Risk of mortality from infecting COVID-19 is age sensitive and is likely to be gender sensitive*

Ziyu Jiang

2022/4/27

Abstract

This report presents an analysis of the City of Toronto's data on the outcomes of the cases of infection of Toronto residents. The results indicate that patients' age and gender affect their risk of mortality from infecting COVID-19. The analysis is conducted with logistic regression model created with the statistical programming language R. The results reinforce our understanding of the probable variables related to the consequencies of COVID-19 infections, which provides new probable ideas of preventing COVID-19 infections from the social perspective.

Keywords: COVID-19, COVID-19 cases Toronto, COVID-19 risk of mortality, COVID-19 gender impacts, COVID-19 age impacts, toronto residents, open data toronto

Contents

*Code and data are available in this GitHub repository: Owlll11/Analysis-about-Covid-19-cases.

1 Introduction

During the two years of COVID-19 pandemic, people are getting more and more worried about the consequences of infecting COVID-19. It was being said that the COVID-19 pandemic cases happened on children with a markedly low proportion than people in other age groups (Davies et al. 2020). At the same time, it was said that there are increased number of cases and higher risk of severe disease for people with higher ages for infecting COVID-19 (Davies et al. 2020). These suggests that people with higher age should pay more attention on preventing COVID-19 infection. Now a similar question has occurred after the conclusion of age sensitive for COVID-19: Would COVID-19 also be gender sensitive? Gender-related factors may cause differences for an individual's likelihood of exposure to COVID-19, but they may also influence whether an individual tries to acquire a test and whether they are given one (Tadiri et al. 2020). Analyzing this topic would help people have some ideas about society impacts on COVID-19 infection while they tried to have prevention from it.

After analyzing the data in the certain period of time, we can see that the patients' age and gender would have some effects on the symptoms and final results from infecting COVID-19. In this report, we will use the COVID-19 Cases in Toronto data from Toronto Open Data Portal (Gelfand 2020) to investigate the relationship between the final results of infecting COVID-19 and other independent variables, such as age and gender. We will discuss the limitation of the data and the potential bias from the data, apply the logistic regression model on the data. At last, we will discuss the distributions of the independent variables in interests, the resulting influences of the final model, the weakness of our researching processes and the lessons learned from the research and the possible improvement in future.

2 Data

R statistical programming language(R Core Team 2021) is used for analyzing, and the package `tidyverse`(Wickham et al. 2019) is used for data visualizing and data manipulating in this project. The package `broom`(Robinson, Hayes, and Couch 2022) is also used in this project to convert the statistical objects into tidy tibbles, the package `knitr`(Xie 2014) is used to knit the R markdown file to pdf form and ro create the tables, to modify the tables, the package `kableExtra` (Zhu 2021) is also used in this project, the package `ggplot2`(Wickham 2016) is used for creating graphs, and the package `ggpubr`(Kassambara 2020) is used for combining several plots into one plot. The packages `dplyr`(Wickham et al. 2022) and `janitor`(Firke 2021) are used to proceed the data cleaning processes. Finally, the package `car`(Fox and Weisberg 2019) is used to creating regression models of the variables.

The data is available to the public through the Open Data Toronto Portal hosted by the City of Toronto. The data was last refreshed on April 20, 2022, and is listed in the "Health" catalogue published by the Toronto Public Health. The COVID-19 cases in Toronto data were collected by the provincial Case & Contact Management System (CCM), a central data repository for COVID-19 case and contact management, and reporting in Ontario. According to the source, the data contained "demographic, geographic, and severity information for all confirmed and probable cases reported to Toronto Public Health" since January 2020(Toronto 2022).

Since this data contains all of the COVID-19 cases which happened in Toronto and were reported to the Toronto Public Health, it is not able to represent all cases of COVID-19 happened in Toronto. The data contains not only the confirmed COVID-19 cases, but also the probable COVID-19 cases reported, which may cause inaccuracy in the analysis processes. However, these bias created by the situation would not impact the project a lot, since we believe the main part of population which is reported to the Toronto Public Health is a good representative of the population in this report. Since we are still experiencing the COVID-19 pandemic, the dataset is subjected to change as future cases report and continuous quality improvement. More precisely, the data will be "completely refreshed and overwritten on a weekly basis, extracted at 8:30 AM on the Tuesday of a given week, and posted on the Wednesday"(Toronto 2022). This report will only use the version updated on April 20, 2022 to conduct analysis.

Table 1: A preview of the COVID-19 cases in Toronto dataset.

