What factors can affect COVID-19 results?*

Statistical Analysis on COVID-19 cases in Toronto

Qingya Li

26 April 2022

Abstract

The paper studied the COVID-19 cases in Toronto. By creating some visualizations and applying binomial regression models, we found outbreak-associated, age, gender, infection source, classification type, ever hospitalized, ever in ICU, ever intubated are correlated to the recovery/fatality of the COVID-19 cases. It is necessary for us to conduct such research since it can provide a pathway for the prevention of the pandemic.

Keywords: COVID-19, pandemic, Toronto, Toronto Open Data Portal, Canada

Contents

*Code and data are available at: LINK.

1 Introduction

The global death toll attributable to COVID-19 has increased since the outbreak, arousing people's attention. The official data presented over 5.5 million confirmed death cases due to the pandemic, but the actual death tolls were more than the official counts (Adam 2022). In this case, we intended to study whether the COVID-19 fatal/resolved is associated with age, gender, source of infection, outbreak-associated, and other factors. We imported data from the Toronto Open Data Portal website (Health 2022). The data contained information about the Toronto COVID-19 cases between 2020 to 2022.

The paper mainly included four sections :2, 3 , 4 and 5. 2 section covered the data cleaning process and removed the unused variables in this paper. We performed some exploratory data analysis (EDA) for the variables with the cleaned dataset. In the 3 section, we conducted a binomial regression model since we considered the results of the COVID-19 cases linearly correlated to the analyzed variables. We applied COVID-19 results as the response variables (either resolved or fatal), other variables were explanatory variables in the models. Then we ended up with the findings in the 4 section. Lastly, we discussed the bias and weakness of the dataset in the 5 section.

We can learn more about COVID-19 through research. The findings can study the factors that may increase the probability of pandemic deaths. Additionally, providing a pathway for preventative and treatment initiatives.

2 Data

2.1 Data Source

The dataset was published by Toronto Public Health (TPH). The organization is responsible for the public health issue and well-being of over 2.9 million Toronto residents since 1883 (Toronto 2019). TPH has primarily addressed protecting and enhancing the residents' health. There are three principal objectives of TPH: 1) Preventing the disease spread and promoting the public health; 2) Monitoring the population's health status by applying surveillance, to respond to the ongoing health requirement; 3) Developing and applying public policy to promote the individuals, communities and the whole society' health (Toronto 2019).

2.2 Methodology and Data Collection

The dataset reported the ongoing and emerging COVID-19 pandemic outbreak in Toronto by TPH (Health 2022). The data was collected from the provincial Case & Contact Management System (CCM) (Health 2022). It included information about the demographic, geographic, and other individuals (age, gender, etc) information for the confirmed COVID-19 cases in Toronto from the first case recorded in January 2020 (Health 2022). The population of the data frame was the residents of the entire Toronto city, and the sample was the infected individuals. In addition, TPH mentioned that some limitations existed. The data we used may not be the latest. For the reason that the entire dataset will be refreshed and updated weekly (Health 2022). The data will be overwritten on the given Tuesday at 8:30 AM and released on Wednesday (Health 2022). The data in the provided dataset keeps updating as the TPH continues to report the new confirmed cases and improve the quality of the initiatives (Health 2022). In our paper, we used the dataset was released on April 13th. Another reason is the data may be different from the dataset reported by other organizations since the dataset was collected from different periods and numerous sources (Health 2022). Our report may come up with different conclusions from others as well.

2.3 Data Characteristics

The dataset is available at Toronto Open Data Portal, which can be accessed by using `opendatatoronto` (Gelfand 2020). Our report was written using the `R` statistical language (R Core Team 2020) and the following packages: dplyr (Wickham et al. 2021) and tidyverse (Wickham et al. 2019). In the visualizations, we plotted the bar charts using ggplot2 (Wickham 2016). The tables were conducted by knitr (Xie 2021) and the tidy function from the package of tidytext (Silge and Robinson 2016).

The dataset contains 18 variables and 32,000 observations. The dataset can be directly downloaded in CSV format. We imported the dataset with the provided sample codes: First, we imported the packages with the show_packages function. Second, we extracted the resources with list_package_resources. Third, we needed to identify the data resource with the filter function. Lastly, load the raw dataset with both filter and get_resource. The dataset was named raw_df. During cleaning the dataset, we created a new data frame called df which was saved with the write_csv function. Then we used the select function to keep the variables that we needed, including "Outbreak Associated", "Age Group", "Neighbourhood Name", "Source of Infection", "Classification", "Client Gender", "Outcome", "Ever Hospitalized", "Ever in ICU", "Ever Intubated". Then we used filter to exclude the "ACTIVE" cases from "Outcome" because we wanted to focus on the fatal and resolved cases. Also, we used filter to exclude the other options in "Client Gender" except males and females. We created a new variable called "result" to distinguish these two cases with mutate and as factor functions. 1 indicated the COVID-19 cases were resolved and 0 represented fatal cases. Moreover, we rename the rest variables into lower letters: "Outbreak Associated" as "outbreak associated", "Age Group" as "age", "neighbourhood" as "Neighbourhood Name", "infection" as "Source of Infection", "Classification" as "classification", "Client Gender" as "gender", "Outcome" as "outcome" "Ever Hospitalized" as "ever hospitalized", "Ever in ICU" as "ever in ICU", and "Ever Intubated" as "ever intubated". The variables are shown in the following.

