Impacts from Covid-19*

Characteristics of people infected by Covid-19 in Toronto since 2020

Qiuhan Wang

26 April 2022

Abstract

Covid-19 is considered a highly contagious and hard to prevent virus and it breakouts in 2019, which profoundly influenced global economics and peoples' lives. The study is based on the Toronto covid-19 dataset currated by the Provincial Case & Contact Management System. To help the government and citizens control the spread of the virus, I analyzed the characteristics of people who are more likely to be infected and their major sources of infection. I find that covid-19 can infect people of any age, and their main infected source is traveling, and it brings more severe results to the elder.

1 Introduction

Covid-19 is an infectious disease caused by the SARS-Cov-2 virus. The virus can spread from an infected person's mouth or nose in small liquid particles when they cough, speak or breathe. It has been an outbreak since 2019, which influences both global economic systems and peoples' lives.(Lazarus 2020) According to the statistics, over 479 million people have been infected by the virus, and 6 million people died because of Covid-19.(Lupton & K. Willis 2021) In addition, Covid-19-related lockdowns were very common during 2020-2021, which impacted economic activities directly. Most nations experienced the most significant hit to their gross domestic product(GDP) growth.

Measures that can effectively control the spread of the virus and prevent covid-19 are largely unkown. In the paper, I aim to help the government on publishing the measures to control the virus spread and prevent covid-19. I will summarize the characteristics of the patients who have been confirmed or probably infected by the virus in Toronto since January in 2020 and find the impact of the virus on the healthy condition of a different group of people.

The data was extracted from the provincial Case & Contact Management System(CCM). This dataset contains demographic, geographic, and severity information for all confirmed and probable cases reported to and managed by Toronto Public Health since January 2020. The data include the cases that occur in the community and outbreaks. The results conclude that most patients can recover without any special treatment, but older people are more likely to develop the severe illness than young persons. Also, we find that the primary source of infection is traveling, and their outbreak-associated is sporadic.

The article is constructed as follows. First, I introduce the dataset, including data resources, data collection methods, and variables in data. Second, I divide people who are infected into different groups based on their age and gender to find who is more likely to be infected. Also, I discuss the sources of infection and patients' state of illness. Finally, some suggestions are provided to the government and people on covid-19 prevention.

*Code and data are available at: https://github.com/wangq166/Covid-19-Research.git

2 Data

2.1 Data collection

All the data used in this paper is downloaded from the website "open.toronto.ca". On the website, you can click open data portal home and then search the keyword "Covid" you can get the catalog called "About Covid-19 Cases in Toronto". Finally, press downloads the "COVID19 cases" dataset from the download data section. Compared with other data relating to covid-19 cases, all the information from the data is extracted from the provincial Case Contact Management System (CCM), which means that it is more informative and authoritative. This data contains demographic, geographic, and severity information for all confirmed and probable cases reported to Toronto Public Health since January 2020. Sporadic and outbreak-associated cases are also included.

2.2 Data processing

All analyses are done with a statistic programming language R (R Core Team 2020). I downloaded the dataset with 295104 observations from the website "opentoronto.ca". The variables in the data set include "Assigned_ID", "Outbreak.Associated", "Age_group", "Neighbourhood.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". I removed the missing value of patients' age using the R function "filter". There are 7319 missing values in Neighborhood.Name column and 3804 missing values in FSA column, but I don't study these variables in the paper.After deleting people whose age is unknown, I have 294842 observations remaining. R package "tidyverse"(Wickham et al. 2019) is used to manipulate the data and "ggplot2" (Wickham 2016) is used for data visualization. "lubridate" (Grolemund and Wickham 2011)is installed to build a linear regression model and logistic model to analyze the relationship among patients' gender, age and their health condition.

The statistical summary of all variables are in Appendix and some details are in the "data characteristics" section. I also made various plots on ages, gender, source of infection, accept medical treatment condition. (see them in the "Results" section)

2.3 Survey Method

By 2020, all the people in Toronto whose self-covid-19-test result is positive have the responsibility to report to Toronto Public health consciously. Additionally, the Toronto Public Health system will automatically record the covid-19 test results for people who have been in walk-in-clinic or similar medical institutions.

2.3.1 Population and sample

The population size is same as the sample size in this case. It's a census; it includes all the people infected by covid-19 of different gender and different age in Toronto since 2020. There are 295104 observations collected by the provincial Case & Contact Management System.