Age Group	FSA	Source of Infection	Classification	Client Gender	Outcome	Ever Hospitalized
30 to 39 Years	M2L	Close Contact	PROBABLE	FEMALE	RESOLVED	No
60 to 69 Years	M1E	Outbreaks, Healthcare Institutions	CONFIRMED	MALE	RESOLVED	No
60 to 69 Years	M5V	Community	CONFIRMED	MALE	FATAL	Yes
20 to 29 Years	M5V	Community	CONFIRMED	MALE	RESOLVED	No
90 and older	M2K	Outbreaks, Healthcare Institutions	CONFIRMED	FEMALE	RESOLVED	No
90 and older	M1E	Outbreaks, Healthcare Institutions	CONFIRMED	MALE	FATAL	No
30 to 39 Years	M5G	Outbreaks, Other Settings	CONFIRMED	MALE	RESOLVED	No

There were 295104 observations in the raw dataset and there were 7581 observations contain missing values. The data has 18 variables which were unique row ID for database, assigned ID for cases, Outbreak associated, age group, neighborhood name, FSA, source of infection, classification, episode date, reported date, client gender, outcome, currently hospitalized, currently in ICU, currently intubated, ever hospitalized, ever in ICU, ever intubated. This dataset contains excessive different topics of variables, it would be simpler for the analysis if cleaning applied. The starting of the dataset cleaning process is applying the `clean_names` function (Firke 2021) to clean the column names of the dataset. After that, the function `drop_na`(Wickham et al. 2022) was used to drop rows containing missing values since we got many of them. Then, to get a binary outcome(fatal or non-fatal) for all the cases, a new variable called `outcome_res` was created, which has value "1" when the value of outcome is "Fatal", and has value "0" when the value of outcome is "Resolved" or "Active". Also, a new variable called `symptom_levels` generated all the information from the last 6 variables into 4 symptom levels: the value is "1" when the client was never hospitalized during the infection; the value is "2" when the client was hospitalized but never in ICU during the infection; the value is "3" when the client was in ICU but never intubated during the infection; the value is "4" when the client was intubated during the infection. To get the information about if the episode date and the reported date for the cases were the same, a new variable called `date_diff` was created, which compared the values in the episode date and reported date variables; it has value "0" if the two dates were the same, and it has value "1" if the two date were different. At last, variables unrelated to the topic interested were removed from the dataset. A preview of a few columns and rows from the original dataset can be found in Table 1.

From the dataset, this paper focuses on the following: (1) age group, measured at the ordinal level; (2) source of infection, categorical; (3) client gender, categorical, (4) outcome, categorical, (5) symptom levels, an ordinal variable, (6) if outcome is fatal, nominal, (7) if the dates are different, nominal.

Figure 1 shown the plots of each of the variables mentioned above(4 and 6 are similar so only plotted 4). From figure 1, we can see that the case outcome of clients were mostly "RESOLVED", with tiny parts of "ACTIVE" and "FATAL"; the age group of clients were slightly right-skewed with main part on the left side; the most source of infection was "No Information", the second most was "Community", and the least was "Pending"; the client gender were mainly "MALE" and "FEMALE", with all other gender values; the symptom levels were mainly concentrated at level 1, and will small part of level 2, 3, 4; and lastly, mostly there are date differences between the episode date and the reported date in all cases.

3 Methodology

Because this study aims to find and verify whether gender and age have effects on risk of mortality from infecting COVID-19, a binary outcome, the model it relies upon is logistic regression. The probability of a particular outcome is linked to the probability distribution: $Pr(Y_i = 1|X_k) = \frac{e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}{1+e^{\beta_0+\beta_1 X_1+\beta_2 X_2}}$

The model estimates the probability of the outcome variable Y to be successful(y = 1) given the predictor variable(s) X_k. In this case, it explains the relationship between the dependent variable, binary outcome of the cases, and the independent variables age group, client gender, source of infection, symptom level, and date difference estimating the parameter values(β_k).