Variables	Description
result	0 is the fatal cases, 1 is the resolved cases
age	cases's age group by 10
gender	self-reported biological gender
outbreak-associated	outbreak associated with healthcare institutions and other settings
neighbourhood	names of the 140 neighbourhoods
infection	the ways of infection of the COVID-19 cases
outcome	classify the cases into fatal and resolved
classification	classify the cases into confirmed and probable
ever hospitalized	The cases ever hospitalized due to COVID-19
ever in ICU	The cases ever in ICU due to COVID-19
ever intubated	The cases ever intubated due to COVID-19

2.4 Figures

Figures 1-2 illustrate the age at the time of the COVID-19 infection group by 10 and gender, respectively. The age distribution follows a normal distribution and is slightly skewed. We observed the mode is in the column of the age of 20-29 years old, follows by 30-39 years old, 19 and younger. Over 6,000 individuals are between 20 and 29 years infected. As the age gets older, the number of infections falls. The distribution of gender shows the number of infections between male and female groups is most likely the same. The results indicate the cases for both females and males are over 15,000, while the number of COVID-19 cases for males is slightly higher than for females.

Figures 3-7 demonstrate the distribution of COVID-19 cases outcome grouped by outbreak-associated, gender, age, infection, and classification, respectively. The results investigate the factors that may influence the outcome. We noticed that most of the fatal/ resolved cases are sporadic (more than 30,000) instead of the outbreak associated in figure 3. Figure 4 shows there seems no significant difference between the number of females and males in fatal and resolved cases. But we need to consider further investigation. We observed that most of the resolved cases occurred between 19- younger and 50-59 years, similar to the fatal cases. Figure 6 indicates the ways of infection in resolved and fatal cases. We found community as one of the infection sources that most significantly affects the outcome. Figure 7 displays the distribution of COVID-19 by classification. Most of the resolved cases are confirmed when all the fatal cases are confirmed. Moreover, the distribution of the COVID-19 outcome by every hospitalized, ever in ICU, ever intubated, respectively are shown in Section 5.2.4.

Therefore, we predicted that the outcome of the COVID-19 cases in Toronto is associated with the outbreak-associated, age, gender, infection, classification, ever hospitalized, ever in ICU, and ever intubated. With a further estimation, we intended to conduct a generalized linear model (binomial regression) that consists of all the explanatory variables mentioned above, and the response variable was the variable called " results".

Before we built the model, we checked whether the assumptions for the binomial regression were held. First, the response variable should be binary the result of COVID-19 is either fatal or resolved. Second, the observations are not correlated and independent of each other. Third, the sample size is not too small with 31,717 observations. Most importantly, the linearity assumption holds.