2.3.2 Strength

All the data are provided by the provincial Case & Contact Management System, thus they are informative and authoritative. Compared with the data from other resources, it records all reported confirmed cases or probable cases in Toronto accurately. Also, the data is completely refreshed and overwritten every week, which means that it updates quickly. It's a benefit to our research.

2.3.3 Weakness

Although the data records the Covid-29 cases as many as possible in Toronto in 2020, it still includes some blank or no information that may influence our results. For example, 261 patients don't provide their age

Table 1: The number of infected people in different age group

Var1	Freq
19 and younger	45838
20 to 29 Years	62478
30 to 39 Years	55723
40 to 49 Years	42774
50 to 59 Years	38615
60 to 69 Years	23808
70 to 79 Years	11836
80 to 89 Years	8871
90 and older	4899

Table 2: The number of infected people from different sources of infection

Var1	Freq
Close Contact	16213
Community	65698
Household Contact	37676
No Information	137959
Outbreaks, Congregate Settings	4410
Outbreaks, Healthcare Institutions	18248
Outbreaks, Other Settings	10673
Pending	59
Travel	3906

information, and over 40% of patients don't answer their resources of infection. We are unable to replace or delete these data because of the large amount, so we have a sampling error here.

2.4 Data characteristics

There are 294842 observations in our data and all the variables we used in the report are categorical variables, including gender, age, source of infection, reported date, ever/currently in IC, ever/currently hospitalized, ever/currently incubated. We can divide people in 9 groups according to age, such as "19 and younger","20 to 29 years old","30 to 39 years old" and "90 and older". Also, for the source of infection, we have 9 categories, including close contact, community, healthcare institution,congregate Settings, household contact, no information, pending, travel and other sitting.The range of reported date is from 2020-01-23 to 2022-03-22. We can see more details from table 1 and table 2.