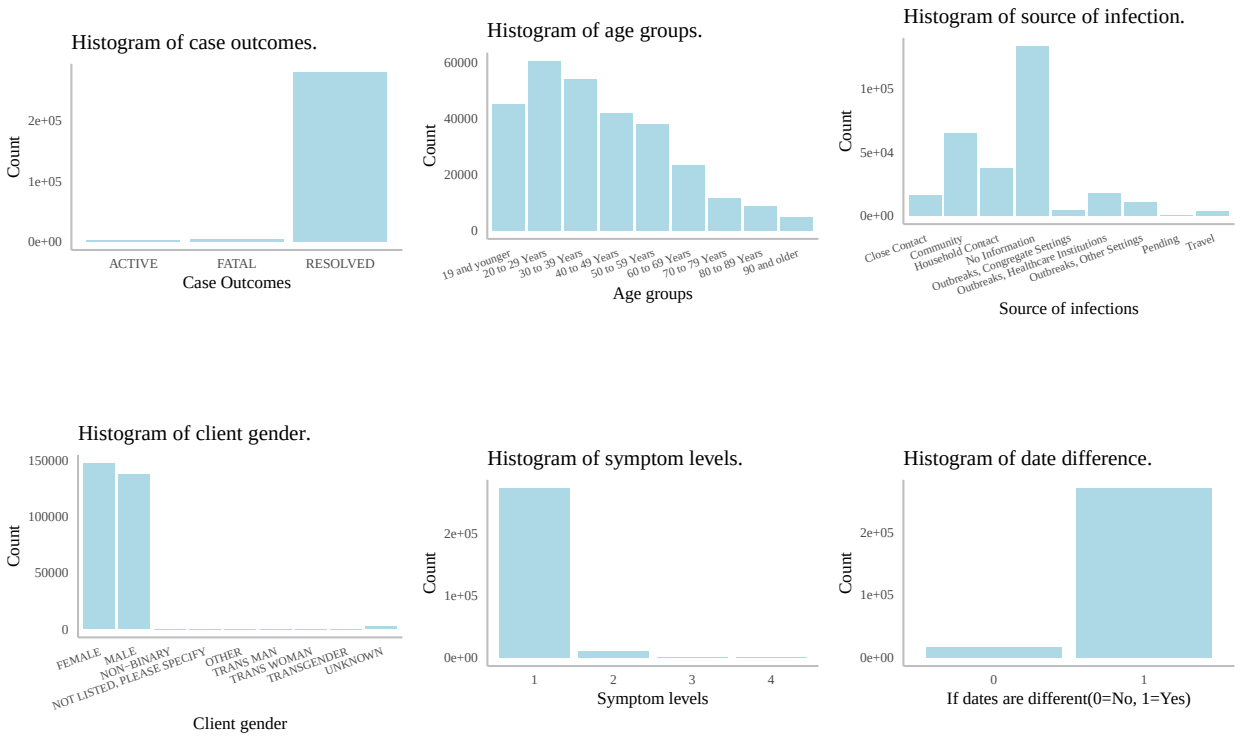

Figure 1: Histograms of the variables interested.

Logistic regression model was selected because it is easier to implement, interpret, and very efficient to train; it not only provides a measure of how appropriate a predictor(coefficient size)is, but also its direction of association (positive or negative); it can interpret model coefficients as indicators of feature importance. To use the logistic regression model, we need to satisfy the assumptions. Two of them can be easily verified for this dataset: The response variable is binary and the observations must be independent of one another, since the dependent variable is the binary outcome of the cases(Fatal or Non-fatal), and each observation in this dataset is an individual person in Toronto, which has relatively independence with other observations. However, the model also assumes linearity of independent variables and log odds, which requires careful considerations of variable inclusions. We need to consider which independent variables could significantly affect the outcome and include them in the model. In addition, the client gender variable was reduced to only have two values "Male" and "Female" in the following analysis and discussion. Because in the starting model, only gender "Male" has significant difference compare to the gender "Female"; reducing the gender variable would make the analysis process simpler and easier to investigate the topic in interest, which is the relationship between client gender and the binary outcome of the cases. Logistic regression is also highly sensitive to abnormal values in data, so the observations who has source of infection as "Pending" were removed. The observations who has source of infection as "No Information" were considered to be removed, however, the number of observations has "No Information" source of infection is around 130 thousands, which is way too much to be removed from the dataset and to keep the data as representative as possible at the same time.

All results got from the models in this analysis might not be precise due to the data limitations and even slight mistaken decisions in the steps of analyzing. Because of this, we need to interpret the model results as the small findings and patterns only from the data itself. It is wrong to treat the model results as the truths to the research questions or the topic in interest.

4 Result

The reference group for the starting model created for the analysis is 10 to 19 years old, female, infection from close contact, symptom level 1, episode date and reported date are the same. The results indicate whether there are significant difference in the outcome between the reference group and other conditions. For age group and symptom levels, the coefficient estimate represents the effect of a one-unit increase in these groups - from "10 to 19 years" to "20 to 29 years," for example. For client gender, source of infection and date difference, the estimated coefficients indicate the effect between identifying as male or female or infected from one source versus another or the episode date and reported date same or not.

Table 2 is the summary table for the starting model, the coefficients in it are exponentiated and indicate the odds ratio rather than log odds. According to the starting model, client gender, age group, source of infection and symptom levels have a statistically significant(p value less than the significance level 0.05) effect on the likelihood of having a case outcome as "Fatal". And in the starting model, as compared to 10-19 years old clients, 30-39 years old clients have 2.11 times higher odds of having "Fatal" as the outcome of cases, 40-49 years old clients have 2.94 times higher odds of having "Fatal" outcome, 50-59 years old clients have 3.79 times higher odds of having "Fatal" outcome, 60-69 years old clients have 4.67 times higher odds of having "Fatal" outcome, 70-79 years old clients have 5.67 times higher odds of having "Fatal" outcome, 80-89 years old clients have 6.56 times higher odds of having "Fatal" outcome, lastly, 90 and older years old clients have 7.11 times higher odds of having "Fatal" outcome, all keeping other variables constant.