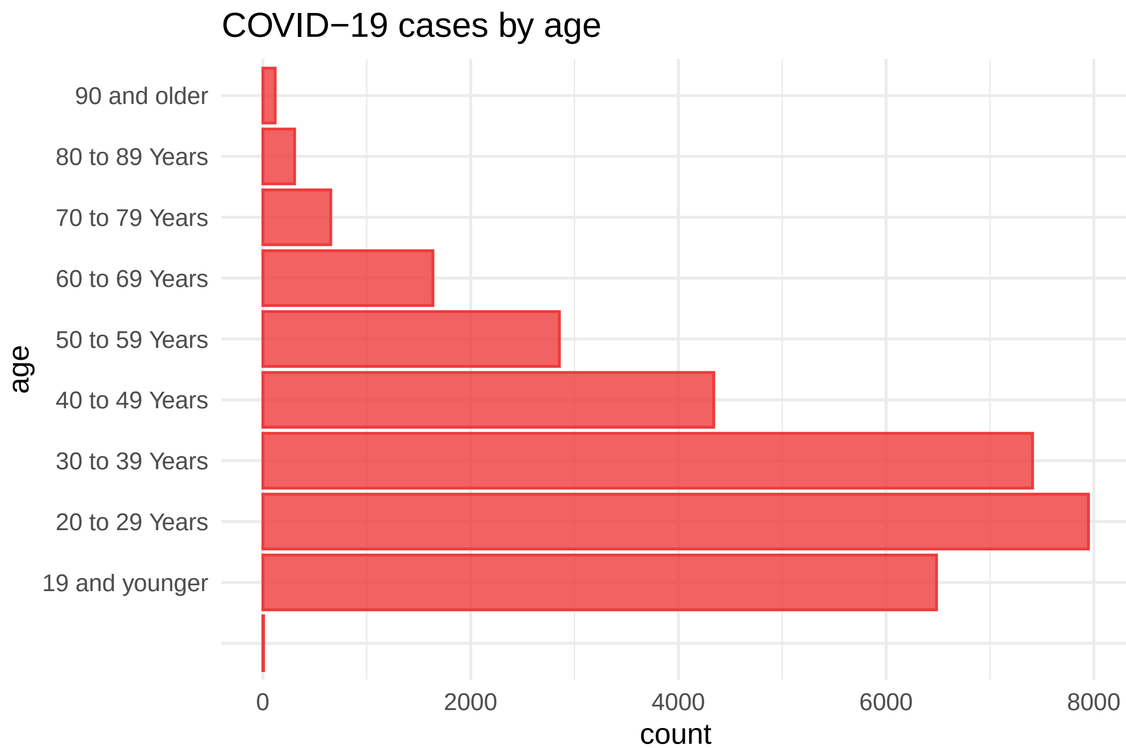

Figure 1: Distribution of COVID-19 cases by age

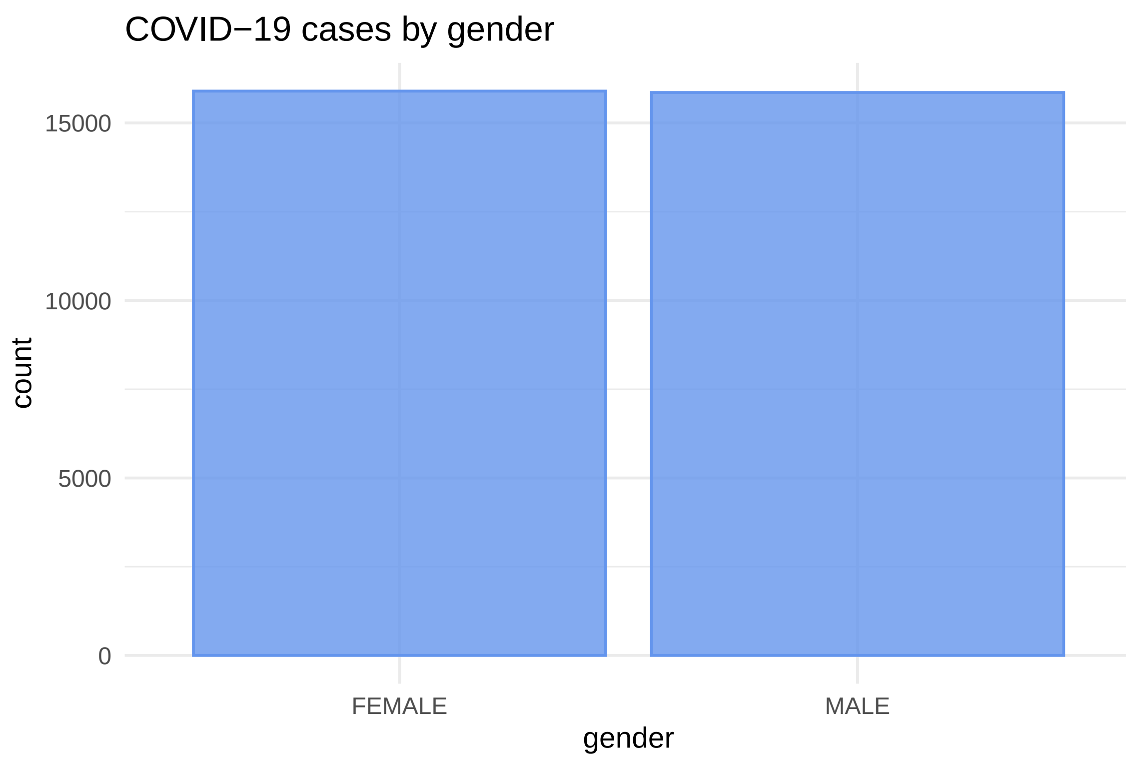

Figure 2: Distribution of COVID-19 cases by gender

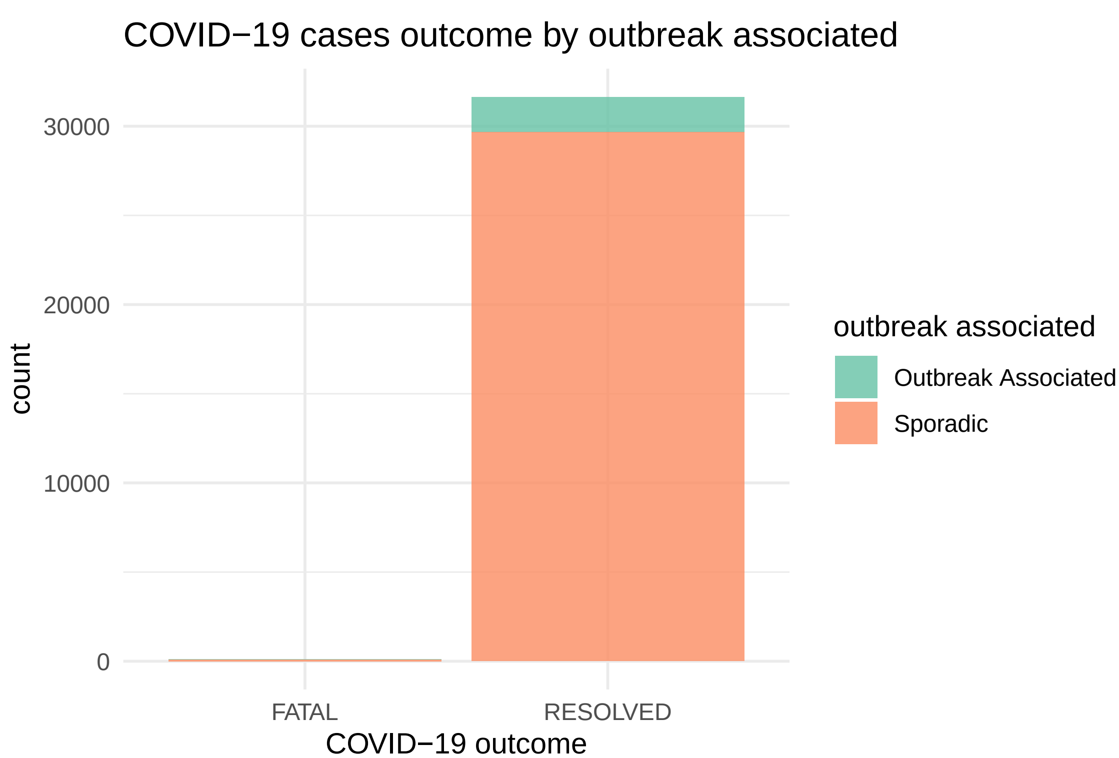

Figure 3: COVID-19 outcome by outbreak associated

