Investigating the Impact of Gender and Age on the Severity of COVID-19*

Shuhan Yang

2024-01-25

COVID-19 has greatly impacted the global society for the past few years. This research aims to discover infection trends of COVID-19, aiding the development of more effective strategies for future pandemic preparedness, including enhanced control plans and hospital infrastructure constructions. Downloading the COVID cases data from Open Data Toronto, graphical analysis including bar charts and line charts are conducted to access distribution patterns, severity levels, and trends. The findings conclude that elderly females are most severly infected and the younger generation tends to be infected more frequently.

Table of contents

*Code and data are available at: https://github.com/Jasmineee35/Investigating-the-Impact-of-Gender-and-Age-on-the-Severity-of-COVID-19.git

1. Introduction

COVID-19 has significantly impacted our society on a global scale. Typically, widespread viruses possess a pattern of targeting certain demographics (Piret and Boivin 2020). For instance, males can be are more susceptible to infection in the case of virus x, while children are more susepible of catching virus y. Therefore, this research serves the purpose to investigate whether COVID-19 has targeted demographics, more specifically, the research will focus on the effects of gender and age.

The COVID-19 data retrieved from opendatatoronto provides a public record of individuals infected by the virous (tested positive). Analyzing trends in this dataset by employing a statistical approach can possibly help the toronto public health system improve virus prevention strategies, such as identifying high-risk populations and insights to vaccine developments, in case of future pandemics or breakouts.

In the following paragraphs, I will first introduce the dataset, then move on to the initial statistic analysis of the data. I will use two bar charts to illustrate the overall distribution of the two main variables I want to study for: age and gender. Then, using another bar chart, it will indicate the COVID cases by both age and gender groups, showing the female and male distribution for each age interval side by side. Moving on to the Severity Analysis section, I will conduct additional analysis to study the specific impact that a person's age and gender has on the severity of their COVID infection. The final section will touch on trend analyses to summarize COVID development with the passage of time.

2. COVID-19 Data

To investigate the COVID infectious trends in Toronto, I retrieved the dataset "COVID19 cases.csv" from the Toronto Open Data Portal (Gelfand 2022). Other similar data exists, but I want this research to base on the most recent updated one, as of year 2024, so "COVID19 cases.csv" is specifically chosen. This unedited data (n.d.a) contains 413474 observations and 15 variables. The 15 variables include the id, outbreak associated, age group, neighbourhood name, source of infection, gender, and other related variables for each observations. Among these 15 variables, this report will focus on four major variables to conduct further analysis: age, gender, ever_hospitalized, and date. Variable age is measured by a number range for instance 20-29; gender contains 9 different types in total with one of them listed "Not Listed"; ever-hospitalized is either "Yes" or "No". By using R (R Core Team 2023), and R packages "tidyverse" (Wickham et al. 2019), "janitor" (Firke 2023), "lubridate" (Grolemund and Wickham 2011), "kableExtra" (Zhu 2021), knitr" (Xie 2023), I first cleaned the dataset by checking missing values and convert the dates format to "Year-Month-Day", then using (Wickham 2016) to conduct graphical analysis based on the dataset. The following table (see Table 1) will illustrate 6 rows for the COVID-19 cases data, to gain a general idea of what the dataset looks like.

Table 1: Illustration of the COVID-19 Cases Dataset

X_id	Assigned_ID	Outbreak.Associated	Age.Group	Neighbourhood.Name	FSA	Source.of.In
1	1	NO	50 to 59 Years	Willowdale East	M2N	Travel
2	2	NO	50 to 59 Years	Willowdale East	M2N	Travel
3	3	NO	20 to 29 Years	Parkwoods-Donalda	M3A	Travel
4	4	NO	60 to 69 Years	Church-Yonge Corridor	M4W	Travel
5	5	NO	60 to 69 Years	Church-Yonge Corridor	M4W	Travel
6	6	NO	50 to 59 Years	Newtonbrook West	M2R	Travel

3. Graphical Analysis

Having a foundation knowledge of the dataset, we will now come to the graphical analysis part. We will look at the direct distribution of age and gender using bar charts. A bar chart can provide clear view on the total distribution of the data: mean, median, skewness, shape.

3.1. Distribution Analysis

The following graph (see Figure 1) displays the distribution of COVID-19 cases across different age groups.From the bar chart above, the data is positively skewed since it has a longer tail on the right hand side. The skewness aligns with the shape of the distribution that majority of the dataset is concentrated on the left hand side, meaning that more COVID cases are younger generations.

Looking at another bar chart (see Figure 2) that shows the distribution of COVID-19 cases between different types of genders, the graph shows that within this dataset, more COVID cases are distributed in the female group. Now we can summarize that based on the initial analysis, female and younger individuals are more likely to get infected by COVID.

