Coronavirus Disease (COVID-19) Dynamics: Age and Gender-based Analysis of Surveillance Variables

Malik Khizar Hayat [a], Ali Daud [b], Rabeeh Ayaz Abbasi [a], Tehmina Amjad [c], Xiuzhen Jenny Zhang [d]

[a] Department of Computer Science, Quaid-i-Azam University, Islamabad, Pakistan
[b] Department of Information Systems and Technology, College of Computer Science and Engineering, University of Jeddah, Jeddah, Saudi Arabia
[c] Department of Computer Science and Software Engineering, International Islamic University Islamabad, Islamabad, Pakistan
[d] School of Science (Computer Science and IT), RMIT University, Melbourne, Australia
khizerhayat92@gmail.com; alimsdb@gmail.com; rabbasi@qau.edu.pk; tehminaamjad@iiu.edu.pk; xiuzhen.zhang@rmit.edu.au
Corresponding author: Malik Khizar Hayat (email: khizerhayat92@gmail.com)

Abstract

COVID-19 emerged in Wuhan and is later declared as a pandemic by World Health Organization. Different-aged people have varying gender-wise immunity control properties that necessitates understanding COVID-19 impact on age and gender which does not exist, currently. In this paper, COVID-19 surveillance variables are extensively studied along with hospitalization, tests-performed, and recovery data. Dataset is curated from three sources; however, age and gender data belong to Belgium, particularly. Visualizations, frequencies, Pearson's and polyserial correlation, student's t-test, and Cramer's V are used for enhanced analysis. Results show higher mortality rate in males and need of more ventilators to combat severe symptoms.

Keywords: Coronavirus disease (COVID-19); Correlation and analysis; Age and Gender; Hospitalization; Tests performed; Recovery

1. Introduction

Influenza is one of the main so far epidemics known to the world. Due to its swift mutation and transmission rate, not only controlling its spread is difficult, but it also has a strong influence on the daily routine operations of society making it a challenge for health and development [1]. In the 21st century, the world has seen two major outbreaks in the human population. One in the form of Severe Acute Respiratory Syndrome Coronavirus (SARS-CoV) in 2003, and the other in the Middle East Respiratory Syndrome Coronavirus (MERS-CoV) in 2012. Over 8000 cases of SARS from 26 countries and 2200 cases of MERS from 27 countries were reported to the WHO [2]. Till the end of December 2020, several cases of unknown pneumonia were reported in Wuhan, Hubei, a Province of China. The novel COVID-19 was identified to be the cause of these cases by the Chinese government and WHO, which belongs to the same family of coronaviruses. By the end of January 2020, 7000 infectious cases of COVID-19 were reported from mainland China. When the countries including Thailand, Japan, United States of America, and South Korea also reported the infectious cases, WHO declares the COVID-19 outbreak as a Public Health Emergency of International Concern (PHEIC).

Currently, two hundred and ten countries across the globe are affected by the disease with almost 2,252,370 reported cases. To prevent the disease spread, the governments of the affected regions have taken serious measures to ensure social distancing. Intending to understand the disease spread, several researchers analyzed the COVID-19 surveillance variables and made predictions based on features like the reported cases, recovered cases, and

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

fatalities. Moreover, they recommended ensuring strong social distancing measures for disease prevention.

In a society, there exist different-aged people who may not have the same strong immunity power. It also varies gender to gender. However, except hospitalization data [3], there has been no study to identify the effects of the novel COVID-19 with the age and gender of the community. Besides, a number of reliable COVID-19 information sources exists including the mostly followed Worldometers as one of them. However, it does not correlate the date-wise, country, age and gender-based data with the surveillance variables.

The daily changing figures of COVID-19 are important for the authorities to take the righteous steps as per the varying situation. These phenomena have motivated us to bring the two important external variables under consideration. In this study, we analyze the data that is curated from multiple sources. However, the study focuses on the correlation of age and gender with the novel COVID-19 surveillance variables for Belgium, in particular. Along with the visualization, COVID-19 surveillance variable frequencies and Pearson's correlation, polyserial correlations, and Cramer's V are used for the analysis of variables.

Finally, we note that the extensive experimentation results signify the greater mortality trend in males. However, more reported cases belong to females. In terms of age, the people who are more than 45 years old are more susceptible to the novel COVID-19. To sum up, the current analysis provides efficient insights for better health implications for Belgium. In all, the contributions made in this paper are as follows:

- Data curation from multiple sources for better understanding the dynamics of COVID-19
- In-depth, date-wise age and gender analysis
- Analysis of the recovery and tests performed data
- Analysis of the hospitalization data
- Enriched the country-level COVID-19 dynamics

The rest of the paper is organized as follows: Section 2 provides a review of the related literature, Section 3 gives the details of methods and dataset, section 4 demonstrates the results and detailed discussions on the findings of the study, and section 5 finally concludes the study with some important future directions and recommendations.

2. Literature Review

With the purpose of understanding and preventing the novel COVID-19, substantial research is being carried out around the globe. Efforts have been made to know the impact of COVID-19 – economy, business decision making, crude oil reserves, social media, surveillance variables, prediction, including a comprehensive review [4]. However, the significant findings from the existing COVID-19 studies are presented next.

In order to identify policies to control viral transmission and prevent economic loss during the current Coronavirus epidemic, authors [1] studied the economic effect of epidemics such as SARS, H1N1, and Ebola. The outcomes of the study emphasize that authentic information can significantly help in fighting against the epidemic. Another economic study [2] of COVID-19 revealed the uncertainties that have disrupted the global trade and supply chain. The businesses are in dire need to make correct decisions at the correct time. The authors suggested using big data analytics tools for effective business decision making.

Oil is a major building block of any society, these days. COVID-19 has squeezed the oil demand resulting in the decline in oil prices, internationally. Authors made an effort [5] to

study the effect of Coronavirus and crude oil prices on the United States Economic Policy Uncertainty (EPU). It is observed that the global number of reported cases and the death rate has no significant effect on the United States EPU. However, the decrease in crude oil demand is causing higher uncertainty. The US has a business relation with China, the reason, the cases outside China has a positive impact on US EPU as well as the crude oil prices. Regarding the surveillance impact of COVID-19 in Arab countries, Qatar is at the top followed by Bahrain, Saudi Arabia, Egypt, and Iraq [6]. It is predicted that the infection rate in the Arab region is likely to increase in the near future.

Social networking sites play an important role for the customers, businesses and organizations by sharing information for making the right decisions [2]. However, inaccurate information can cause panic in the public, while underestimation can cause unawareness, hence a balance is highly required in information revelation. A massive dataset [7] from Twitter, Instagram, YouTube, Reddit and Gab is collected to analyze information diffusion regarding Coronavirus. Using mathematical formulation, the authors identified misinformation in huge volumes from various social media sources. In a nutshell, instead of surveillance variables, the fear, disgrace, and discrimination propagated by social media sources found to be the major reasons behind the COVID-19 economic impacts.

A day level forecasting is also performed for Coronavirus cases [8]. The authors performed a time series analysis on data that contains day level data of countries, cities, provinces, including the number of reported cases, confirmed cases, and the number of deaths available till that date. The experimentation revealed that countries which do not take disease prevention measures like quarantines, travel ban, or social distancing, are most likely creating an increased number of cases. The results show that the exponential increase is due to virus transmission, however, the number of tests performed can help in diagnosing the infectious cases.