Also, in the starting model, as compared to female clients, male clients have 0.42 times higher odds of having "Fatal" outcome, keeping other variables constant; as compared to the clients had source of infection "Close Contact", the clients have source of infection "No Information" have 0.43 times lower odds of having "Fatal" outcome(which kind of make no sense but will be explained in the discussion section), the clients have source of infection "Outbreaks, Healthcare Institutions" have 1.30 times higher odds of having "Fatal" outcome, all keeping other variables constant. Finally, as compared to the clients never be hospitalized during the infection, the clients who were hospitalized but were not in ICU have 2.20 times higher odds of having "Fatal" outcome, the clients who were in ICU but were not intubated have 3.84 times higher odds of having

Table 2: Summary table for the starting model.

term	estimate	std.error	statistic	p.value
(Intercept)	-10.0165404	0.7174569	-13.9611728	0.0000000
as.factor(age_group)20 to 29 Years	0.7007186	0.8171719	0.8574924	0.3911728
as.factor(age_group)30 to 39 Years	2.1066209	0.7339286	2.8703350	0.0041004
as.factor(age_group)40 to 49 Years	2.9448991	0.7191286	4.0950939	0.0000422
as.factor(age_group)50 to 59 Years	3.7854716	0.7121630	5.3154565	0.0000001
as.factor(age_group)60 to 69 Years	4.6720215	0.7103185	6.5773618	0.0000000
as.factor(age_group)70 to 79 Years	5.6657578	0.7097085	7.9832186	0.0000000
as.factor(age_group)80 to 89 Years	6.5551764	0.7093503	9.2410989	0.0000000
as.factor(age_group)90 and older	7.1065204	0.7097483	10.0127331	0.0000000
as.factor(client_gender)MALE	0.4240370	0.0398899	10.6301819	0.0000000
as.factor(source_of_infection)Community	-0.0525082	0.1132277	-0.4637398	0.6428342
as.factor(source_of_infection)Household Contact	-0.2021091	0.1348482	-1.4987895	0.1339282
as.factor(source_of_infection)No Information	-0.4323524	0.1144083	-3.7790311	0.0001574
as.factor(source_of_infection)Outbreaks, Congregate Settings	-0.1073098	0.2402437	-0.4466706	0.6551129
as.factor(source_of_infection)Outbreaks, Healthcare Institutions	1.2996527	0.1119098	11.6133919	0.0000000
as.factor(source_of_infection)Outbreaks, Other Settings	-0.1758754	0.2024479	-0.8687437	0.3849873
as.factor(source_of_infection)Travel	-0.3447580	0.2586586	-1.3328689	0.1825748
as.factor(symptom_levels)2	2.1979320	0.0454735	48.3343481	0.0000000
as.factor(symptom_levels)3	3.8387316	0.0922929	41.5929135	0.0000000
as.factor(symptom_levels)4	5.2158656	0.0748836	69.6529878	0.0000000
as.factor(date_diff)1	-0.0995074	0.0657499	-1.5134244	0.1301719

"Fatal" outcome, and the clients who were intubated during the infection have 5.22 times higher odds of having "Fatal" outcome, all keeping other variables constant.

This paper was focused on answering two research questions: (1) How is age group related to the likelihood to have fatal outcome from infecting COVID-19? (2) How does gender affect the likelihood to have fatal outcome from infecting COVID-19?

The null hypothesis for the investigation were: (1) The factor age group does not have significant effects on the likelihood to have fatal outcome from infecting COVID-19, under the condition that other factors remain unchanged. (2) The factor gender does not have significant effects on the likelihood to have fatal outcome from infecting COVID-19, under the condition that other factors remain unchanged.

Based on the results from the logistic regression model and with the significance level 0.05, we have strong evidence to reject the null hypothesis that age does not affect the likelihood to have fatal outcome from infecting COVID-19. Also, we have strong evidence to reject the null hypothesis that gender do not affect the likelihood to have fatal outcome from infecting COVID-19.

5 Discussion

The results of the model only provide some patterns and information generated from the data, to get more life relative conclusions, we need to discuss the results in a larger background.

Table 3: Table of quantities of three case outcomes.

Case outcome	Count
ACTIVE	2803
FATAL	4021
RESOLVED	278436

5.1 Findings

As we can see, the date difference between the episode date and the reported date does not have significant effect on the likelihood of having a fatal case outcome. This is understandable because infecting COVID-19 is not lethal at majority of time, people were asked to be self-isolated when they were suspected exposed to COVID-19 or had COVID-19 symptoms. In the dataset, it is also being verified that majority of people did not have fatal case outcome from infecting COVID-19. Table 3 is a summary of number of clients who have each case outcome, we can see that the proportion of clients have fatal outcome is only $\frac{2803}{285260} = 0.98\%$. However, it is not saying that we can leave the COVID-19 with no worries, "COVID-19 symptoms can sometimes persist for months. The virus can damage the lungs, heart and brain, which increases the risk of long-term health problems"(Staff 2021). Therefore, the prevention to avoid infecting the COVID-19 is still very important to protect people's own health and public health.

