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

Prasad Bankar*, Subhasis Panda*, Vaibhav Anand*, Vineet Kumar*
Indian Institute of Technology
Kharagpur, WB, India, 721302
{prasadbankar33,subhasispanda94,vaibhav.hk.anand,vntkumar8}gmail.com

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.

Introduction

The Novel Coronavirus is spreading fast. While India is attempting to unlock the economy fully, the number of new cases per day is still 40,000. Actionable insights to contain the spread 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, Cripps, and Tanner 2020).

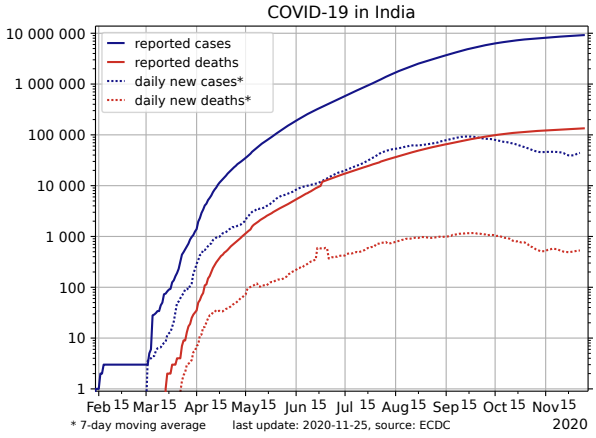

Figure 1: Daily COVID-19 cases in India (Log Scale)

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 COVID19 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 delays in such processes are often over 50% of the total time and standard techniques exist to reduce the cycle time drastically.

We also used survival analysis to compute the infection fatality rate (IFR) in the two critical states of India, viz. Tamil Nadu and Karnataka. 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 tried to analyse the effect of comorbidity on the COVID-19 fatality rate. We performed a cluster analysis with the different districts of Karnataka to understand the disease spread.

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.

*Equal Contribution, Names appear in alphabetical order.

Covid Risk Analysis

When COVID-19 struck, the governments across the globe started enforcing the policy of lockdown. The Government of India followed suit 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, via 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 / 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.

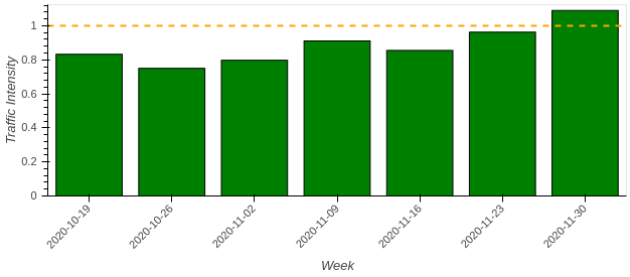

Figure 2: Traffic Intensity across last 7 weeks

  - Points plot of Tests in Each Week and % Positive in each week: Higher proportion positive test results on a larger test population (possibly less targeting) are anomalous and need investigation. There are at least three possibilities:
    1. A rapid increase in the rate of infection
    2. Carrying out tests in new (hitherto untested) areas with a high but unknown rate of infection
    3. Improper testing leading to many incorrectly positive results

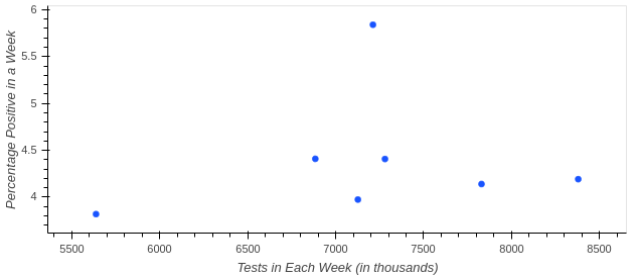

Figure 3: Assessing adequacy of testing in India

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.

Cluster Analysis

In any nation, when a 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. 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 meaning and interpretation. 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, could give a better picture of how a less affected state shifts from one cluster to another as cases increase and vice-versa.

Covid Survival Analysis

In this section, we will discuss our approach of performing survival analysis (Pocock, Clayton, and Altman 2002) on the COVID-19 cases of India. There has been 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.