To further testify this conclusion, analysis for the data that include both variable: age, gender is conducted. This bar chart (see Figure 3) will show individuals' gender side by side within each age group interval. Analysing the graph, I found that female typically have more chance to get infected compare to male especially in younger and middle age: 20-59. It also worth noticing that female aging 90 years or older are approximately twice as likely to get COVID than male. Thus if the nation encounter another similar type of virus, the health system can implement provision strategy with a focus on old age females as well.

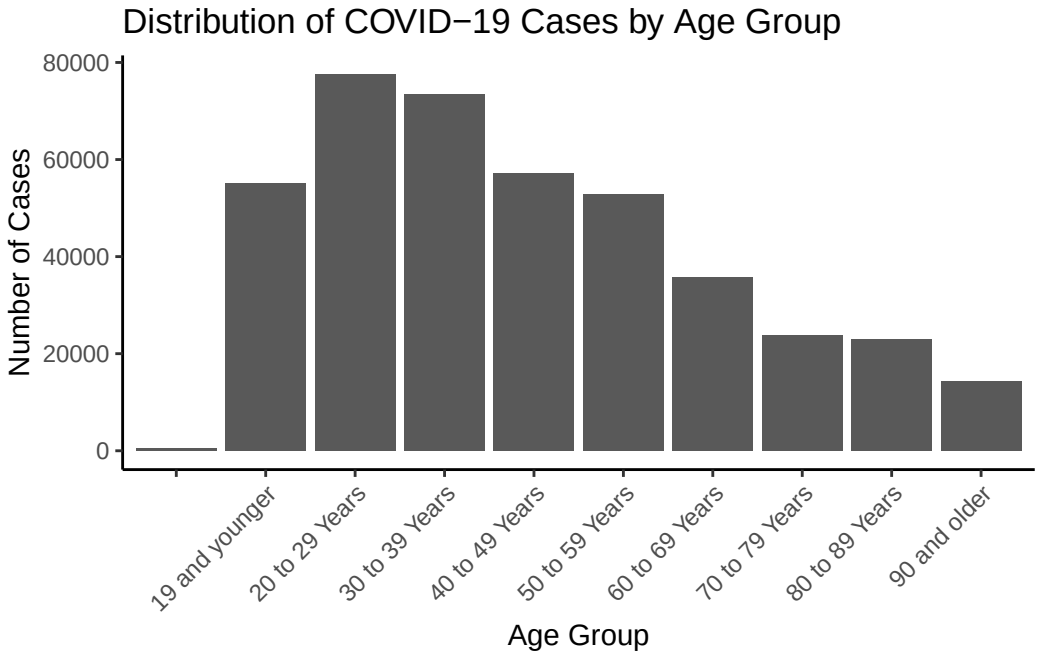


Figure 1: Bar Chart for Age Group Distribution

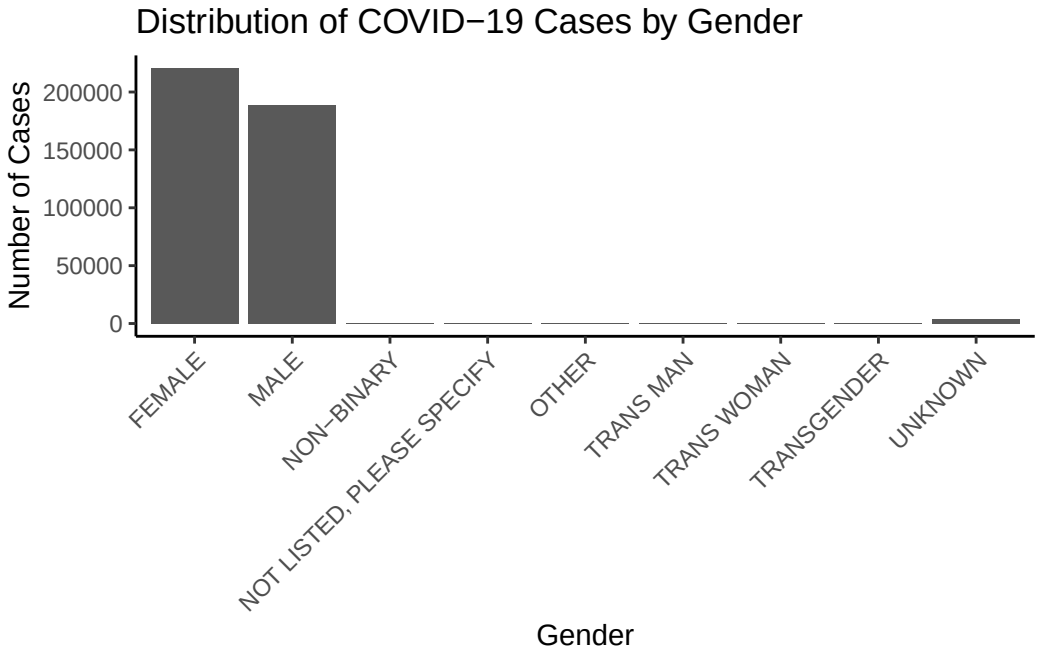


Figure 2: Bar Chart for Gender Distribution

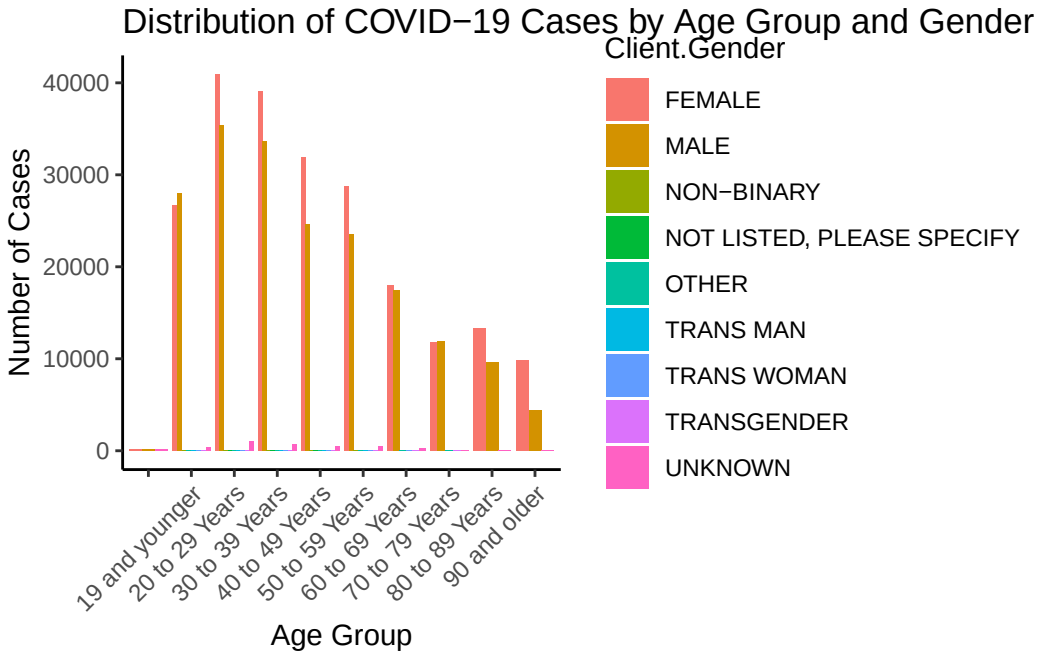


Figure 3: Bar Chart for Age Group and Gender Distribution

3.2. Severity Analysis

Having a foundation knowledge about the distribution of our dataset, we will now look at how will age group affect the severity of COVID infections. By developing the group most at risk, the public health system know how to allocate resources efficiently, such as the number of beds in hospital needed or quarantine plans, to prepare any future virus outbreaks. For this analysis, I assume that severity is based on hospitalization rates, with in hospital considered severe. Observing from the graph (see Figure 4), older age group especially 70 to 79 and 80 to 89 generation are severely affected by Covid infections. They exhibit the most hospitalized rate. This aligns with the study indicating older generations' immune system and somatic function are weaker than younger generations (n.d.b).

3.3. Trend Analysis

As COVID elapsed for several years, in this section, we will also use line graphs to study its trends with the progress of time. The below graph (see Figure 5a) shows COVID-19 started in 2020, with the infection number peaked around year 2021: approximately 175,000 cases. With the world getting adept to the virus, several vaccines were developed thus we see in the graph there is a sharp decrease from 2023 to 2024. In the current year, 2024, COVID cases nearly vanished as the World Health Organization announced the end of COVID emergency (n.d.c).