3 Results

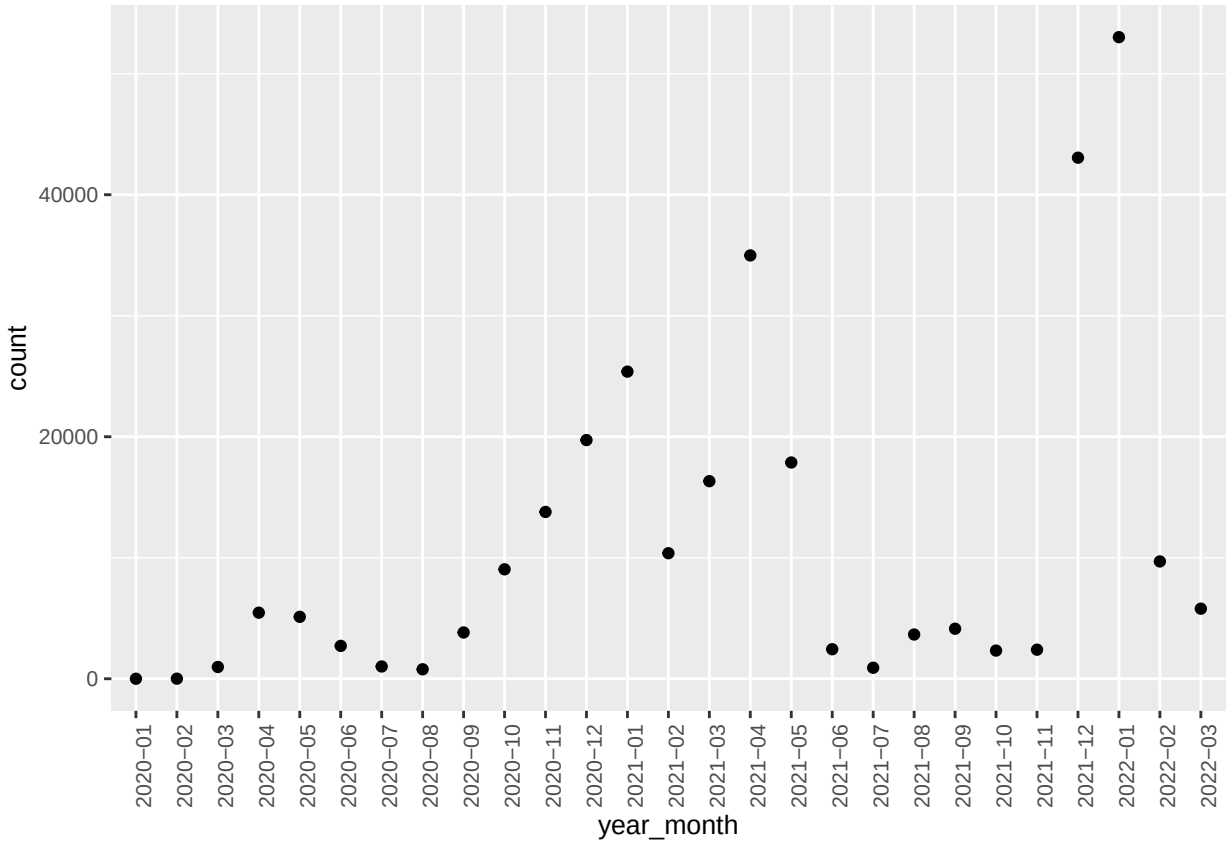

Figure 1: the number of confiremed case or porble cases each month

From Figure 1, I observe the number of confirmed cases and probable cases each month in Toronto from January 2020 to April 2022.There was a significant increasing trend from 5000 cases to 27000 cases between August 2020 and February 2021. Although the number of infected people decreased since June 2021 and remained below 5000 in the following months, I still had an explosive growth to 52000 cases in February 2022.

From Figure 2, I observe that the number of patients aged between 20 and 29 is above 60000, ranking the first. Also, I find that the number of patients aged between 30 and 39 is roughly 55000. The fatal rate increases as the increase of age, so the graph displays that about one of four people older than 90 cannot recover successfully. What's more, the active rate for people younger than 40 years old is higher than for those more senior than 40.