5.1.1 Age group

It is being reported that age has kind of certainly significant effect on the likelihood to have fatal outcome from infecting COVID-19 and in this report we verified that. Figure 2 shows the quantities of the fatal and non-fatal outcomes for different age groups. We can clearly see that as the age group become larger and larger, the proportion of clients who have fatal outcome also become larger and larger in the total number of their age groups. It was being said that "adults over 65 years of age represent 80% of hospitalizations and have a 23-fold greater risk of death than those under 65"(Mueller, McNamara, and Sinclair 2020). From the model results we got before, each 10 years increasing of age would cause around 1 time increase of odds to have fatal outcome. For having effective actions preventing the infection of COVID-19, we must consider the age factor.

5.1.2 Gender

The factor gender is usually kind of uncertain thing when we analysis the relationship between gender and COVID-19 infection related outcome. Some people have thoughts that the physical qualities of males might overall be better than females', so they think if the likelihoods of having fatal outcome are different between male and female, it might be that male have less risk of having fatal outcome. However, the result of model told us that in this data, male clients have 0.42 times higher odds of having "Fatal" outcome, keeping other variables constant. Table 4 shown the case outcomes' quantities between gender; the female fatal rate is $\frac{1883}{147657} = 1.27\%$, and the male fatal rate is $\frac{2138}{137603} = 1.55\%$. It is not certain in all the reports about the relationship between gender and risk of fatal, what we got from the results is only a pattern from the data. But why would this situation happen in the COVID-19 cases in Toronto data?

One possible reason is that "differences in the expression of ACE2 caused by sex hormones may help in explaining the sex disparities in COVID-19 infection, severity, and fatality"(Mukherjee and Pahan 2021). This possible reason is kind of natural physical difference between male and female and have many dispute about it.

Another possible reason is that "gender may influence an individual's exposure to infectious disease through occupation in essential services, risk-taking behaviours and employment of precautions, or have an impact on an individual's ability to seek and receive testing and care, through norms for health-seeking behaviour,

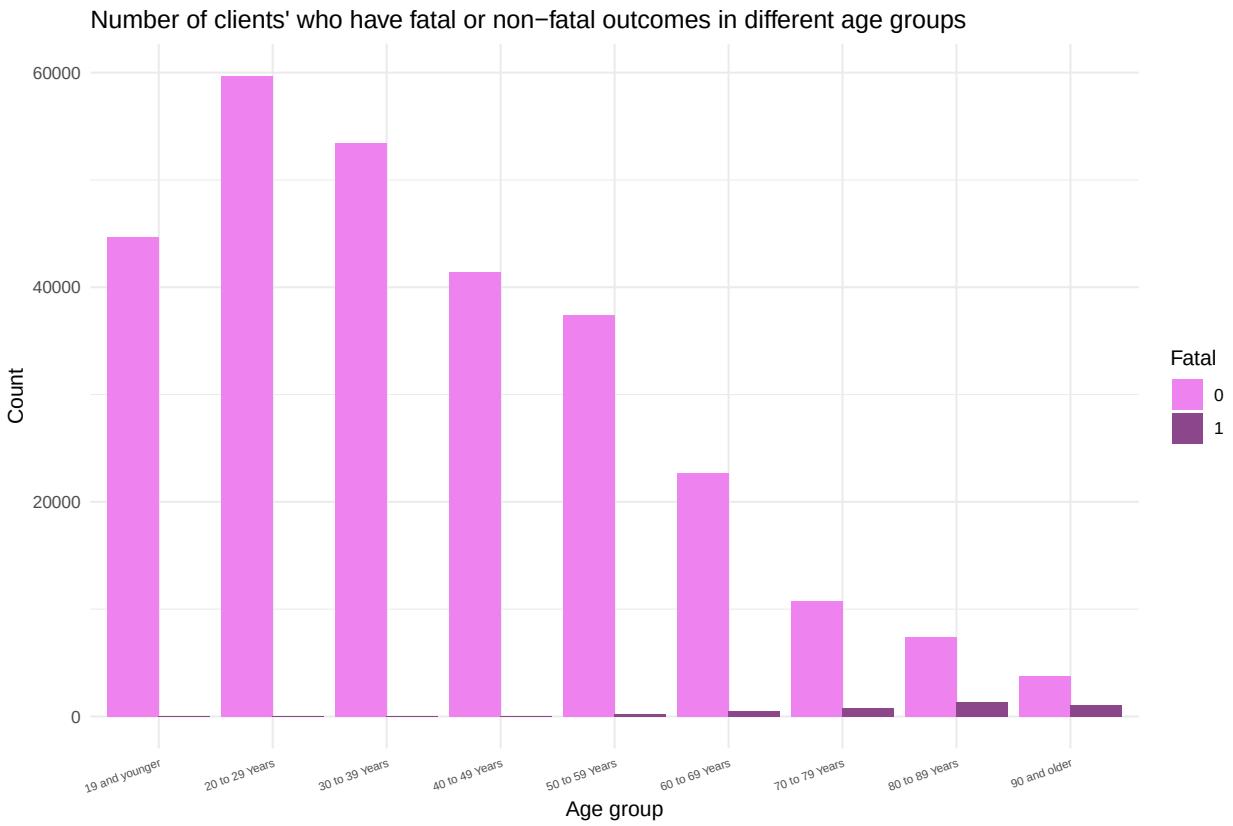

Figure 2: Histogram of number of clients who have fatal or non-fatal case outcomes in different age groups.