In another related research, the researchers studied three mathematical models including the Logistic model, model and Gompertz model to the first fit and analyze the trends of SARS and then applied the results to study the situation of Coronavirus [9]. The obtained results are different with the use of varying features, and different for different regions. The results also show that the fitting model of logistic regression was best among the three models. The authors [10] studied the relationship between the number of reported infectious cases with the number of fatalities. It is observed that the rate of fatalities was 2% till January, however, January onwards, it was quadratic, not exponential. The same trend was prevalent at the time of publishing of the study and authors suggested that keeping this trend in mind the fair predictions can be performed for the upcoming 1 to 2 months.

A case study using Composite Monte-Carlo (CMC) method is performed for decision making in circumstances of high uncertainty during Coronavirus outbreak. CMC was further enhanced with the use of deep learning and fuzzy rule induction [11]. Experimentation was performed on data gathered by a Chinese government agency and it is observed that CMC in conjunction with Group of Optimized and Multisource Selection (GROOMS) methodology performs best deterministic forecasting and generated results were better than the results achieved by the traditional methods.

A simple mathematical time-dependent SIR model was proposed to predict the trends of Coronavirus [12], [13]. The authors designed various experiments to simulate the spread of Coronavirus based on infection rate and removal rate under different levels of anti-spread measures and medical care. In prediction results, the authors emphasized that serious attention is required towards the control measures in terms of medical services and the availability of more space for hospitalization to decrease the spread of the pandemic. In a similar study, SIR

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

is applied to predict the Coronavirus spread in China [14]. The study predicts the number of infected cases, prone cases, and removed cases with respect to time. The authors suggested that further research should be conducted to improve the outcomes of prediction with more complex mathematical models on updated data.

Moreover, using accurate simulations, a statistical model is built on the data of Hubei Province to study the process of virus transmission [15]. The results show that trend identified by the model for the number of confirmed infections, the number of cured people and the number of deaths matches the actual official data very closely. The authors have also studied the parameters like the incubation period and the average number of days of cure and found that these measures can control the impact of the disease.

Social distancing is proved to be an important tool to fight against COVID-19. A study [16] examined COVID-19 data from South Korea to find the strategies for preventing the disease spread. Data ranging from January 20, 2020, to February 26, 2020, is provided by the Korea Centers for Disease Control and Prevention. By calculating the reproduction rate, the authors suggested that social distancing strategies can be helpful to limit the spread of the disease. Authors [17], [18] have also analyzed the Chinese provincial data - Hubei, Zhejiang, Guangdong, Hunan and Henan, and the city data - 50 cities of China ranging from January 29, 2020, to February 9, 2020. The authors used bootstrap and computational methods for diagnosis rate identification. It is observed that the diagnosis rate of 15 cities in Hubei Province was lower than that of 35 cities outside Hubei. However, in cases of provinces, human to human transmission is found. It is also observed that the measures of isolation and quarantine were effective in controlling the spread of disease.

3. Age and Gender-based Analysis of Surveillance Variables

3.1 Dataset

In this study, the data is curated from multiple sources. The current study is primarily focused on correlation analysis of age and gender with the novel COVID-19 in Belgium. In order to acquire the data of two external variables – age and gender, Epistat – Covid19 Monitoring[1] is used as the primary data source. Epistat is a data repository which contains Epidemiology of Infectious Diseases Statistics maintained by Sciensano[2] – the Belgian research institute of health. The second source is Worldometers[3] that is being used for the latest and live COVID-19 updates, these days. With the intention of verification of different variable counters in the data, a widely referred COVID-19 API[4] is used which is sourced from John Hopkins CSSE.

The Epistat data source has information for COVID-19 total cases and total deaths in terms of age and gender. To analyze the correlation of age and gender, following are the variables that are merged as a single database: daily deaths, daily recovery from Worldometers and COVID-19 API, cases by age and gender, deaths by age and gender, hospitalization, tests performed from Epistat, and counters for all these variables are verified from COVID-19 API. Table 1 presents the data summary. However, the variables are also further classified into sub-classes as per data pre-processing requirement.

[1] https://epistat.wiv-isp.be/covid/

[2] https://www.sciensano.be/en/about-sciensano

[3] https://www.worldometers.info/coronavirus/

[4] https://covid19api.com/

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

The recovery data is not available in terms of age and gender. The statistics such as how much people recovered by gender, what is the recovery rate by age groups cannot be addressed in the study. Table 1 shows the data summary.

***Table** 1: Data Summary*

Variables	Source	Time Period
Cases by Age Groups and Gender	Epistat	Mar 01 - Apr 10
Deaths by Age Groups and	Epistat	Mar 10 - Apr 10
Deaths and Recovery	Worldometers and Epistat	Mar 01 - Apr 10
Tests Performed	Epistat	Mar 01 - Apr 10
Hospitalized	Epistat	Mar 15 - Apr 10
Recovery Cases	apicovid19.com	Mar 01 – Apr 10

3.2 Hypothesis

The null hypothesis H_o is: there is no statistically significant relationship between age and gender, and COVID-19 surveillance variables. Considering the continuous nature of the variables, the Pearson correlation analysis is used to test the hypothesis. In all, the following are the null sub-hypothesis:

1. Age groups are not statistically significantly related to the number of deaths
2. In terms of death, there is no statistically significant relationship between age groups and gender
3. Recovery rate is not statistically significantly related to the death rate
4. Tests performed are not statistically significantly related to the number of deaths
5. Hospitalization has no statistically significant effect on the number of deaths

3.3 Correlations

In the underline data, death is considered as the dependent variable. Regarding the concerned surveillance variables, the following correlations are calculated:

- Age and death
- Gender and death
- Recovery and death
- Tests performed and death
- Admitted and death
- ICU and death
- Respiratory and death

We use mainly two widely used statistical methods to analyze the dataset presented in section 3.1, first frequency and second correlation. Frequencies are obtained from the dataset available from the various sources.

3.3.1 Correlations among numeric variables

For correlations among numeric variables (Admitted, ICU, Respiratory, Recovery, Tests Performed, Deaths), Pearson's correlation is used as formulated in Eq. (1). The variables are

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

represented by x and y and $\bar{x}$ and $\bar{y}$ represent the means of the variables. The correlation value is between -1 and 1. Negative 1 represents a strong negative correlation, 1 represents a strong correlation, and 0 represents no correlation between the two variables.

$$r_{xy} = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2} \sqrt{\sum_i (y_i - \bar{y})^2}} \quad (1)$$

3.3.2 Correlations among categorical variables

For finding the correlations among categorical variables (age group and gender), we use Cramer's V. Cramer's V is based on chi-squared statistic and is used for categorical variables. Given two variables x and y, Cramer's V returns a value between 0 and 1 representing no and strong correlations, respectively. v is defined in Eq. (2), where χ^2 is defined in Eq. (3), n is the total observations of the variables x and y, r and k represent the unique values, n_i and n_j represent the frequencies of the i^{th} and j^{th} unique values observed in the variables x and y, respectively, and n_{ij} represents the number of times i^{th}, the unique value of variable x is observed with the j^{th}, the unique value of the variable y.

$$v = \sqrt{\frac{\chi^2/n}{min(r-1, k-1)}} \quad (2)$$

$$\chi^2 = \sum_{i,j} \frac{\left(n_{ij} - n_i n_j / n\right)^2}{n_i n_j / n} \quad (3)$$

3.3.3 Correlations among a categorical and a numerical variable

Pearson's correlation and Cramer's V allow to find correlation among two numeric and two categorical variables, respectively. To find the correlation between a numeric and a categorical variable, we use the polyserial correlation ρ [19] is used as formulated in Eq. (4) . Given a numeric (like number of deaths) and a categorical variable (like age group or gender), polyserial correlation returns a value between -1 and 1 representing a strongly negative and a strongly positive correlation, respectively.

$$\rho = \frac{\sum_{j=1}^{r} \mu_j \left[\phi(\tau_j - 1) - \phi(\tau_j)\right]}{\sum_{j=1}^{r} \frac{\left[\phi(\tau_j - 1) - \phi(\tau_j)\right]^2}{p_j}} \quad (4)$$

where μ_j is the expectation of the observed continuous variable X given $Y = p_j$, and τ_j is the location of thresholds. The base article [19] can be referred for the detailed explanation of the Eq. 4.

