The influence of COVID-19 on the world and Canada[*]

Defining the population most infected by Covid-19 since 2020

HaoCheng Xu

27 April, 2022

Abstract

Covid-19 is a highly contagious and difficult to avoid the virus that broke out in 2019, having a tremendous effect on world economics and people's lives. The study utilizes the Toronto covid-19 dataset, which is maintained by the Provincial Case and Contact Management System. To assist the government and public in restricting the virus's spread, I evaluated the characteristics of those who are more prone to contracting the virus and their primary sources of infection. I've discovered that covid-19 can infect anyone of any age, that their primary source of infection is travel, and that it has a more severe effect on the elderly.

Contents

[*] Code and data are available at: https://github.com/Jasonfallen/The-influence-of-COVID-19.git

1 Introduction

Covid-19 is a viral-borne infection caused by the SARS-Cov-2 virus. When an infected person coughs, speaks or breathes, the virus can spread through the mouth or nose in microscopic liquid particles. Since 2019, it has been a crisis affecting both global economic systems and individual life. (Lazarus 2020) According to studies, the virus has infected about 479 million individuals and killed over 6 million people. (Lupton & K. Willis 2021) Additionally, lockdowns associated with Covid-19 occurred often between 2020 and 2021, directly affecting economic activity. The majority of countries saw their gross domestic product (GDP) growth suffer the most substantial setback.

The mechanisms through which the virus can be efficiently controlled and covid-19 prevented are mainly unknown. I intend to use this document to assist the government with publicizing methods to contain the virus's transmission and avoid the emergence of covid-19. I clean and analyzed the data, by using R (R Core team, 2020), tidymodels (Kuhn and Wickham, 2020), ggplot2 (Wickham, Hadley, 2016), readr (Hadley Wickham, 2022) packages, and I will outline the features of patients who have been confirmed or are very likely to have been infected with the virus in Toronto since January 2020 and determine the virus's influence on the health of a distinct set of individuals.

The information was derived from the provincial Case and Contact Management System (CCM). This dataset includes demographic, geographic, and severity data for all confirmed and probable cases reported to and managed by Toronto Public Health between January 2020 and December 2020. Cases occurring in the community and epidemics are included in the statistics. The findings indicate that while the majority of patients recover without additional treatment, older adults are more likely than younger adults to acquire a serious disease. Additionally, we discover that travelling is the principal source of infection, and their outbreaks are intermittent.

The following paragraphs comprise the article. To begin, I describe the dataset, which includes data sources, data gathering techniques, and data variables. Second, I categorize affected individuals according to their age and gender in order to determine who is more prone to get the virus. Additionally, I address infection origins and the state of disease of patients. Finally, recommendations on covid-19 prevention are made to the government and the public.

2 Data

2.1 Data collection

The data for this article was obtained from the website "open.toronto.ca." On the internet, select open data portal home and then search for the term "Covid" to obtain the catalogue titled "About Covid-19 Cases in Toronto." Finally, pressing the download data section gets the "COVID19 cases" dataset. In comparison to other data on covid-19 cases, this data is more informative and authoritative because it is derived directly from the provincial Case Contact Management System (CCM). This dataset includes demographic, geographic, and severity data for all confirmed and suspected cases reported to Toronto Public Health from January 2020. Additionally, sporadic and outbreak-related instances are included.

2.2 Data processing

All studies are carried out in R (R Core Team 2020), a statistical computing language. I acquired the dataset, which has 295104 observations, from the opentoronto.ca website. The data collection contains the following variables:

"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".

The missing value of patients' age is filtered by using the R function "filter".There are 7319 values in the Neighborhood that are missing. Although I investigate the Name column and the 3804 missing values in the FSA column in the manuscript, I do not study these variables. After eliminating individuals whose ages are unclear, I am left with 294842 observations. The data manipulation software "tidyverse" (Wickham et al. 2019) is used, and the data visualization package "ggplot2" (Wickham 2016) is utilized. "lubridate" (Grolemund and Wickham 2011) is installed to create a linear regression and logistic model for analyzing the association between a patient's gender, age, and health status.

The Appendix has a statistical overview of all variables, and the "data characteristics" section contains further information. Additionally, I created multiple plots based on age, gender, source of infection, and willingness to seek medical care. (they may be viewed in the "Results" section)

2.3 Survey Method

By 2020, all residents of Toronto with a positive self-covid-19 test will be required to voluntarily report to Toronto Public Health. Additionally, the Toronto Public Health system will automatically record covid-19 test results for individuals who visit walk-in clinics or other comparable medical facilities.

2.3.1 Population and sample

In this scenario, the population size is equal to the sample size. It is a census; it covers all persons infected with covid-19 in Toronto since 2020, regardless of gender or age. The province Case & Contact Management System has collected 295104 observations.

2.3.2 Strength

Because the provincial Case & Contact Management System provides all of the data, they are both informative and authoritative. In comparison to other sources of data, it correctly tracks all confirmed or probable cases in Toronto. Additionally, the data is totally reset and wiped each week, ensuring that it remains current. It's advantageous for our research.