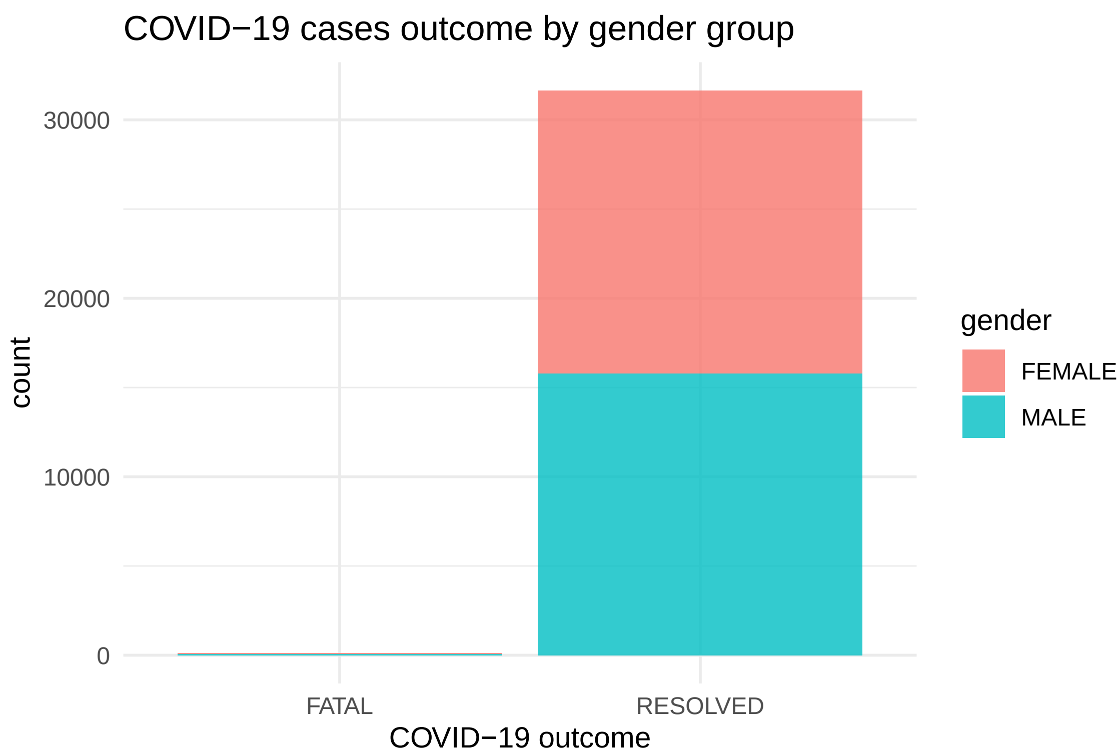

Figure 4: COVID-19 cases outcome by gender

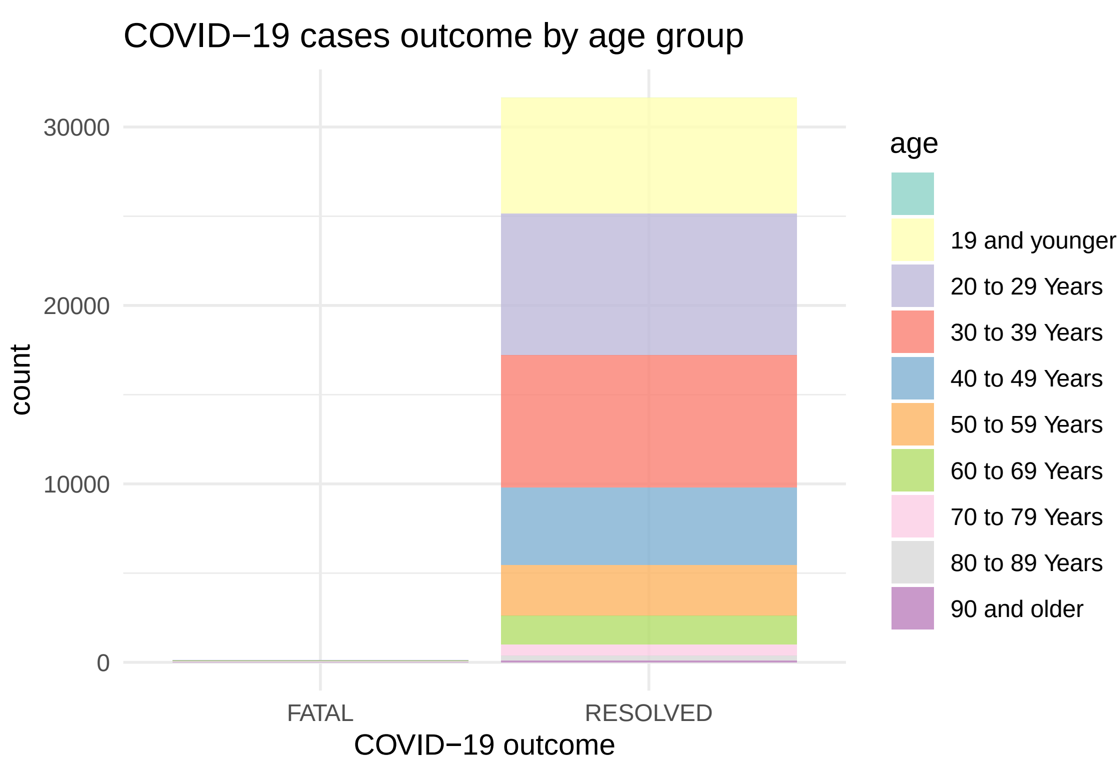

Figure 5: COVID-19 outcome by age

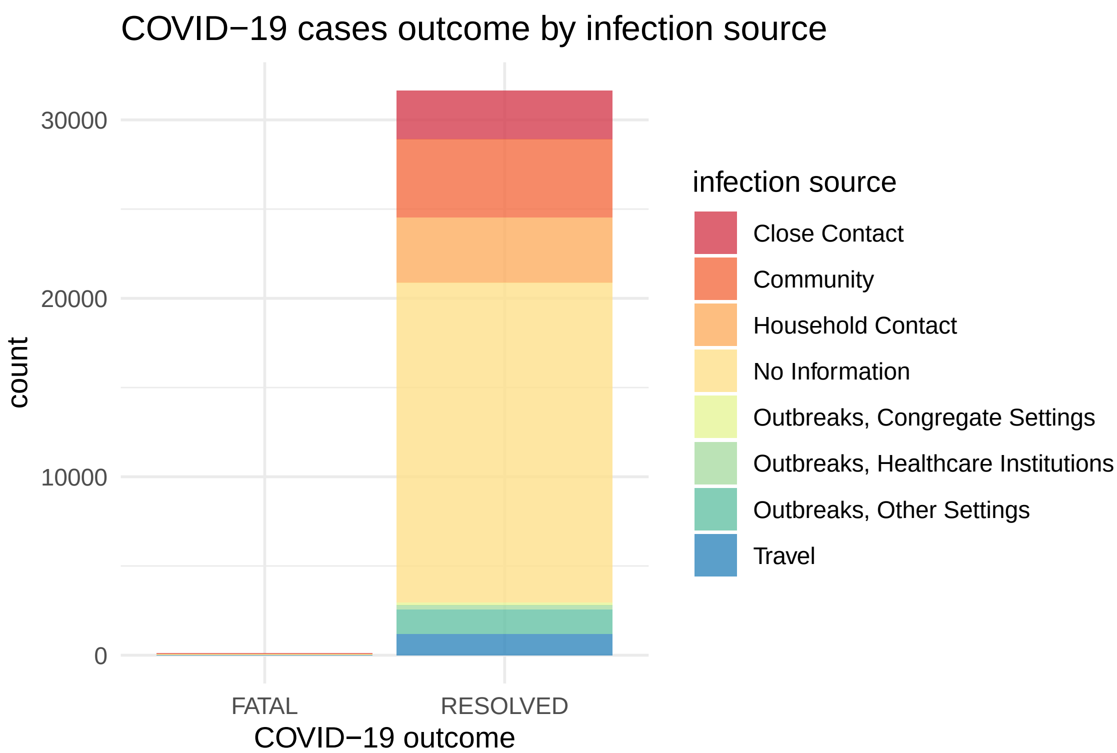

Figure 6: COVID-19 outcome by infection

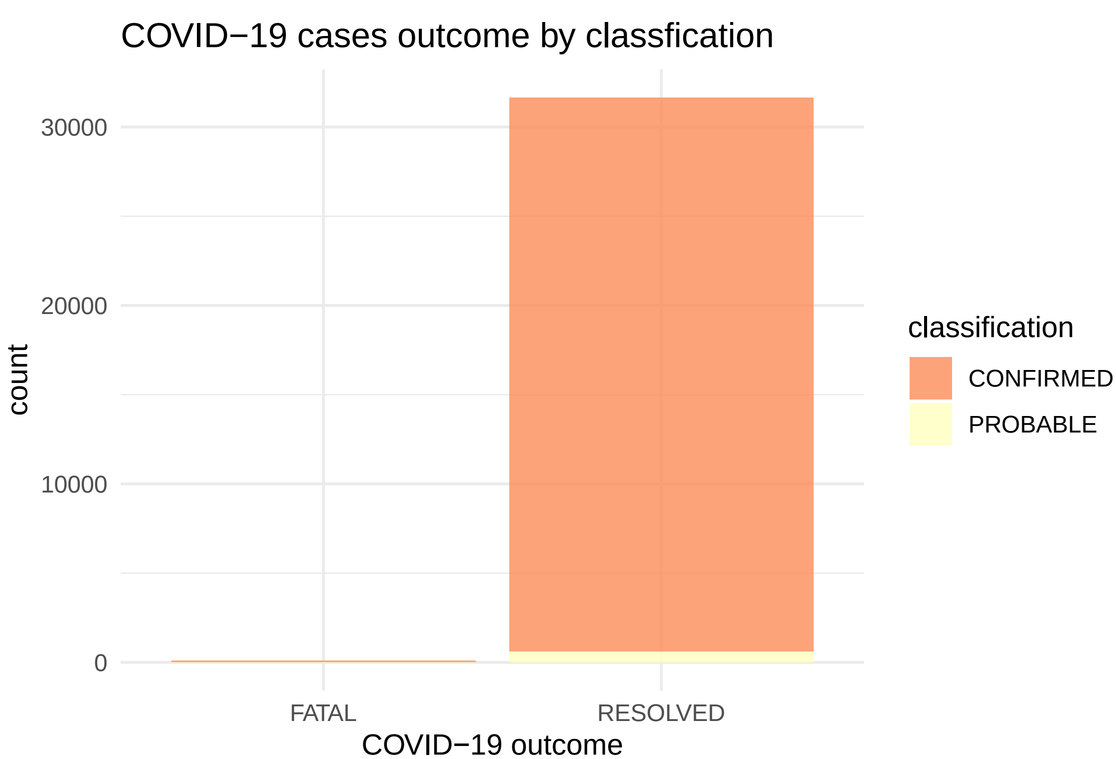

Figure 7: COVID-19 outcome by classification

3 Model

$$Y_i \sim Binomial(N_i, p_i)$$
$$log(\frac{p_i}{1-p_i}) = X_i\beta$$

The function follows the Binomial distribution, where β_0 is the log odds for X = 0, β_1 is the log odds ratio comparing X = 0 and X = 1. We fitted our model as shown.