3.4 Pseudocode

Data curation is the key component of the underline methodology of this research work. Several data sources contain information about COVID-19 surveillance variables, however, some of them are reliable. For this research, the data is curated from three sources discussed in section 3.1. In each different data source, the data for the same variable was not available for the selected time period i.e., Mar 01, 2020, to Apr 10, 2020. For instance, hospitalization data is available from Mar 15, 2020, onwards. For the sake of computations, the problem is solved by using the available-ranged data. After merging data from the aforementioned sources, the pairs of variables are analyzed for correlations. P^n denotes the numerical/continuous and numerical/continuous pair, P^c denotes the categorical and categorical pair, and P^x denotes the categorical and the numerical pair. Subsequently, the t-test is also performed using student's t-

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

test to verify the significance of the calculated correlation. In summary, the research methodology is presented in the form of the following pseudocode:

Pseudocode: Correlation and Significance Matrices

Input: D_e, D_w
Output: **C**, **M**
BEGIN

1. Curate: $D_M \leftarrow D_e$ and D_w
2. Pre-process and Clean: $D_M \leftarrow D_c$
3. **for each** $s_i \in S$
4. **if** $s_i \in P^n$
5. Compute correlation value for s_i using Eq. 1
6. **C** $\leftarrow s_i$
7. **else if** $s_i \in P^c$
8. Compute correlation value for s_i using Eq. 2
9. **C** $\leftarrow s_i$
10. **else**
11. Compute correlation value for s_i using Eq. 4
12. **C** $\leftarrow s_i$
13. Compute p-value for s_i using t
14. **M** $\leftarrow s_i$
15. **end for**
16. **return C**, **M**

END

***Table** 2: Notations for pseudocode*

Notation	Description
D_c	apicovid19 data source
D_e	Epistat data source
D_w	Worldometers data source
D_M	Merged data
S	Variable pairs in D_M
t	Student's t-test
C	Correlation matrix
M	Significance matrix

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

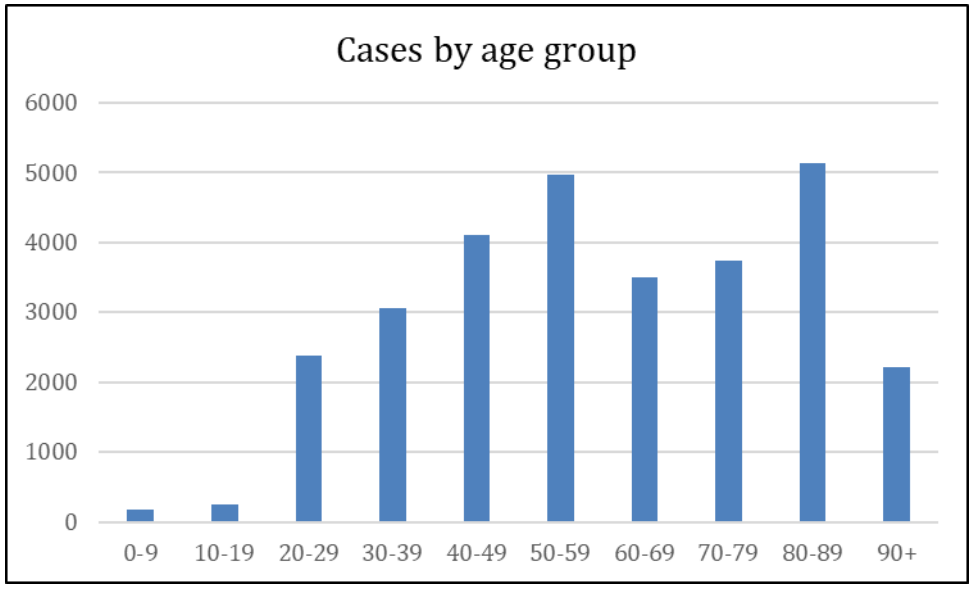

***Figure** 1: Number of cases of coronavirus*

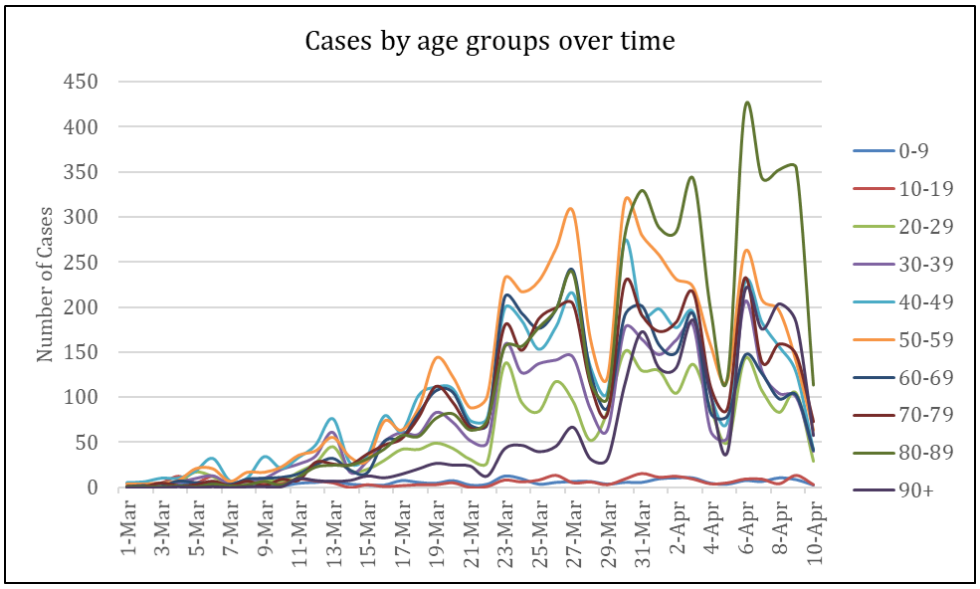

***Figure** 2: Cases of coronavirus of different age groups over time*

4. Results and Discussion

4.1 Age groups analysis

To understand the impact of age on coronavirus, we plot the number of cases reported in Fig. 1. We can observe that the coronavirus affects the people above the age of 50 the most. Its impact on people under the age of 20 is minimal, impacting only 434 people out of a total of 29,524 confirmed cases. Most of the population (around 39%) of Belgium is in the age group of 25 – 54 years[5], but the cases reported in these age groups are less than the cases reported in higher age groups. There are around 13% of people in Belgium in the age group of 55 – 64 years and around 19% of the age 65 or above. Despite having a relatively lower population in higher age groups, the number of coronavirus cases reported in these age groups is higher.