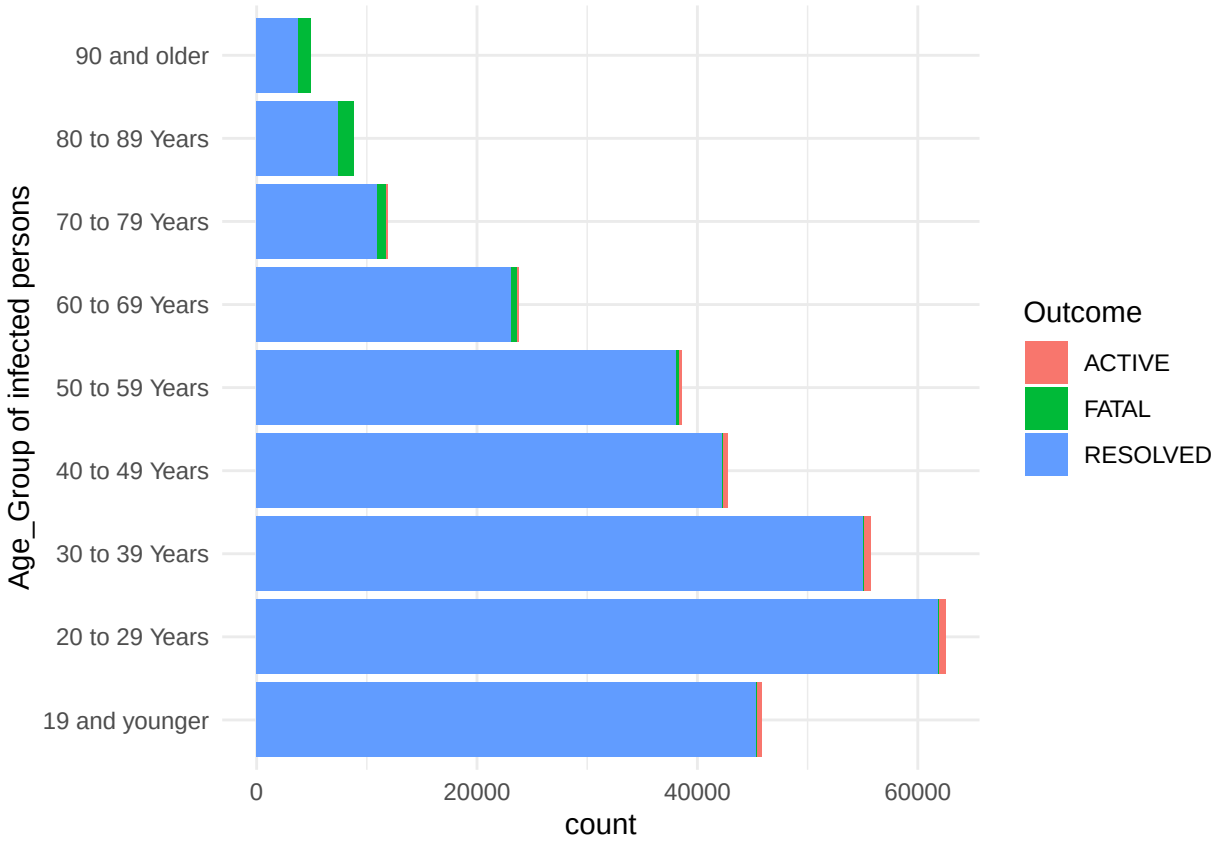

Figure 2: outcome for infected people in different age groups

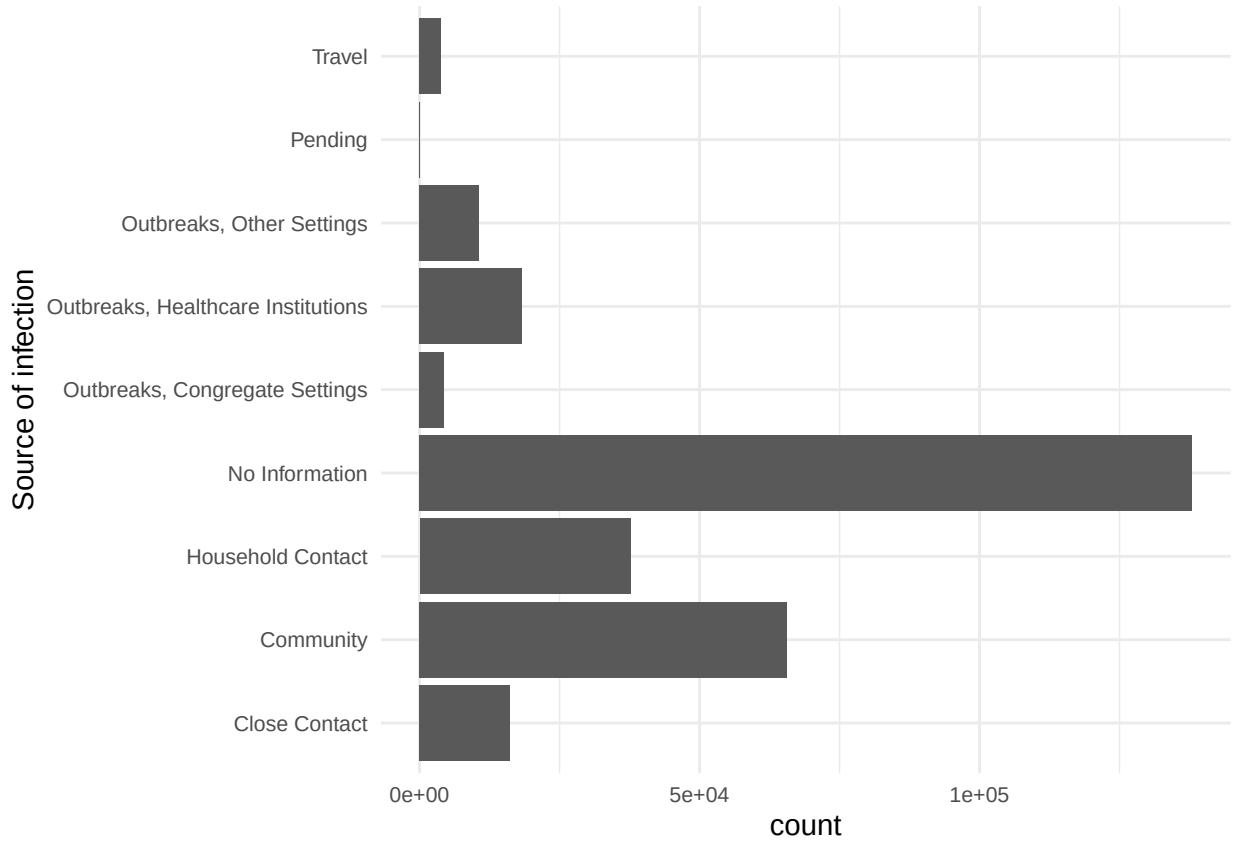

Figure 3: sources of infection distribution

From Figure 3, I find that patients can get covid-19 from various sources, including travel, close contact, community, and others. Besides those people who don't provide their information about the source of infection, community, household contact, and outbreaks health care institutions are considered three primary sources.

```
## < table of extent 0 >
```

From Figure 4, I find that 86.75% of cases are sporadic, and 13.25% are outbreak-associated. It means that sporadic cases are nearly seven times that of outbreak-associated cases.