$$\begin{aligned} log(\frac{p\_resolved}{1-p\_resolved}) = \beta_0 + \beta_1 age19andyounger + \beta_2 age20to29Years + \beta_3 age30to39Years \\ +\beta_4 age40to49Years + \beta_5 age50to59Years + \beta_6 age60to69Years \\ +\beta_7 age70to79Years + \beta_8 age80to89Years + \beta_9 age90andolder \\ +\beta_1 0genderMALE + \beta_1 1outbreakassociatedSporadic + \beta_1 2infectionCommunity \\ +\beta_1 3infectionHouseholdContact + \beta_1 4infectionNoInformation + \beta_1 5infectionOutbreaks, CongregateSettings \\ +\beta_1 6infectionOutbreaks, HealthcareInstitutions + \beta_1 7infectionOutbreaks, OtherSettings \\ +\beta_1 8infectionPending + \beta_1 9infectionTravel + \beta_2 0classificationPROBABLE \\ +\beta_2 1everhospitalizedYes + \beta_2 2everinICUYes + \beta_2 3everintubatedYes \end{aligned}$$

The intercept β_0 represents the log odds of the COVID-19 cases being resolved. β_2 to β_9 represent the log odds of the resolved COVID-19 cases when age is getting older and keeping the other variables constant. $\beta_1 0$ represents the log odds being the resolved cases when the gender is changed from female to male. $\beta_1 1$ is the log off being the resolved cases when outbreak associated changes to sporadic. $\beta_1 2$, $\beta_1 3$, $\beta_1 4$, $\beta_1 5$, $\beta_1 6$, $\beta_1 7$, $\beta_1 8$, $\beta_1 9$ represents the log odds being the resolved cases when the source of infection changes from close contact to the community, household contact, no information, congregate settings, healthcare institutions, other settings, pending, travel, respectively. $\beta_2 0$ is the log odds of resolved cases when the case is changed from confirmed to probable. Similarly, $beta_2 1$, $beta_2 2$, $beta_2 3$ means log odds being the resolved cases when they are ever hospitalized, ever in ICU, ever intubated changed from no to yes.

4 Results

$$\begin{aligned} log(\frac{p\_resolved}{1-p\_resolved}) = 23.07 - 0.55age19andyounger - 0.42age20to29Years - 15.43age30to39Years \\ -16.16age40to49Years - 16.94age50to59Years - 17.40age60to69Years \\ -18.40age70to79Years - 19.47age80to89Years - 20.26age90andolder \\ -0.34genderMALE + 0.22outbreakassociatedSporadic + 0.19infectionCommunity \\ +0.40infectionHouseholdContact + 0.24infectionNoInformation + 1.26infectionOutbreaks, CongregateSettings \\ -0.30infectionOutbreaks, HealthcareInstitutions + 0.07infectionOutbreaks, OtherSettings \\ +18.73infectionPending + 1.77infectionTravel - 1.12classificationPROBABLE \\ -2.31everhospitalizedYes - 1.70everinICUYes - 2.07everintubatedYes \end{aligned}$$

Overall, we observed that the log odds for the recovery of COVID-19 cases is 23.07. The log odds of the resolved cases are negatively correlated to the age groups, which are -0.55, -0.42, -15.43, -16.16, -16.94, -17.40, -18.40, -19.47, and -20.26, respectively. The recovery decreases and the fatality rate increases as the age gets older. When the gender switches from female to male, we found the log odds being resolved

Table 2: Summary table of cofficients of COVID-19 cases's result model

term	estimate	std.error	statistic	p.value
(Intercept)	22.42	3.2e+04	0.00	1.00
age19 and younger	0.02	3.2e+04	0.00	1.00
age20 to 29 Years	0.10	3.2e+04	0.00	1.00
age30 to 39 Years	-14.90	3.2e+04	0.00	1.00
age40 to 49 Years	-15.61	3.2e+04	0.00	1.00
age50 to 59 Years	-15.98	3.2e+04	0.00	1.00
age60 to 69 Years	-16.98	3.2e+04	0.00	1.00
age70 to 79 Years	-18.34	3.2e+04	0.00	1.00
age80 to 89 Years	-19.74	3.2e+04	0.00	1.00
age90 and older	-19.84	3.2e+04	0.00	1.00
genderMALE	-0.17	2.6e-01	-0.66	0.51
‘outbreak associated‘Sporadic	0.97	1.2e+00	0.83	0.41
infectionCommunity	0.48	4.5e-01	1.06	0.29
infectionHousehold Contact	0.79	6.1e-01	1.29	0.20
infectionNo Information	0.52	4.7e-01	1.11	0.27
infectionOutbreaks, Congregate Settings	18.89	5.4e+03	0.00	1.00
infectionOutbreaks, Healthcare Institutions	0.45	1.2e+00	0.37	0.71
infectionOutbreaks, Other Settings	1.18	1.2e+00	0.97	0.33
infectionTravel	0.02	8.4e-01	0.02	0.98
classificationPROBABLE	-1.24	8.3e-01	-1.50	0.13
‘ever hospitalized‘Yes	-2.81	3.5e-01	-8.00	0.00
‘ever in ICU‘Yes	-1.66	4.9e-01	-3.42	0.00
‘ever intubated‘Yes	-3.29	5.7e-01	-5.75	0.00

decreases. Meaning male patients may have a higher fatality. Additionally, as outbreak-associated changes to sporadic, the log odds being resolved raised by 0.22. The log odds being recovery changes from close contact to the community, household contact, no information, congregate settings, other settings, pending, travel is positive. Meaning the log odds being resolved increases. However, we also noticed that the log odds being recovery drops when close contact with healthcare institutions. Last but not least, the log odds being recovery declined as classification changes to probable, ever hospitalized, ever in ICU, ever intubated changed from no to yes, respectively. The reason is obvious: the patients who ever hospitalized, ever stayed in ICU, or ever intubated, his case is more serious than others, which leads to a higher fatality.