[5] https://www.cia.gov/library/publications/the-world-factbook/fields/341.html#BE

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

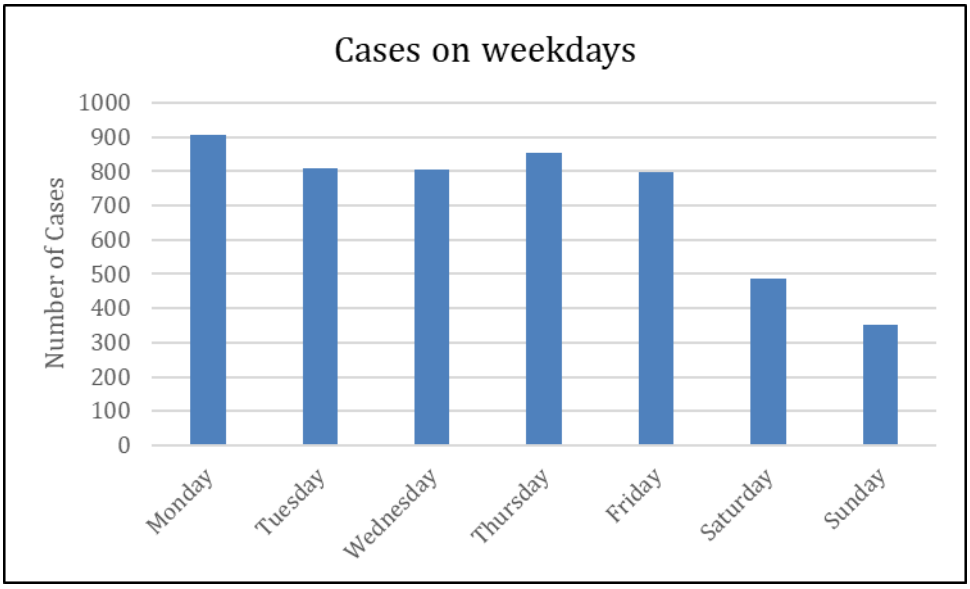

***Figure** 3: Number of cases reported on weekdays*

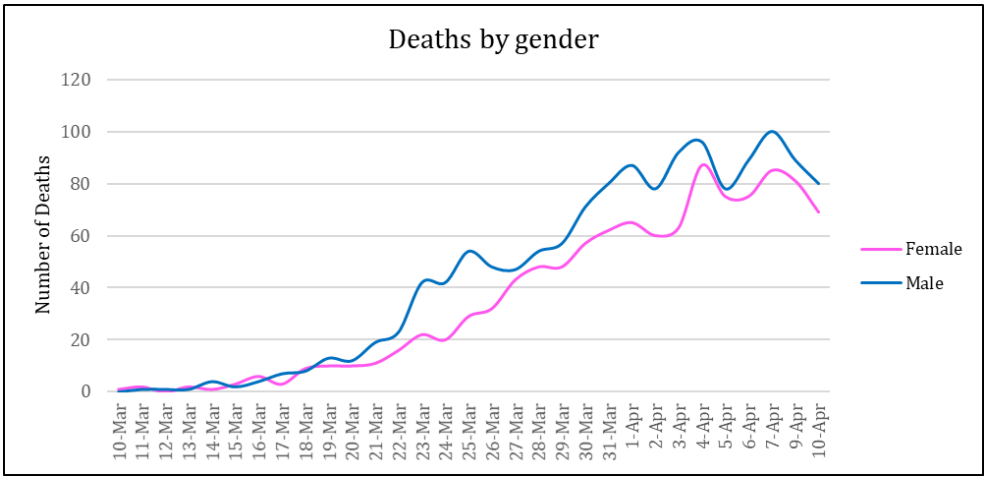

***Figure** 4: Deaths over time for both the genders*

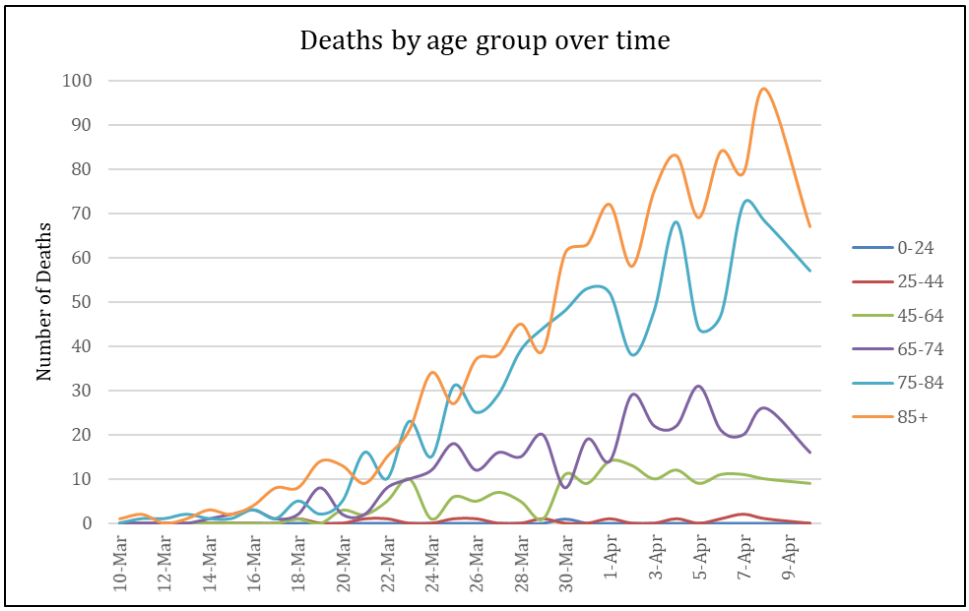

***Figure** 5: Deaths due to coronavirus over time*

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

Fig. 2 shows the number of cases reported over time between March 1 and April 10. The cases keep increasing until April 10, with periodic dips. The dips are due to the weekends, as a lesser number of cases are reported over the weekends. Fig. 3 shows the average number of cases reported over days of the week. We can observe that the least number of cases are reported over the weekends, whereas the highest number of cases are reported on Mondays.

To understand the vulnerability of people belonging to different genders over time, we show the people died over time in Fig. 4, and for various age groups in Fig. 5. The number of depths kept increasing over time. Although, a number of cases reported for females have been more as compared to the males, however, the mortality rate among males is higher than females and this observation is almost consistent over time (except for the start of the pandemic). The higher the age group, the greater the number of people died due to the virus. Most affected people were in the age group of 85 years and above, followed by the people in the age group of 75 and 84. The deaths of people below 44 years were minimal. This shows that the people above the age of 45 must be protected the most, as they are the most vulnerable. Fig. 6 shows the aggregated number of deaths for different age groups. Apparently, age has a strong correlation with mortality.

4.2 Age group and gender analysis

To analyze the impact of coronavirus on age and gender, we analyze the cases reported with respect age and gender. Fig. 7 shows the number of cases reported for different age groups against their gender. We can observe that for most of the age groups (except between 60 – 80 years old), more female cases are reported than males. Around 58% of the cases reported in Belgium were of females and 42% of males. The data for Belgium reports contrary to gender impact as compared to other places[6], for example, in New York City around 38% of the patients were females and 62% males.