Table 4: Table of case outcomes quantities between gender

client_gender	outcome_type	n
FEMALE	Fatal	1883
FEMALE	Non-Fatal	145774
MALE	Fatal	2138
MALE	Non-Fatal	135465

responsibilities at home or work, reduced availability of health services or institutional biases and policies"(Tadiri et al. 2020). This possible reason is kind of social construct difference between male and female, which is a concept to understand and connect with some social concepts. As some additional references, figure 3 shown the proportions of different source of infections between gender and table 5 and 6 shown the numbers of male and female clients having the four symptom levels.

Overall, gender is also an important factor to be consider and included in the COVID-19 relative analysis with both physical and social differences between gender.

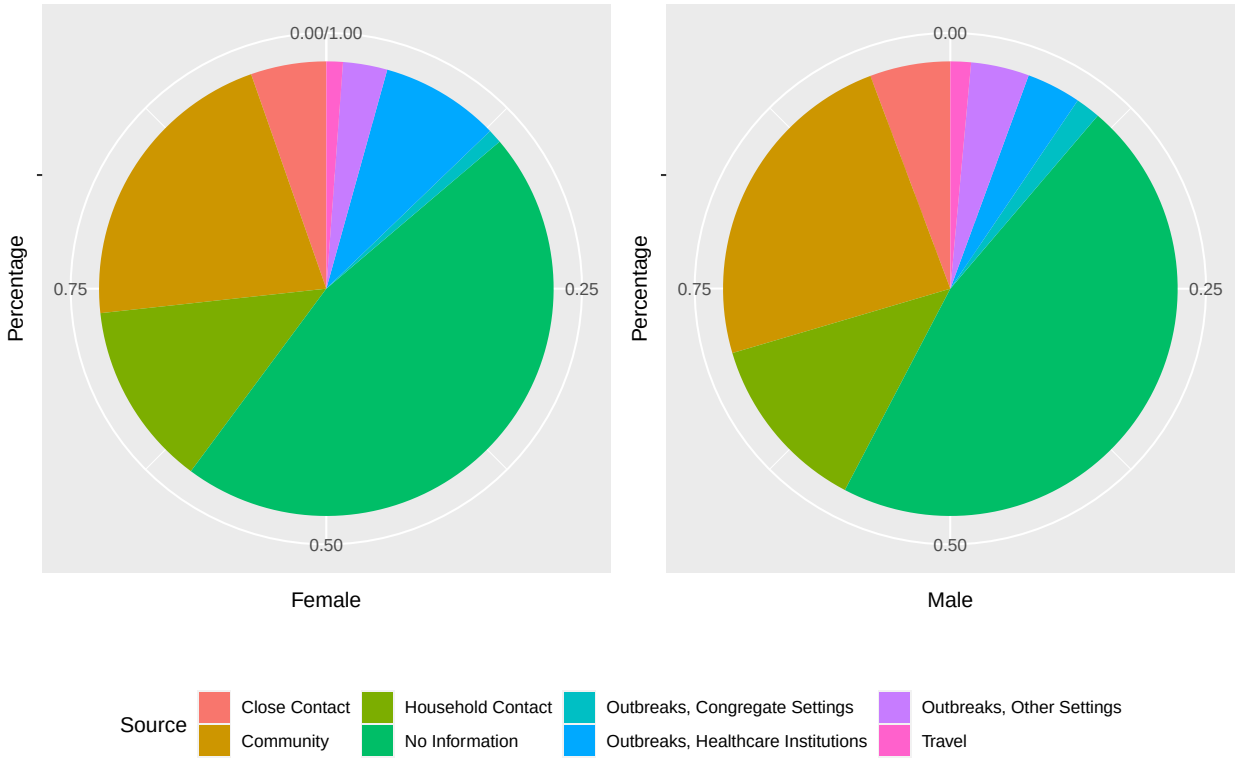

Figure 3: Pie charts of proportions of different source of infections between male and female.

Table 5: Table of numbers of female clients having the four symptom levels.

Symptom_levels	n
1	141518
2	5190
3	423
4	559

Table 6: Table of numbers of male clients having the four symptom levels.

Symptom_levels	n
1	130293
2	5717
3	637
4	977

5.2 Limitations

As mentioned in the former section, there are some limitations that directly impact the data. The variable client gender originally had 9 variables and was reduced to only has male and female in this paper. The gender variable in the original data is representative since it has comprehensive gender types, removing the other genders except male and female in this paper is only for convenience of analyzing the topic of interest and it is not rigorous to do that.

The source of infection variable were also been modified to remove the observations with value "Pending", since the logistic regression model is highly sensitive to abnormal values in data. The value "No Information" were also considered to be removed, however, it is not safe to do so since the number of clients who has source of infection as "No information" is too much(more than 1/3 in the whole population). Compare to the number of "No Information", "Pending" has really small quantities in the total number of population in this data. This leads to the occurrence of "No Information" as one of the term in the logistic regression model; although the coefficients are kind of non-sense to be interpret, these observations still need to be within the model to make sure the accuracy.