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

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

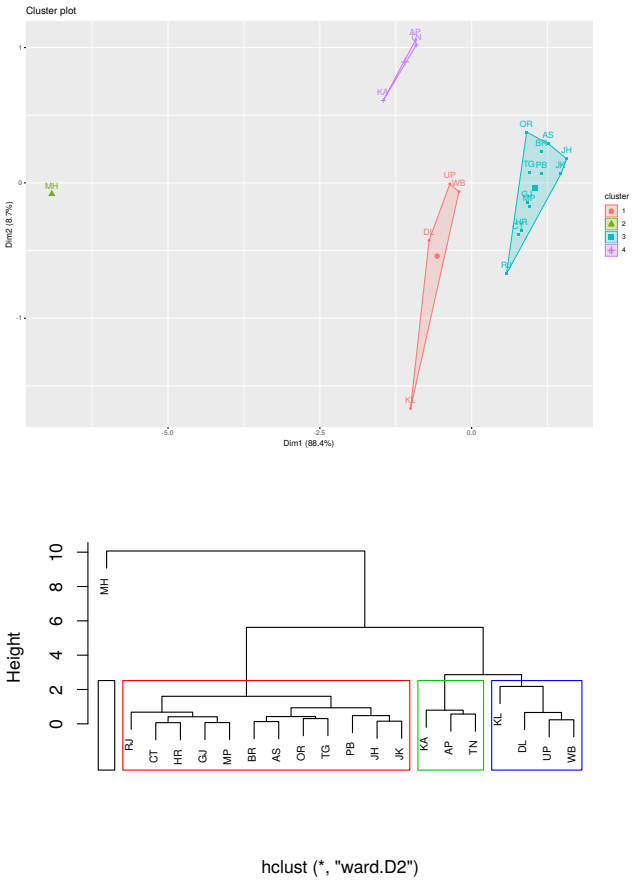

Figure 4: Clustering analysis (K-means and AGNES)

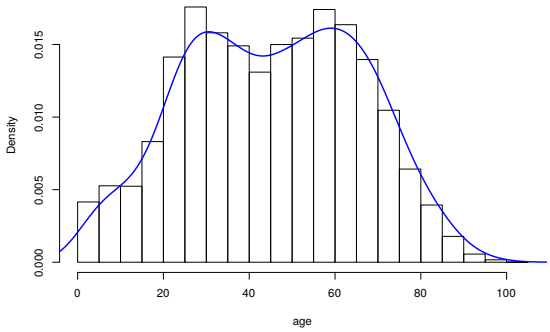

Figure 5: Age distribution of patients

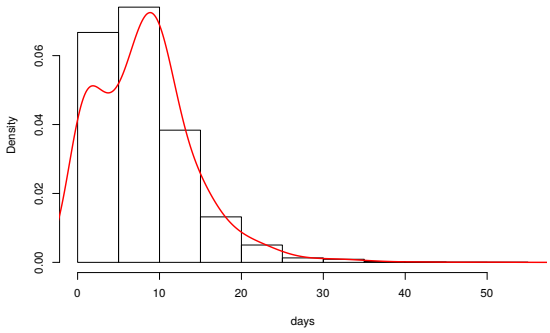

Figure 6: Stay distribution of patients

Kaplan-Meier estimator of the survival function. Our particular work aims to study the difference in recovery time among gender and various age groups.

Cohort Selection We downloaded the patient-specific data containing patient id, age, gender, admit date and cure/deceased date from Siva Athreya and Mishra (2020). These data were sourced from official government medical bulletins and recorded neatly into a spreadsheet. 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. We selected only such patients who had a definite outcome – either recovered or deceased. We excluded the active cases. Our final cohort size was 26,741 patients. Age distribution of patients is shown in Fig: 5.

In our selected cohort, the mean age is 37 years (median 36) and the mean time to recovery is 8 days. Similarly, the stay distribution is shown in Fig: 6.

Gender Stratified KM Estimate

To perform the Kaplan-Meier survival estimate (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: 7. Median survival probability time for Male and Female is 14 and 21 days, respectively.