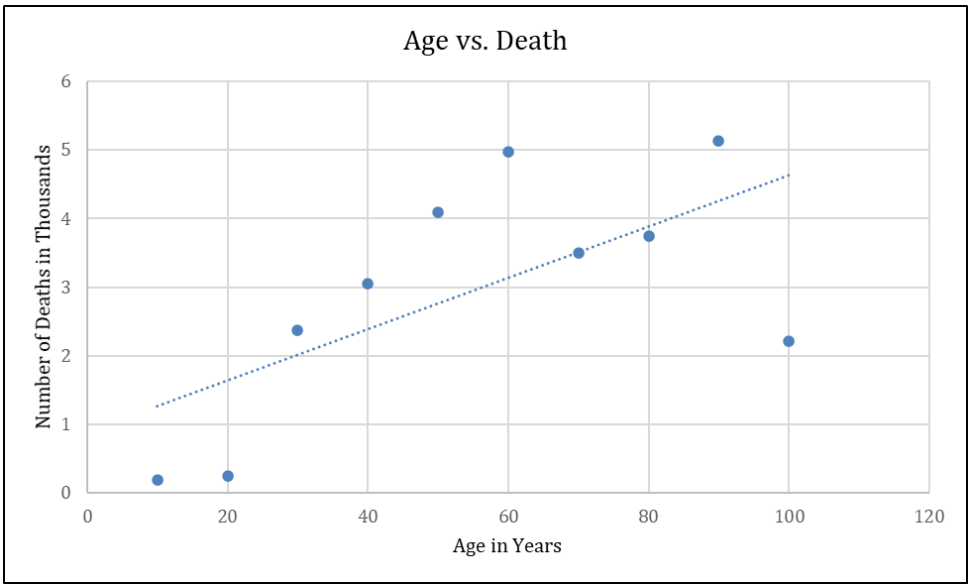

***Figure** 6: Number of deaths in thousand for different age groups*

[6] https://www1.nyc.gov/assets/doh/downloads/pdf/imm/covid-19-daily-data-summary-deaths-04102020-1.pdf

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

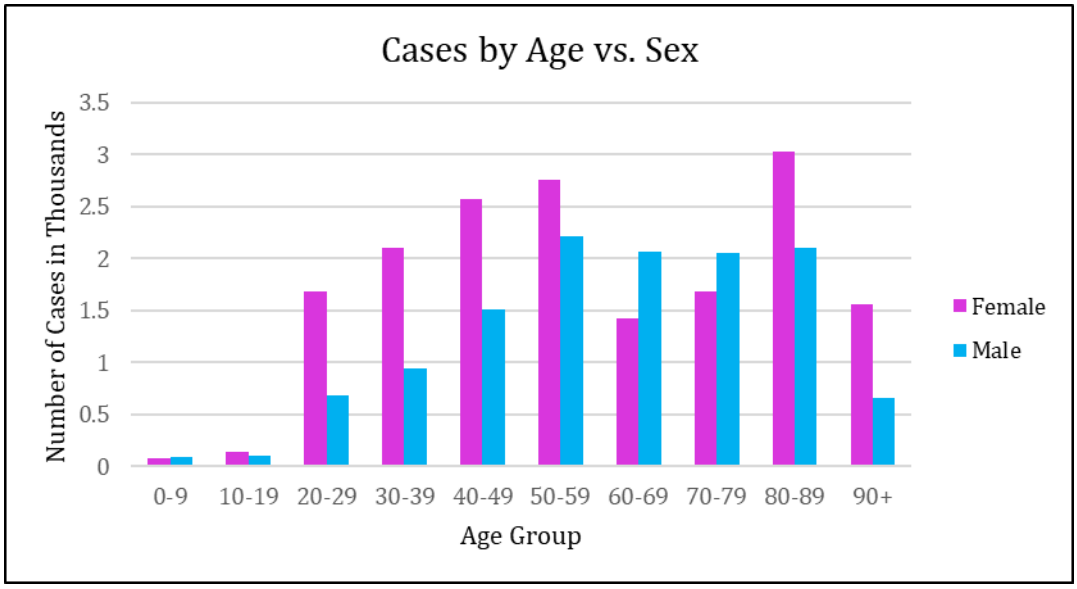

***Figure** 7: Number of coronavirus cases for different age groups and genders*

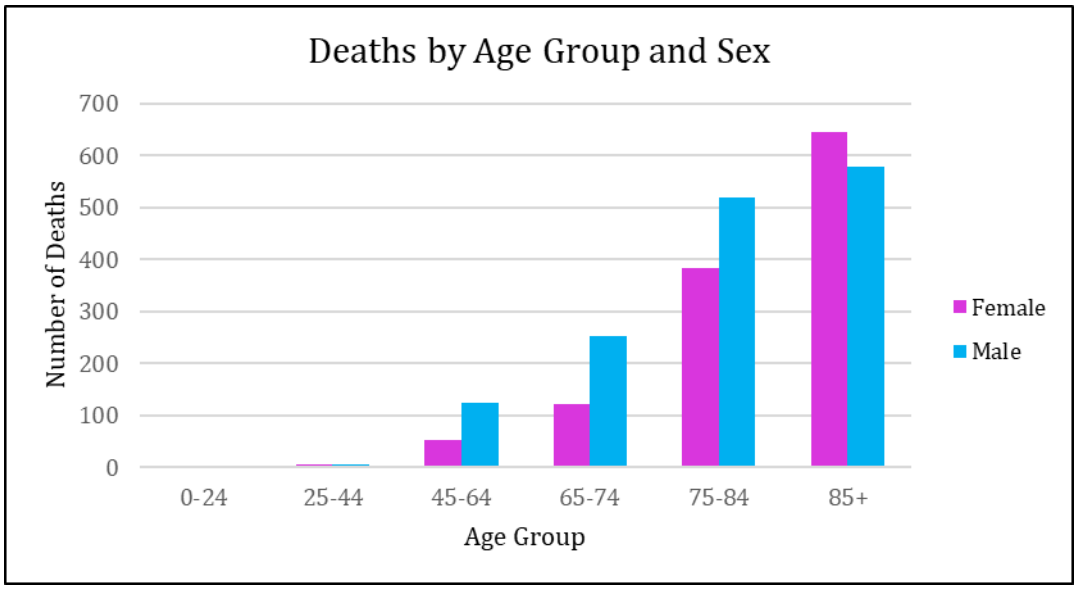

***Figure** 8: Number of deaths due to coronavirus for different age groups and genders*

Fig. 8 shows the number of deaths for various age groups and genders. Females above 85 years of age have a higher mortality rate as compared to the males in the same age group. However, for all the other age groups (above 45 years), males have a higher mortality rate.

To understand the impact of the disease on gender and age group, Fig. 9 shows the increase in cases over time. We can observe that, towards the later stage of the pandemic, females of all age groups except 0-19 and 60-79 are impacted in the same way as males, but for all the other cases, females are significantly more impacted than the males. The number of male cases for almost all the age groups and time periods were either less than or equal to the female cases.