The Table 3-5 describe the number of people who ever or currently in ICU, ever or currently hospitalized, ever or currently intubated in different age groups. From table1 and table3, overall, the proportion of infected people in each age group who accept ICU or intubated treatment is pretty low, lower than 5%. However, I can see that the proportion of people who have been hospitalized increases as the patient's age grows. The table2 points out that the proportion of people who are in the age group "19 and younger" and "20 to 29" and be hospitalized is lower than 1%, but the proportion of people who are in the age group "70 to 79", "80 to 89" and "90 and older" and be hospitalized reach 23%, 30%, and 24.7%.

4 Model

Because the response is binomial, it's better to fit a logistic regression to investigate the relationship of age, sex, and source of infections on the likelihood of being in ICU. According to the model summary (Table 6), I conclude that the males aged 70-89 who are infected by the community are more likely to accept treatment in ICU. I also find that the 70-79 age group and 80-89 age group have 62.8 and 78.4 fold of odds of being in ICU compared to the under 19 age group. Males have 1.77 fold of odds of being in ICU compared to females.

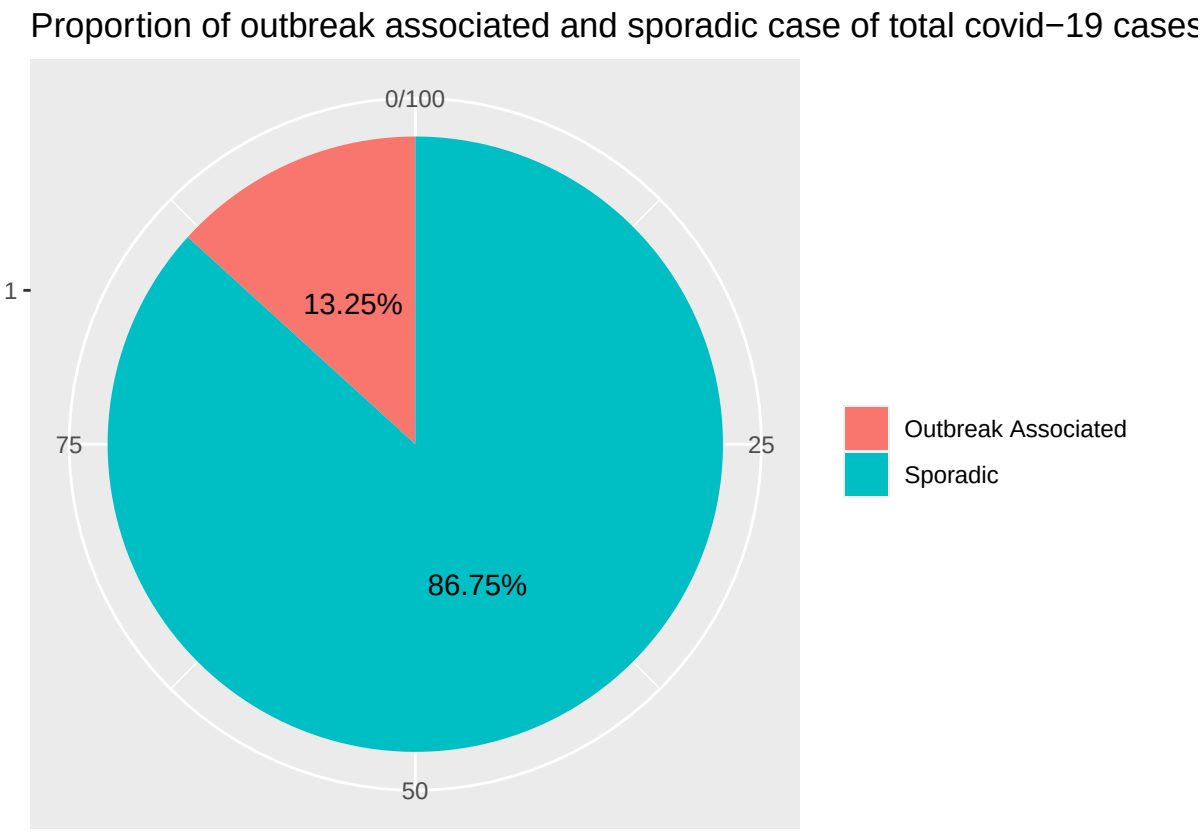

Figure 4: proportion of outbreak associated and sporadic cases

Table 3: The number and proportion of people ever/currently in ICU in different age group