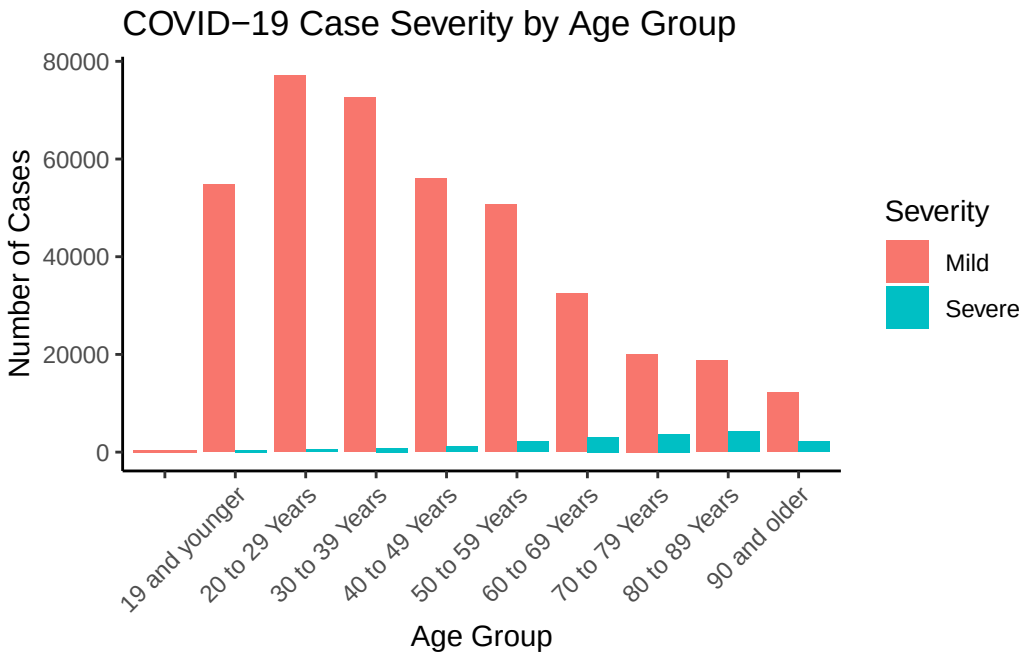
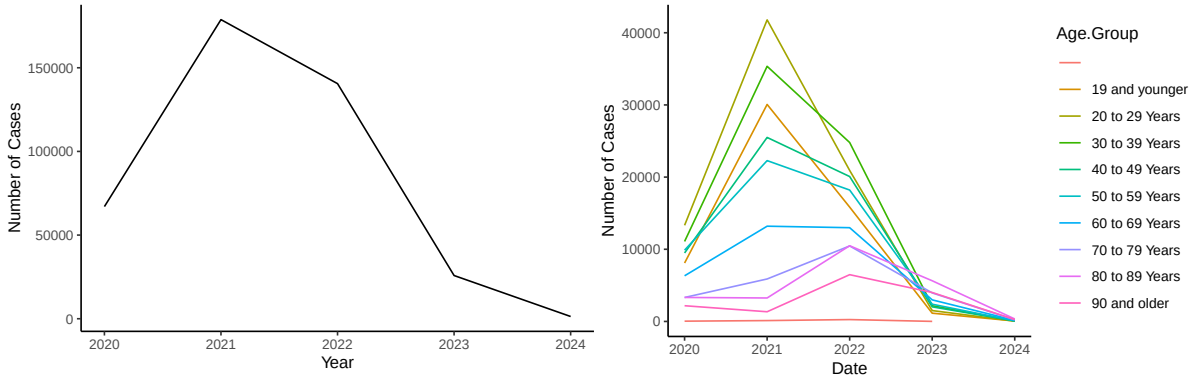


Figure 4: Bar Chart for Severity Analysis



(a) Trend of COVID-19 Cases Over Time

(b) Trend of COVID-19 Cases Over Time by Age Group

Figure 5: COVID Trends Analysis Over Time

To have a more detail look at each age segment contributing to the overall trend line that we previously discussed, this graph (see Figure 5b) shows a trend line for each age group with the passage of time on the x-axis. Individuals aging between 19 to 39 exhibit the most change in shape. The trends actually align with the overall trend line since as discussed in the distribution analysis, the younger generations consist the majority of COVID cases data.

4. Conclusion

Since COVID-19 alters our society in a multitude of ways, studies around this topic is essential as results can improve the overall health system for nations. In conclusion, this investigation on the impact of gender and age on the severity of COVID-19 has yielded several key findings. Firstly, the data indicates younger generations and females are more susceptible to COVID infections. While examining the severity analysis section, the outcome reveals that older age groups typically can contract more severe illnesses as evidenced by their higher hospitalized rate. Additionally, the distribution analysis indicates a notable trend: biological females of ages greater than 90 are almost twice as likely to be infected as their male counterparts. Based on the data analysis in this paper, the group that is most at risk of contracting severe COVID-19 infections in Toronto are elderly females.

5. Reference

n.d.a. *City of Toronto Open Data Portal*. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

———. n.d.b. *WebMD*. WebMD. https://www.webmd.com/healthy-aging/seniors-boost-immunity.

———. n.d.c. *Who Chief Declares End to COVID-19 as a Global Health Emergency - UN News*. https://news.un.org/en/story/2023/05/1136367.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://sharlagelfand.github.io/opendatatoronto/.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

Piret, Jocelyne, and Guy Boivin. 2020. "Pandemics Throughout History." *Frontiers*. Frontiers. https://www.frontiersin.org/articles/10.3389/fmicb.2020.631736/full.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. http://haozhu233.github.io/kableExtra/.