Log Rank Test We performed the well-known non-parametric Log Rank Test (Bland and Altman 2004) to statistically compare the difference between the survival probability of male and female strata. 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

	Infected	Recovered	Median Survival Time (days)
Total infected	26741	15231	10 (8-13)
Male	17451	9472	10 (8-13)
Female	9290	5759	10 (8-13)
Total infected age $\leq$ 18	2295	2258	10 (8-14)
Total infected age $\leq$ 18 Male	1223	1200	10 (8-14)
Total infected age $\leq$ 18 Female	1072	1058	10 (8-14)
Total infected 18<age <60	16380	11809	10 (8-13)
Total infected 18<age <60 Male	10715	7574	10 (8-13)
Total infected 18<age <60 Female	5665	4235	10 (8-13)
Total infected age $\geq$ 60	8066	1164	10 (8-13)
Total infected age $\geq$ 60 Male	5513	698	10 (8-13)
Total infected age $\geq$ 60 Female	2553	466	10 (8-13)

Table 1: Median recovery time of infected population

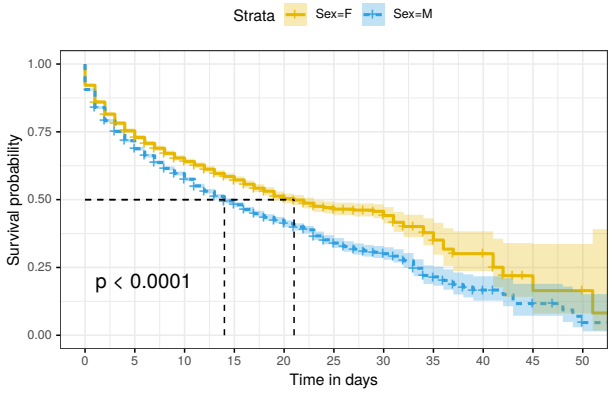

Figure 7: Gender stratified KM estimate

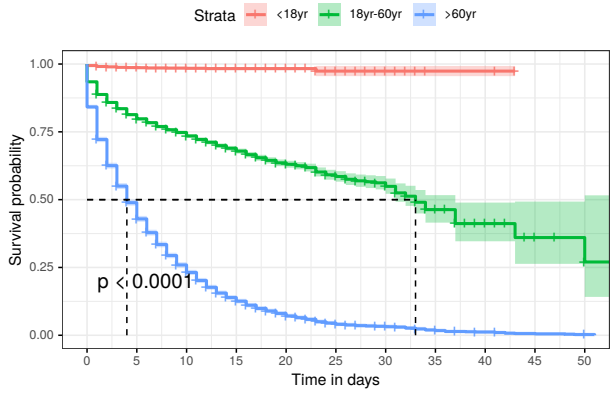

Figure 8: Age stratified survival curve estimation

significant prognostic/treatment factors and form a stratum for each level.

For gender-based KM curve (Kaplan and Meier 1958) (Figure 7) 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.

Age Stratified KM Estimate

To better understand the effect of age on the recovery time of patients, we stratified our 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: 8.

There was a large difference in median survival probability time for Adults and Old as 4 and 33 days, respectively. As expected, the **log-rank test** also validated the visual evidence with p-value < 0.0001.

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 COVID-19 patients via detailed state bulletins. Hence, we performed survival analysis on the available data. Our cohort had 26,741 patients (refer Table: 1). 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 ($>$18 years old) of 2295 patients, 98% of them recovered. Among the old cohort ($>$60 years old) of 8066 patients, mere 14% recovered. In our knowledge this is largest retrospectivecohort based survival analysis (Clark et al. 2003) study on COVID-19 outbreak in Indian context. The data had many problems like for recovered patients co-morbidity is not recorded in the bulletins. Had this been the case, we could have fitted a Cox-proportional hazard model to assess the impact of co-morbidity in prognosis. Still, this is an incremental contribution in understanding the many facets of COVID-19.

References

Bland, J. M.; and Altman, D. G. 2004. The logrank test. *Bmj* 328(7447): 1073.

Clark, T. G.; Bradburn, M. J.; Love, S. B.; and Altman, D. G. 2003. Survival analysis part I: basic concepts and first analyses. *British journal of cancer* 89(2): 232–238.

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

Ioannidis, J. P.; Cripps, S.; and Tanner, M. A. 2020. Forecasting for COVID-19 has failed. *International journal of forecasting* .

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

Mishra, V.; Burma, A. D.; Das, S. K.; Parivallal, M. B.; Amudhan, S.; Rao, G. N.; et al. 2020. COVID-19-Hospitalized Patients in Karnataka: Survival and Stay Characteristics. *Indian Journal of Public Health* 64(6): 221.

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

Siva Athreya, N. G.; and Mishra, A. 2020. COVID-19 India-Timeline an understanding across States and Union Territories. Ongoing Study at http://www.isibang.ac.in/~athreya/incovid19.