Age_Group	count	prop
19 and younger	34	0.001
20 to 29 Years	62	0.001
30 to 39 Years	136	0.002
40 to 49 Years	238	0.006
50 to 59 Years	465	0.012
60 to 69 Years	766	0.032
70 to 79 Years	607	0.051
80 to 89 Years	294	0.033
90 and older	50	0.010

Table 4: The number and proportion of people ever/currently hospitalized in different age group

Age_Group	count	prop
19 and younger	292	0.006
20 to 29 Years	484	0.008
30 to 39 Years	792	0.014
40 to 49 Years	1126	0.026
50 to 59 Years	1955	0.051
60 to 69 Years	2587	0.109
70 to 79 Years	2662	0.225
80 to 89 Years	2668	0.301
90 and older	1210	0.247

Table 5: The number and proportion of people ever/currently intubated in different age group

Age_Group	count	prop
19 and younger	6	0.000
20 to 29 Years	27	0.000
30 to 39 Years	67	0.001
40 to 49 Years	141	0.003
50 to 59 Years	296	0.008
60 to 69 Years	471	0.020
70 to 79 Years	376	0.032
80 to 89 Years	165	0.019
90 and older	20	0.004

Compared to the people infected from close contact, people infected from the community have 1.3 fold of odds of being in ICU, and people infected from health care have a lower odds of being in ICU (0.355).

5 Discussion

5.1 Purpose and Summary of the report

The covid-19 outbreak in 2019 brought significant influence to global economic systems and people's health. This virus has highly contagious and is mainly transmitted when people breathe in air contaminated by droplets and small airborne particles containing the virus. Before April 16, 2022, more than 503 million confirmed cases had been reported globally, and over 6.196 million patients died because of it. The group of people who Covid-19 infects may have fever, cough, breathing difficulties, loss of smell and taste symptoms, but some may suffer dyspnea, hypoxia, shock or multiorgan dysfunction. After research, we find that the confirmed cases exceed 40000 cases each month in January and February in Toronto, so it seriously threatens people's health.

In this report, I used the data extracted from the Case Contact Management System to identify the characteristics of people who are more likely to be infected by a virus. Also, I analyzed what kind of people will have severe symptoms e.g., ever being in ICU. My target population is all the confirmed or probably infected citizens in Toronto. All the information about the patients is reported to and managed by Toronto Public Health.

In this paper, I accumulated the number of confirmed cases each month in Toronto since 2020 to assess the spread of the virus. Also, I discussd the number of people infected by different sources in Toronto since 2020 and the potential relationship among gender, age, and ever accepting treatment in the ICU. Finally, the resolved rate and mortality rate are mentioned as well.

In conclusion, the total number of confirmed cases and probably cases display a declining trend in these two months. The males who are 70-89 years old and infected with the virus from the community have a higher possibility of being in ICU.

5.2 Learned about the world

From the result of research, I clearly see that Covid-19 is highly contagious and spreads quickly, especially at the beginning of 2021 and January and February 2022. Therefore, I conclude that a lockdown policy plays an essential role in controlling the virus and mandatory vaccination is an excellent way to reduce the risk of infection.

Although the Covid-19 is easily spread, the resolved rate is much higher than the active and fatal rates, which means that the virus won't damage our health severely. However, I find that older people are at a higher risk of developing severe symptoms, especially those 60 and above 60 years old. In addition, the number of old patients who are ever in hospital or ICU to accept treatment is much more significant than the younger. To reduce the number of infected older people, the government should encourage the elder to do more exercise and regular their diet to enhance their immunity.

In addition, I find that community and household contact are two primary sources of infection in existing information. Therefore, I conclude that a person has a higher possibility to infect his roommate or people in the same community. It proves that Covid-19 spreads quickly and is easily transmitted by close contact. According to this finding, I suggest the government force the people who have close contact with infected people to experience seven days of home quarantine and detect their temperature during this period. Also, if a person is infected by the virus, all the people living in the same community should be informed and encourage them to do self-test.

Finally, the logistic regression model displays that among people ever with Covid-19 infections, males are more likely to being in ICU compared to females. Vaccines are essential tools in our fight against the pandemic. In addition, I suggest that all the males reduce the frequency of smoking and drinking because smokers and drinkers are likely to be more vulnerable to Covid-19; they may also already have lungs. Generally speaking, males may take more financial responsibilities and be occupied with their careers, having irregular lifestyles. Quite a few males work overtime or through the night, so their body clock is affected, and their immune system declines. I highly suggest they change their lifestyles, such as falling asleep before 11 o'clock, and improve their immune systems.