4.3 Recovery analysis

As the number of cases increases, we see an increase in the number of deaths and number of people recovered. Fig. 10 shows the people recovered and deaths occurred over time. Fig. 11

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

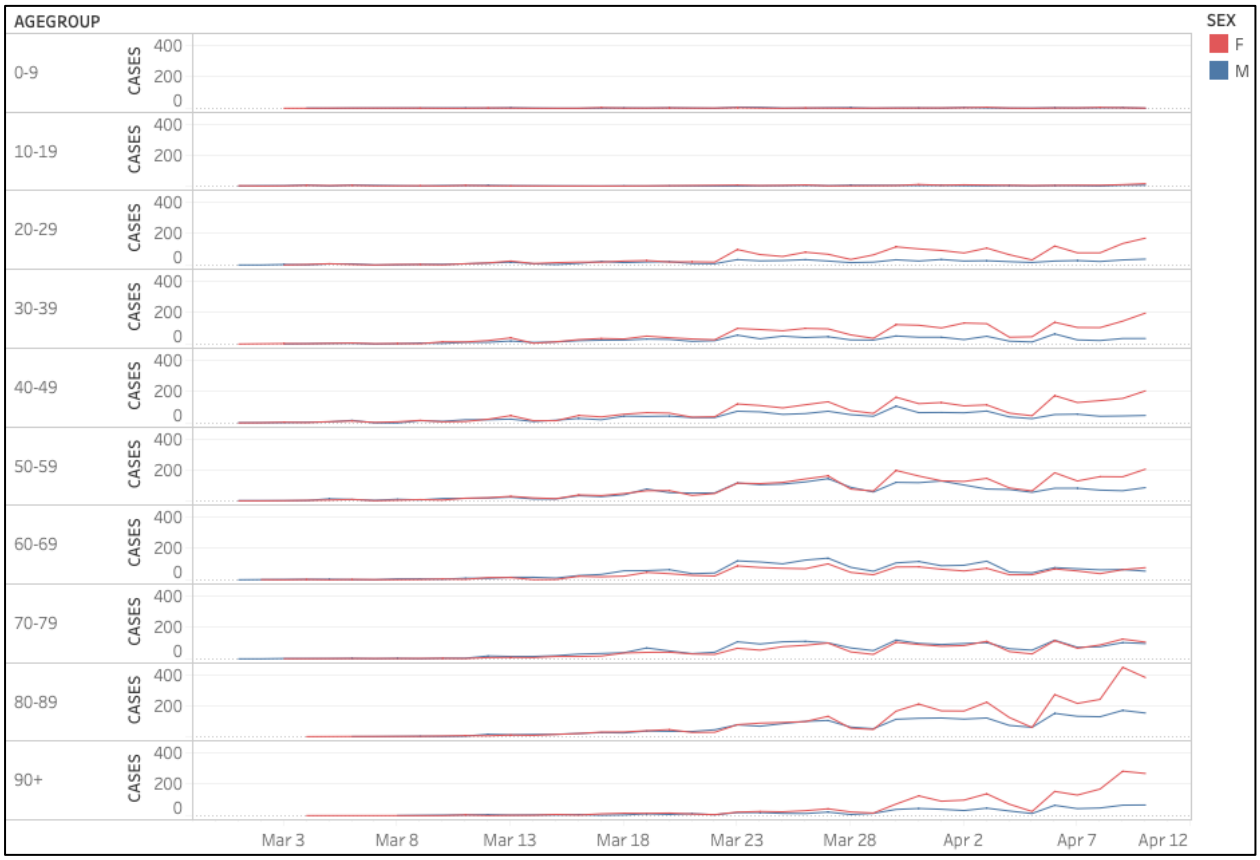

***Figure** 9: Comparison of age groups and gender over time*

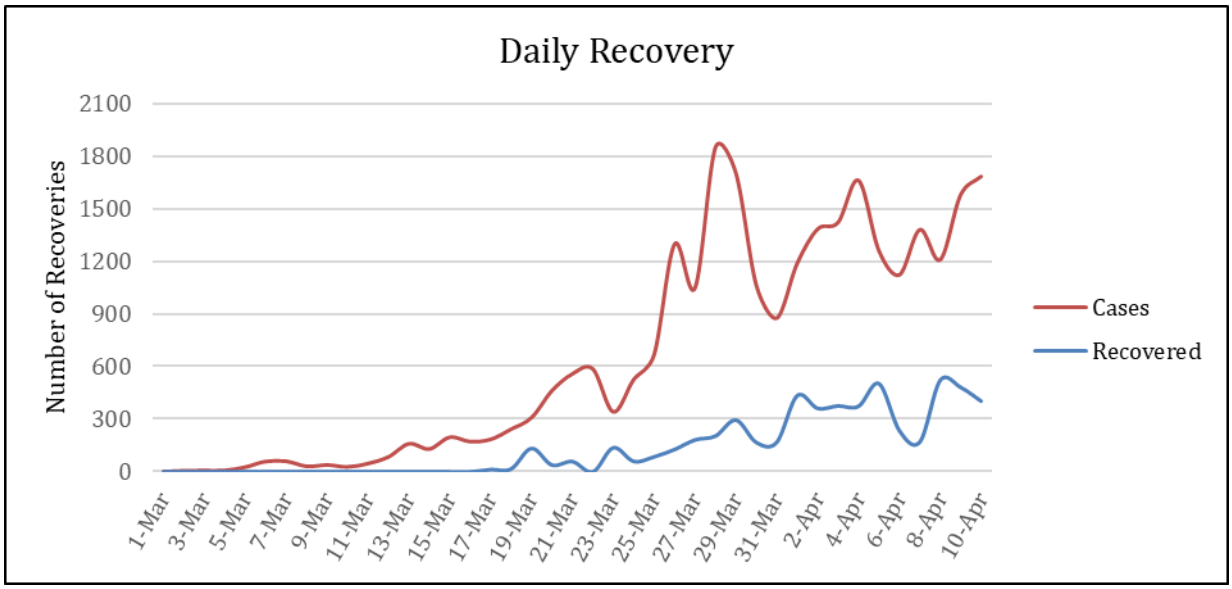

***Figure** 10: Comparison of the total number of cases and people recovered*

shows a scatter plot between the number of deaths and people recovered. Initially there were more deaths reported than the number of people recovered, however later on the number of people recovered was more than the deaths reported. The relation between the numbers of deaths and people recovered is linear, with a strong correlation. It also shows that hospitals with advanced medical equipment can increase the recovery rate.