Appendix

Datasheet

Motivation

1. *For what purpose was the dataset created? Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.*
   - The dataset was created to enable analysis of the COVID-19 cases in Toronto for the public.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Toronto Public Health
3. *Who funded the creation of the dataset? If there is an associated grant, please provide the name of the grantor and the grant name and number.*
   - Toronto Public Health
4. *Any other comments?*
   - None

Composition

1. *What do the instances that comprise the dataset represent (for example, documents, photos, people, countries)? Are there multiple types of instances (for example, movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*
   - COVID-19 cases, Toronto, and clients' information when infecting COVID-19.
2. *How many instances are there in total (of each type, if appropriate)?*

   295104.
3. *Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set? If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (for example, geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (for example, to cover a more diverse range of instances, because instances were withheld or unavailable).*
   - It is a sample of instancec from a larger set. The larger set could be all cases of COVID-19 happened in Toronto.
4. *What data does each instance consist of? "Raw" data (for example, unprocessed text or images) or features? In either case, please provide a description.*
   - Each instance consist of features, those were the case information for each cases.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Reported COVID-19 cases in Toronto.
6. *Is any information missing from individual instances? If so, please provide a description, explaining why this information is missing (for example, because it was unavailable). This does not include intentionally removed information, but might include, for example, redacted text.*
   - Yes, some of the instances contained certain missing values.
7. *Are relationships between individual instances made explicit (for example, users' movie ratings, social network links)? If so, please describe how these relationships are made explicit.*

- They are the reported COVID-19 cases in Toronto.

8. *Are there recommended data splits (for example, training, development/validation, testing)? If so, please provide a description of these splits, explaining the rationale behind them.*
   - No.
9. *Are there any errors, sources of noise, or redundancies in the dataset? If so, please provide a description.*
   - No.
10. *Is the dataset self-contained, or does it link to or otherwise rely on external resources (for example, websites, tweets, other datasets)? If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (that is, including the external resources as they existed at the time the dataset was created); c) are there any restrictions (for example, licenses, fees) associated with any of the external resources that might apply to a dataset consumer? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*
    - It is self-contained.
11. *Does the dataset contain data that might be considered confidential (for example, data that is protected by legal privilege or by doctor-patient confidentiality, data that includes the content of individuals' non-public communications)? If so, please provide a description.*
    - No.
12. *Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety? If so, please describe why.*
    - No.
13. *Does the dataset identify any sub-populations (for example, by age, gender)? If so, please describe how these subpopulations are identified and provide a description of their respective distributions within the dataset.*
    - Yes, all types of gender were used in this data.
14. *Is it possible to identify individuals (that is, one or more natural persons), either directly or indirectly (that is, in combination with other data) from the dataset? If so, please describe how.*
    - No.
15. *Does the dataset contain data that might be considered sensitive in any way (for example, data that reveals race or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)? If so, please provide a description.*
    - No.
16. *Any other comments?*
    - None.

Collection process

1. *How was the data associated with each instance acquired? Was the data directly observable (for example, raw text, movie ratings), reported by subjects (for example, survey responses), or indirectly inferred/derived from other data (for example, part-of-speech tags, model-based guesses for age or language)? If the data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- The data were extracted from the provincial Case & Contact Management System (CCM), which all the cases in the data were the cases reported to Toronto Public Health.

2. *What mechanisms or procedures were used to collect the data (for example, hardware apparatuses or sensors, manual human curation, software programs, software APIs)? How were these mechanisms or procedures validated?*
   - Manual human curation and probably also software programs.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
   - It could be seen as a subset of all actual COVID-19 cases happened in Toronto, it was collected totally depends on if the case was reported to the Toronto Public Health.
4. *Who was involved in the data collection process (for example, students, crowdworkers, contractors) and how were they compensated (for example, how much were crowdworkers paid)?*
   - The provincial Case & Contact Management System (CCM), don't know how were they compensated as this is a provincial institution.
5. *Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (for example, recent crawl of old news articles)? If not, please describe the timeframe in which the data associated with the instances was created.*
   - From January 2020 till now, with the last updated date April 20, 2022.
6. *Were any ethical review processes conducted (for example, by an institutional review board)? If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*
   - No
7. *Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (for example, websites)?*
   - I collected the data from the website of Toronto Open Data Portal.
8. *Were the individuals in question notified about the data collection? If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*
   - Not collected from individuals.
9. *Did the individuals in question consent to the collection and use of their data? If so, please describe (or show with screenshots or other information) how consent was requested and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.*
   - Not collected from individuals. The data was published in the Toronto Open Data Portal.
10. *If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses? If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*
    - Not collected from individuals.
11. *Has an analysis of the potential impact of the dataset and its use on data subjects (for example, a data protection impact analysis) been conducted? If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*
    - No.
12. *Any other comments?*
    - None.