5.3 Weaknesses and learn in the future

However, there are several limitations in my research which may make the data inaccurate and I need to improve in the future.

Firstly, the data I used may change because it is subject to change as public health investigations into reported cases and continuous quality improvement initiatives continue. It is completely refreshed weekly, extracted on Tuesday, and posted on Wednesday. Data are from different sources, so they may differ from those published elsewhere.

Secondly, the data population includes all the confirmed and probable cases reported to the government, and the size is large enough for our research. Still, some citizens may conceal the fact that they are infected. Because of this reason, our data is still incomplete, and the basis has existed. In order to solve this problem, the only measurement is that the government enhances the covid-19 regulation propaganda and punishes those who hide their actual health condition.

Thirdly, some asymptomatic patients may influence the precises of data and bring some basis. Asymptomatic refers to people who are infected but never develop symptoms during the infection period. Therefore, they can transmit the virus to others, but they don't recognize that. These asymptomatic cases are excluded by our data, because asymptomatic patient may not test whether they are infected. To eliminate the inaccurate, I suggest citizens in Toronto test themselves every week and the government should provide enough iHealth test.

Finally, the report mentions that males with Covid-19 infections are more likely to be in ICU than their female counterparts, but we cannot explain this phenomenon. In order to provide more efficient suggestions and measures to control the spread of the virus, it's necessary to do further research. To be more specific, I make some assumptions and try to prove it. Does Covid-19 relate to a smoking habit or some unique cell in males' blood? Does Covid-19 relate to the imbalance between work and life? Another area we can learn is about the source of infection. From the data I used in the paper, there are many cases without providing a source of infection information, which may influence our identification and measurements preventing Covid-19 spread. In the future, I should collect relating information as much as possible.

Appendix

```
## Rows: 294,842
## Columns: 18
## $ X_id                   <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ~
## $ Assigned_ID            <dbl> 1, 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, ~
## $ Outbreak.Associated    <chr> "Sporadic", "Sporadic", "Sporadic", "Sporadic",~
## $ Age_Group              <chr> "50 to 59 Years", "50 to 59 Years", "20 to 29 Y~
## $ Neighbourhood.Name     <chr> "Willowdale East", "Willowdale East", "Parkwood~
## $ FSA                    <chr> "M2N", "M2N", "M3A", "M4W", "M4W", "M2R", "M1V"~
## $ Source_of_Infection    <chr> "Travel", "Travel", "Travel", "Travel", "Travel~
## $ Classification         <chr> "CONFIRMED", "CONFIRMED", "CONFIRMED", "CONFIRM~
## $ Episode_Date           <dttm> 2020-01-22, 2020-01-21, 2020-02-05, 2020-02-16~
## $ Reported_Date          <dttm> 2020-01-23, 2020-01-23, 2020-02-21, 2020-02-25~
## $ Client_Gender          <chr> "FEMALE", "MALE", "FEMALE", "FEMALE", "MALE", "~
## $ Outcome                <chr> "RESOLVED", "RESOLVED", "RESOLVED", "RESOLVED",~
## $ Currently_Hospitalized <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Currently_in_ICU       <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Currently_Intubated    <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Ever_Hospitalized      <chr> "No", "Yes", "No", "No", "No", "No", "No", "Yes~
## $ Ever_in_ICU            <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
## $ Ever_Intubated         <chr> "No", "No", "No", "No", "No", "No", "No", "No",~
```

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 record the information about confirmed covid-19 cases and probable cases. No specific gap needs to be filled.
2. *Who created the dataset (for example, which team, research group) and on behalf of which entity (for example, company, institution, organization)?*
   - Qiuhan Wang created the dataset on behalf of University of Toronto.
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.*
   - No funding was received for this project.
4. *Any other comments?*
   - No

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.* By 2020, all the people in Toronto whose self-covid-19-test result is positive have the responsibility to report to Toronto Public health consciously. Additionally, the Toronto Public Health system will automatically record the covid-19 test results for people who have been in walk-in-clinic or similar medical institutions.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There 295104 observations and 17 variables in total. Gender had 6 options; age group had 9 options;source of infection had 9 options and other variables "Currently_Hospitalized", "Currently_in_ICU", "Currently_Intubated", "Ever_Hospitalized", "Ever_in_ICU", "Ever_intubated" have 2 categories.
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).*

- The dataset contains all possible instances rather than a larger set, because all the people in Toronto who are infected should be recorded in dataset.