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

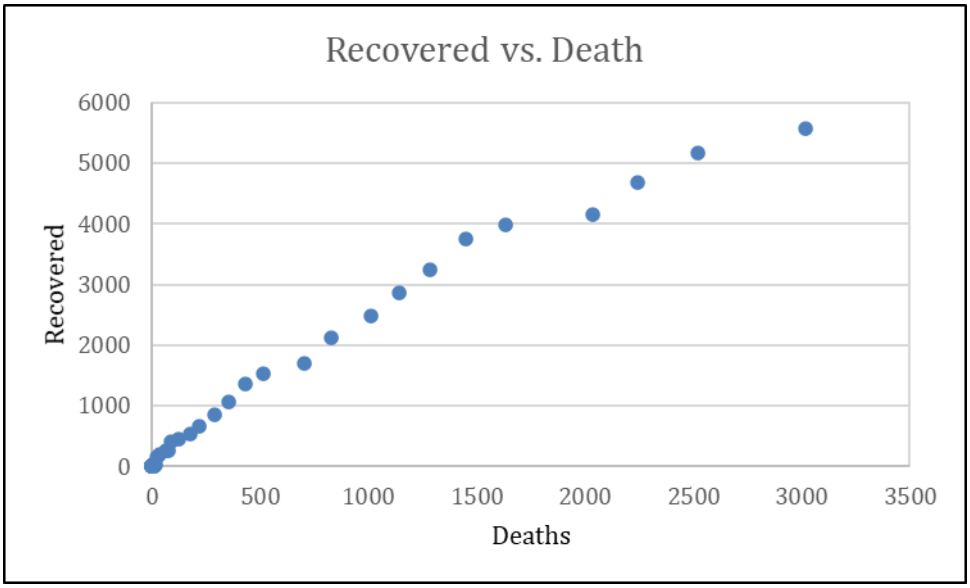

***Figure** 11: Comparison of the number of deaths and people recovered*

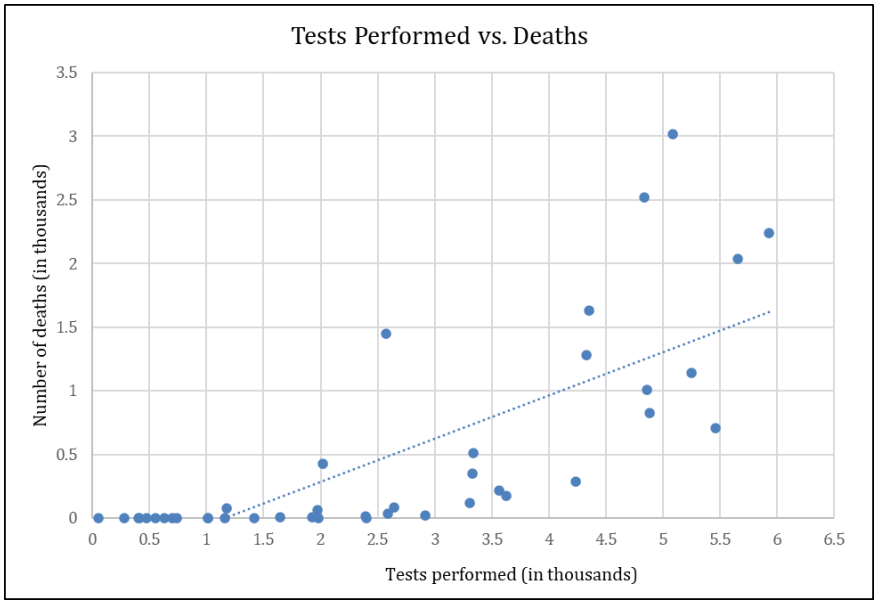

***Figure** 12: Correlation between the number of deaths and tests performed*

4.4 Tests performed analysis

It is important to conduct tests to identify positive cases and take preventive measures to minimize the spread of the disease. Fig. 12 shows the correlation between the number of tests performed and deaths reported on that day. There is a strong correlation between the two variables and it is also expected, as the number of cases increases, so are the number of tests performed and deaths.

4.5 Hospitalization analysis

People having symptoms of the disease or who are expected to be infected (like people in contact with an infected person) need to be tested, if they are tested positive, they might be

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

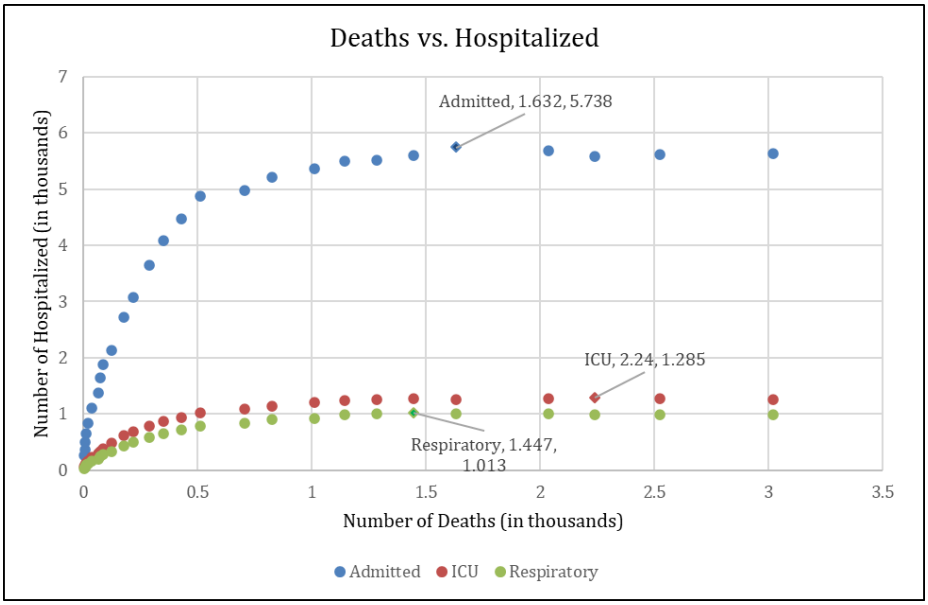

***Figure** 13: Comparison of the number of deaths with people 1) admitted, 2) in ICU, 3) getting respiratory care*

hospitalized. Among the hospitalized, some might just be quarantined without the need for intensive care, others might get intensive care in Intensive Care Units (ICU), or might be put on ventilators (respiratory). Fig. 13 shows the correlation between the number of people hospitalized and deaths occurred on a particular day. There is a strong correlation among the number of deaths and people admitted to the hospital (0.79), treated in ICU (0.81) and supported with a ventilator (0.82). For all the cases, the number of deaths increases with the number of cases. The death rate among people with less severe symptoms (admitted) is much lower than the people getting intensive care (ICU and respiratory). On any specific day, the maximum number of people admitted to a hospital were 5738 with 1632 deaths reported on that day. Similarly, for ICU there were 1285 people admitted with 2240 deaths and for respiratory 1013 people with 1447 deaths.

4.6 Correlation analysis

In this sub-section, we find the correlations and their statistical significance among the numeric variables we have discussed in the earlier sections. Table 3 shows the correlations among the variables calculated using Pearson's correlation and Table 4 shows corresponding significance values calculated using student's t-test. All the variables are statistically significantly strongly correlated with p-values significantly lower than 0.05. The highest correlation is among the variables: Deaths and Recovery, Admitted and ICU, Admitted and Respiratory, and ICU and Respiratory. We can use these correlations to predict how many deaths are expected based on the number of people admitted in the hospital among other variables.

In order to see whether categorical variables – Age Group and Gender correlate with the numeric variable – number of Deaths, we use polyserial correlation (see Section 3.3). Table 5 shows the correlation values along with the significance values. We observe that the correlation between Age Group and number of Deaths is statistically significant, this means that the Deaths can be predicted based on the number of people belonging to a particular Age Group, however, the correlation between the Gender of a person and the number of deaths is, although significant, but low (correlation = 0.04, p=1.36E-12).

Preprints (www.preprints.org) | NOT PEER-REVIEWED | Posted: 20 August 2020 doi:10.20944/preprints202008.0458.v1

***Table** 3: Correlation analysis of the variables*

	Recovery	Tests Performed	Admitted	ICU	Respiratory
Deaths	0.99	0.75	0.79	0.81	0.82
Recovery		0.78	0.86	0.87	0.88
Tests performed			0.8	0.81	0.81
Admitted				1	1
ICU					1

***Table** 4: Statistical significance values of the correlations reported in Table 2*

	Recovery	Tests Performed	Admitted	ICU	Respiratory
Deaths	9.06E-34	1.54E-08	7.73E-07	3.09E-07	1.94E-07
Recovery		1.30E-09	8.81E-09	2.49E-09	1.10E-09
Tests performed			5.11E-07	2.51E-07	3.85E-07
Admitted				2.16E-33	1.23E-29
ICU					1.34E-34

***Table** 5: Correlation and significance values between the categorical variables Age and Gender and the numeric variable Number of Deaths*

	Death	P-value
Age Group	0.56	2.54E-21
Gender	0.04	1.36E-12

The correlation among categorical variables Age Group and Gender are computed using Cramer's V (see Section 3.2), the correlation value is 0.03 (p=0.37), which shows that these two variables do not correlate and cannot be used to predict each other. The correlations and significance matrices show that the null hypothesis declared in section 3.2 are accepted with higher confidence. All the variable pairs are strongly correlated where p-values prove the significance of these relations. All the correlations discussed in this section are computed when the COVID-19 cases were still increasing day by day. A similar analysis over the complete timespan of the pandemic would be required at the end of the pandemic.