Preprocessing/cleaning/labeling

1. *Was any preprocessing/cleaning/labeling of the data done (for example, discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal of instances, processing of missing values)? If so, please provide a description. If not, you may skip the remaining questions in this section.*
   - No.
2. *Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (for example, to support unanticipated future uses)? If so, please provide a link or other access point to the "raw" data.*
   - No.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
   - Do not know.
4. *Any other comments?*
   - None.

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
   - Yes, it had beed used for analyzing the relationships between the COVID-19 cases outcomes and other independent variables in interest.
2. *Is there a repository that links to any or all papers or systems that use the dataset? If so, please provide a link or other access point.*
   - Yes; https://github.com/Owlll11/Analysis-about-Covid-19-cases
3. *What (other) tasks could the dataset be used for?*
   - Analyze which factors could impact the symptom levels after infected COVID-19.
4. *Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses? For example, is there anything that a dataset consumer might need to know to avoid uses that could result in unfair treatment of individuals or groups (for example, stereotyping, quality of service issues) or other risks or harms (for example, legal risks, financial harms)? If so, please provide a description. Is there anything a dataset consumer could do to mitigate these risks or harms?*
   - No.
5. *Are there tasks for which the dataset should not be used? If so, please provide a description.*
   - Some tasks like generating private personal information.
6. *Any other comments?*
   - None.

Distribution

1. *Will the dataset be distributed to third parties outside of the entity (for example, company, institution, organization) on behalf of which the dataset was created? If so, please provide a description.*
   - Yes, it is open for public on the website of the Toronto Open Data Portal.

2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - Github, do not know.
3. *When will the dataset be distributed?*
    - April 27, 2022
4. *Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)? If so, please describe this license and/ or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*
    - Yes, it is under the Open Government Licence – Toronto, https://open.toronto.ca/open-data-license/
5. *Have any third parties imposed IP-based or other restrictions on the data associated with the instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*
    - No.
6. *Do any export controls or other regulatory restrictions apply to the dataset or to individual instances? If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*
    - No.
7. *Any other comments?*
    - None.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
    - Toronto Public Health
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
    - Via this email address: edau@toronto.ca
3. *Is there an erratum? If so, please provide a link or other access point.*
    - No.
4. *Will the dataset be updated (for example, to correct labeling errors, add new instances, delete instances)? If so, please describe how often, by whom, and how updates will be communicated to dataset consumers (for example, mailing list, GitHub)?*
    - Yes, the data will be completely refreshed and overwritten on a weekly basis, extracted at 8:30 AM on the Tuesday of a given week, and posted on the Wednesday.
5. *If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (for example, were the individuals in question told that their data would be retained for a fixed period of time and then deleted)? If so, please describe these limits and explain how they will be enforced.*
    - No.
6. *Will older versions of the dataset continue to be supported/hosted/maintained? If so, please describe how. If not, please describe how its obsolescence will be communicated to dataset consumers.*
    - Yes.

7. *If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so? If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to dataset consumers? If so, please provide a description.*
    - No.
8. *Any other comments?*
    - None.

References

Davies, Nicholas G., Petra Klepac, Yang Liu, Kiesha Prem, Mark Jit, CMMID COVID-19 working group, and Rosalind M. Eggo. 2020. *Age-Dependent Effects in the Transmission and Control of COVID-19 Epidemics*. https://www.nature.com/articles/s41591-020-0962-9.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Kassambara, Alboukadel. 2020. *Ggpubr: 'Ggplot2' Based Publication Ready Plots*. https://CRAN.R-project.org/package=ggpubr.

Mueller, Amber L., Maeve S. McNamara, and David A. Sinclair. 2020. *Why Does COVID-19 Disproportionately Affect Older People?* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7288963/.

Mukherjee, Shreya, and Kalipada Pahan. 2021. *Is COVID-19 Gender-Sensitive?* https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7786186/.

R Core Team. 2021. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David, Alex Hayes, and Simon Couch. 2022. *Broom: Convert Statistical Objects into Tidy Tibbles*. https://CRAN.R-project.org/package=broom.

Staff, Mayo Clinic. 2021. *COVID-19 (Coronavirus): Long-Term Effects*. https://www.mayoclinic.org/diseases-conditions/coronavirus/in-depth/coronavirus-long-term-effects/art-20490351.

Tadiri, Christina P., Teresa Gisinger, Alexandra Kautzky-Willer, Karolina Kublickiene, Maria Trinidad Herrero, Valeria Raparelli, Louise Pilote, and Colleen M. Norris. 2020. *The Influence of Sex and Gender Domains on COVID-19 Cases and Mortality*. https://www.cmaj.ca/content/192/36/E1041.

Toronto, City of. 2022. *About COVID-19 Cases in Toronto*. Toronto Public Health. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

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. 2022. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Xie, Yihui. 2014. "Knitr: A Comprehensive Tool for Reproducible Research in R." In *Implementing Reproducible Computational Research*, edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.