5 Discussion

5.1 Findings

To answer our research question what are the factors that may affect the recovery of COVID-19 cases, we visualized the fatal and resolved cases grouped by various explanatory variables. Besides, we conducted the binomial regression model which is the part of generalized regression model, and came up with the following conclusion. Older people are less likely to be resolved from COVID-19. Males patients are less resolved compared to females. Sporadic cases tends to be resolved. In addition, the sources of infection are positively related to the resolved cases excepts for healthcare institutions. Probable cases are more likely to be resolved. The cases of ever hospitalized/ever in ICU/ever intubated are less likely to be resolved.

5.2 Limitations and Weaknesses

5.2.1 Only including Females and Male in Gender

The original dataset included several gender options based on the self-reported from patients: female, male, non-binary, transgender, and unknown. However, our analysis only considered the cases of both females and males. For the reason that females and males are associated with the most cases. However, non-binary, transgender, and unknown patients cases excluded from the model may cause bias in our results.

5.2.2 The P-values of the Age Groups

In the summary table of coefficients of the model of the COVID-19 case, we saw the p-values of different age groups are approximately 1. The p-values close to 1 represent no difference between the age groups other than due to chance. Nevertheless, some people argued that the p-value maybe not be that important because it does not provide a good measure in a model. We still recommend we need a further investigation to explain why the p-values are small or exclude the variable of age in our model.

5.2.3 Dataset Consisted of Different Sources

As TPH mentioned, they extracted the data from different time periods and different sources (Health 2022). Thus, our analysis conclusions may be entirely different from other research papers. For instance, we estimated that males cases are less likely to be resolved compared to females. But other research proved that the COVID-19 cases fatality is higher in females than males (Dehingia and Ra 2021).

5.2.4 Model Fitting

Applying more complicated models may crush the R studio due to the large dataset and also many variables were included in our model. So we fitted the binomial regression model without the interaction terms. Consequently, we cannot consider all the possible situations.