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.*
   - The data describes the information about confirmed cases and probable cases in Toronto since 2020. It includes "Assigned_ID", "Outbreak.Associated", "Age_group", "Neighbourhood.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".

5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Yes, there are demographic, geographic, and severity information for all confirmed and probable cases reported to Toronto Public Health.
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.*
   - The dataset shows the percentage of women's education level in Turkey in 1998. Individual instances are confidential, so we don't know if there is any missing information from individual instances.
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.*
   - Yes. For example, some patients who are infected by Covid-19 may come from the same household.
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.*
    - The datasheet is available on the website "open.toronto.ca". The data will be refreshed and overwritten weekly. The website is only available for those people who have special username and passward.
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.*
    - The dataset identifies sub-populations. It only records the information of infected people and probabley infected people.
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.

- Yes, it provides patients' neighborhood name, which is private to patients.

16. *Any other comments?*
    - No

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 was derived from "open.toronto.ca". We don't find any error in the data, but we find some blank and "no information" in some variables.
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?*
    - The dataset was collected by provincial Case & Contact Management System and Toronto Public Health.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - No. In this dataset, population size and sample size are equaling.
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)?*
    - Contractors were involved in the data collection process. Every group consisted of one manager, five females, and one male.
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.*
    - The data in the report includes the information about infected people in Toronto before 2022-03-23. However, this data is freshed weekly, which means that you can download the lates one from website "open.toronto.ca".
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)?*
    - The dataset was obtained from third parties. The complete dataset can be obtained from website "open.toronto.ca".
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.*
    - Yes. Because all the information of infected people are recorded by the government and they are informed brfore recording.
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.* -No.
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).*
    - The questionnaire did not mention whether participants can revoke their consent to the collection and use of data.

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?*
    - No

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.*
    - The dataset was split into text lines using the R package "stringi". Then we separated the data into columns and stored it as raw data.
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.*
    - The raw data was saved in addition to the cleaned data. It is available thorugh Github.
3. *Is the software that was used to preprocess/clean/label the data available? If so, please provide a link or other access point.*
    - R was used.
4. *Any other comments?*
    - No

Uses

1. *Has the dataset been used for any tasks already? If so, please provide a description.*
    - The dataset was used to analyze the reasons for dropping out of school for women with different levels of education in Turkey in 1998.
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.*
    - No
3. *What (other) tasks could the dataset be used for?*
    - The dataset can also be used to calculate the number of infected people in Toronto and analyze the characteristics of patients. From discuss the source of infection, the government can find a way to prevent the spread of 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.*
    - No
6. *Any other comments?*
    - No

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.*
    - The dataset and report are available through Github. Code and data are available at: https://github.com/wangq166/Covid-19-Research.git
2. *How will the dataset be distributed (for example, tarball on website, API, GitHub)? Does the dataset have a digital object identifier (DOI)?*
    - The dataset is available through Github. The dataset does not have a DOI.
3. *When will the dataset be distributed?*

- The dataset is available through Github.

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.*
   - The dataset is not distributed under a copyright, IP license, and ToU. The dataset is licensed under the MIT 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?*
   - No

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - Qiuhan Wang
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - qiuhan.wang@mail.utoronto.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.*
   - 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. Please note that these numbers may differ from those posted elsewhere, as data are extracted at different times, and from different sources.
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.*
   - No. Because all the old datas are overwritten.
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.*
   - If others want to extend/augment/build on/contribute to the dataset, they can pull requests on Github. We will credit their contributions.
8. *Any other comments?*
   - No

References

Grolemund, Garrett, and Hadley Wickham. 2011. *Dates and Times Made Easy with lubridate. Journal of Statistical Software*. Vol. 40. https://www.jstatsoft.org/v40/i03/.

Lazarus, S., Ratzan. 2020. "COVID-SCORE: A Global Survey to Assess Public Perceptions of Government Responses to COVID-19" 1 (3): 86. https://doi.org/https://doi.org/10.1371/journal.pone.0240011.

Lupton & K. Willis, Eds. 2021. "The COVID-19 Crisis: Social Perspectives" 2 (2): 168.

R Core Team. 2020. *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 Fran??ois, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.