4.7 Comparison with existing information sources

The scientific community needs information fulfilment to make viable decisions to effectively control the pandemic. There exist a number of COVID-19 information sources, whereas Worldometers is one of them. However, they provide limited information that bounds the authorities to keep an eye merely on the COVID-19 counters. Worldometers[7] has information about age and gender, but it only covers the worldwide figures. However, our analysis covers date-wise age and gender data for surveillance variables which can help in the provision of viable medication with respect to time. Also, in contrast to our sparse age groups analysis, the Worldometers has dense age groups information.

[7] https://www.worldometers.info/coronavirus/coronavirus-age-sex-demographics/

 doi:10.20944/preprints202008.0458.v1

Other well-known information sources including The Guardian[8], Metro[9], Global Health 50/50[10], and so on, also encompass COVID-19 information. In an article, Guardian mentioned that [20] COVID-19 affected more men than women. However, this is a piece of generic meaningful information. Actually, using this information, one cannot distinguish COVID-19 dynamics particularly in terms of age or gender. In summary, the analysis presented in this study is supportive for higher authorities of the country to strengthen the health sector in the wake of the worldwide novel pandemic.

5. Conclusion

The analysis of age and gender is quite correlated with the COVID-19 surveillance variables. It helps the higher authorities to deploy the resources in the right direction in the fight against COVID-19. In Belgium, the mortality is rate higher in males than females, however, females got infected in the majority. Age has a strong correlation with death, nevertheless, 45+ aged people cover the larger portion of fatalities. Death is also strongly correlated with the hospitalized people, COVID-19 tests performed, and the recovery rate. It depicts that the Belgium government has assured the best possible medical facilities to the COVID-19 patients along with the necessary precautionary measures. The correlation between age group and number of deaths is statistically significant that can be used for death prediction based on number of people in a particular age group. However, the correlation between gender of a person and the number of deaths is, although significant, but low (correlation = 0.04). The analysis shows that age group can be used for death prediction, however, gender is weakly correlated with deaths.

In order to better understand the COVID-19 impact over the globe, the same analysis can be extended and compared with the other countries. Moreover, the meteorological variables such as temperature, humidity, and the economic variables such as oil prices, currency rates, or stock market can be analyzed along with the surveillance variables in the future.

Declarations

Funding: Funding information is not available

Conflict of interest: There is not conflict of interest among all of the authors

Availability of data and material: Available on request

Code availability: Available on request

References

[1] B. Gong, S. Zhang, L. Yuan, and K. Z. Chen, “A balance act: minimizing economic loss while controlling novel coronavirus pneumonia,” *J. Chin. Gov.*, pp. 1–20, 2020.

[2] I. Nesteruk, “Statistics based predictions of coronavirus 2019-nCoV spreading in mainland China,” *MedRxiv*, 2020.

[3] B. Smeets, R. Watte, and H. Ramon, “Scaling analysis of COVID-19 spreading based on Belgian hospitalization data,” *medRxiv*, p. 2020.03.29.20046730, Mar. 2020, doi: 10.1101/2020.03.29.20046730.

[4] T. Amjad, A. Daud, M. K. Hayat, M. T. Afzal, and H. Dawood, “Coronavirus Pandemic (COVID-19): A Survey of Analysis, Modeling, and Recommendations,” pp. 1–23, 2020.(submitted)

[8] https://www.theguardian.com/commentisfree/2020/apr/07/coronavirus-hits-men-harder-evidence-risk
[9] https://metro.co.uk/2020/04/05/coronavirus-uk-covid-19-seem-affect-men-women-12509747/
[10] https://globalhealth5050.org/covid19/

[5] A. Ahani and M. Nilashi, “Coronavirus Outbreak and its Impacts on Global Economy: The Role of Social Network Sites,” *J. Soft Comput. Decis. Support Syst.*, vol. 7, no. 2, pp. 19–22, 2020.
[6] C. Albulescu, “Do COVID-19 and crude oil prices drive the US economic policy uncertainty?,” *ArXiv Prepr. ArXiv200307591*, 2020.
[7] O. Reyad, “Novel Coronavirus COVID-19 Strike on Arab Countries and Territories: A Situation Report I,” *ArXiv Prepr. ArXiv200309501*, 2020.
[8] M. Cinelli *et al.*, “The covid-19 social media infodemic,” *ArXiv Prepr. ArXiv200305004*, 2020.
[9] S. J. Fong, G. Li, N. Dey, R. G. Crespo, and E. Herrera-Viedma, “Composite Monte Carlo Decision Making under High Uncertainty of Novel Coronavirus Epidemic Using Hybridized Deep Learning and Fuzzy Rule Induction,” *ArXiv Prepr. ArXiv200309868*, 2020.
[10] A. Brandenburg, “Quadratic growth during the 2019 novel coronavirus epidemic,” *ArXiv Prepr. ArXiv200203638*, 2020.
[11] H. H. Elmousalami and A. E. Hassanien, “Day Level Forecasting for Coronavirus Disease (COVID-19) Spread: Analysis, Modeling and Recommendations,” *ArXiv Prepr. ArXiv200307778*, 2020.
[12] Y.-C. Chen, P.-E. Lu, C.-S. Chang, and T.-H. Liu, “A Time-dependent SIR model for COVID-19 with Undetectable Infected Persons,” 2020.
[13] L. Jia, K. Li, Y. Jiang, and X. Guo, “Prediction and analysis of Coronavirus Disease 2019,” *ArXiv Prepr. ArXiv200305447*, 2020.
[14] L. Zhong, L. Mu, J. Li, J. Wang, Z. Yin, and D. Liu, “Early Prediction of the 2019 Novel Coronavirus Outbreak in the Mainland China based on Simple Mathematical Model,” *IEEE Access*, 2020.
[15] L. Li *et al.*, “Propagation analysis and prediction of the COVID-19,” *Infect. Dis. Model.*, vol. 5, pp. 282–292, 2020.
[16] E. Shim, A. Tariq, W. Choi, Y. Lee, and G. Chowell, “Transmission potential of COVID-19 in South Korea,” *medRxiv*, 2020.
[17] Y. Zheng, Y. Zi-xia, and J. I. A. Zu-yao, “Estimating the Number of People Infected with COVID-19 in Wuhan based on Migration Data,” 电子科技大学学报, vol. 49, pp. 1–9, 2020.
[18] S. Shahid Nadim, I. Ghosh, and J. Chattopadhyay, “Short-term predictions and prevention strategies for COVID-2019: A model based study,” *arXiv*, p. arXiv–2003, 2020.
[19] U. Olsson, F. Drasgow, and N. J. Dorans, “The polyserial correlation coefficient,” vol. 47, no. 3, pp. 337–347, 1982, doi: 10.1007/BF02294164.
[20] P. Ball, “Coronavirus hits men harder. Here’s what scientists know about it | Philip Ball,” *The Guardian*, Apr. 07, 2020.