Appendix

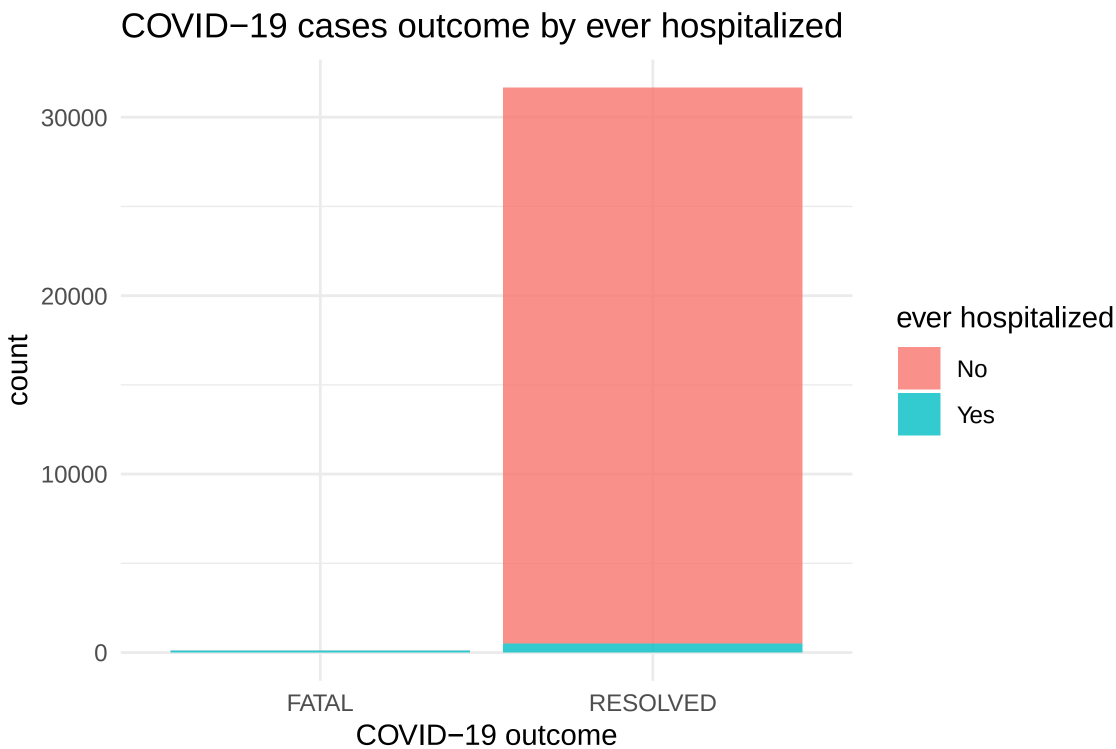

Figure 8: COVID-19 outcome by ever hospitalized

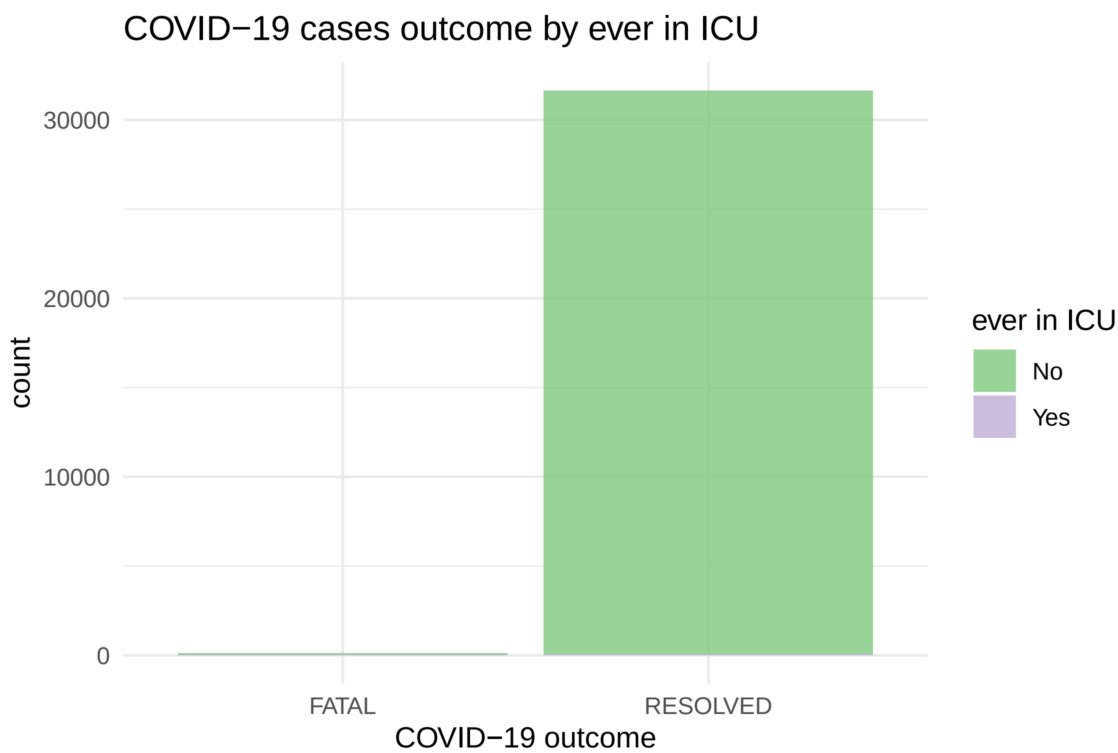

Figure 9: COVID-19 outcome by ever in ICU

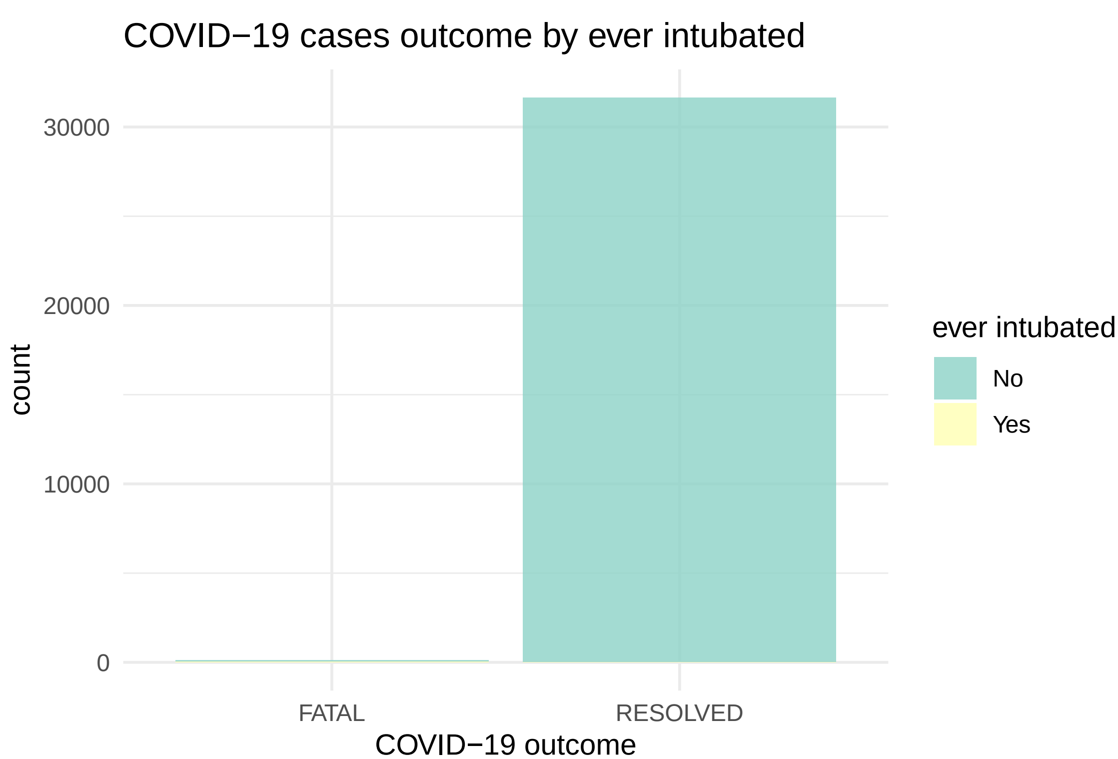

Figure 10: COVID-19 outcome by ever intubated

References

Adam, David. 2022. *The Pandemic's True Death Toll:Millions More Than Official Counts*. https://www.nature.com/articles/d41586-022-00104-8.

Dehingia, Nabamallika, and Anita Ra. 2021. "Sex Differences in Covid-19 Case Fatality: Do We Know Enough?" 9 (1). https://doi.org/10.1016/s2214-109x(20)30464-2.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://sharlagelfand.github.io/opendatatoronto/,%20https://github.com/sharlagelfand/opendatatoronto/.

Health, Toronto Public. 2022. *City of Toronto Open Data Portal*. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Silge, Julia, and David Robinson. 2016. "Tidytext: Text Mining and Analysis Using Tidy Data Principles in R." *JOSS* 1 (3). https://doi.org/10.21105/joss.00037.

Toronto, City of. 2019. *Toronto Public Health*. https://www.toronto.ca/city-government/accountability-operations-customer-service/city-administration/staff-directory-divisions-and-customer-service/toronto-public-health/.

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.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in R*. https://yihui.org/knitr/.