiao, Jian Wang, Rui Chen, Wei Zhang, Yue Yang, Daojun Lv, Chao Qin, Di Gu, Bo Zhang, et al. Sex differences in severity and mortality among patients with covid-19: evidence from pooled literature analysis and insights from integrated bioinformatic analysis. *arXiv preprint arXiv:2003.13547*, 2020.