2.3.3 Weakness

Although the data covers as many Covid-29 instances as feasible in Toronto in 2020, it also contains some blank or missing data that may affect our conclusions. For example, 261 patients do not offer their age, and more than 40% of patients do not respond to their infection resources. Due to the impossibility of replacing or deleting these data due to their volume, we have a sampling error here.

2.4 Data characteristics

Our data set has 294842 observations, and all variables in the report are categorical, including gender, age, source of infection, reported date, and ever/currently in IC, ever/currently hospitalized, and ever/currently incubated. We may categorize people into nine age categories, including "19 and younger," "20 to 29 years old," "30 to 39 years old," and "90 and older." Additionally, we have nine categories for the source of infection, including close contact, community, healthcare facility, communal settings, household contact, unknown, pending, travel, and other sitting. The reported date range is between 2020-01-23 and 2022-03-22. We may see further details in the table 1 and table 2.

Table 1: The total infected people in distinct 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 total infected people from distinct 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

3 Results

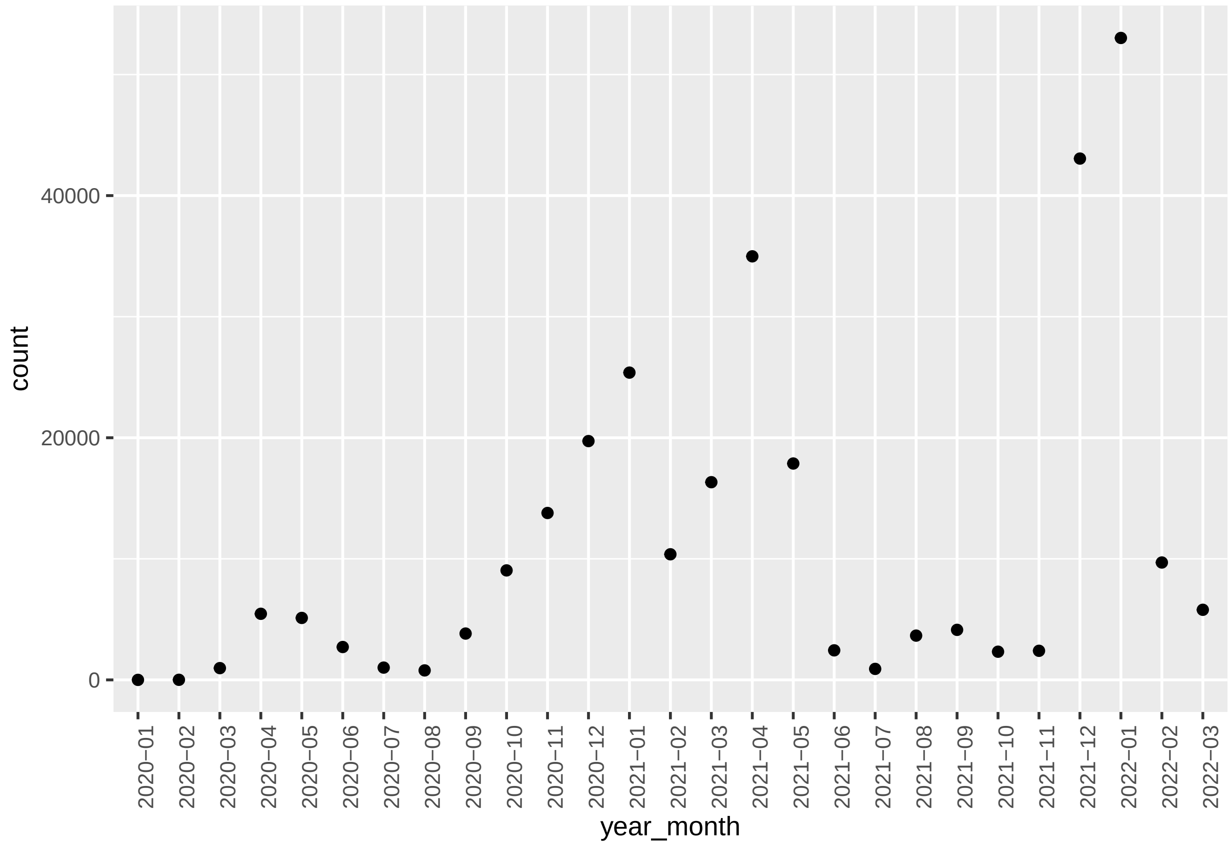

Figure 1: the number of confiremed case or porble cases each month

After charting the number of confirmed and probable cases in Toronto from January 2020 to April 2022 in Figure 1, I figure out that between August 2020 and February 2021, there was a considerable increase in instances, from 5000 to 27000. Although the number of infected individuals has declined significantly since June 2021 and has remained below 5000 in subsequent months, I nonetheless saw an enormous increase to 52000 cases in February 2022.

According to Figure 2, the number of patients aged 20 to 29 exceeds 60000, placing them top. Additionally, I discovered that the number of patients aged 30 to 39 is around 55000. As the fatality rate grows with age, the graph indicates that around one in every four persons over the age of 90 cannot recover effectively. Additionally, persons under the age of 40 had a greater rate of activity than those over 40.

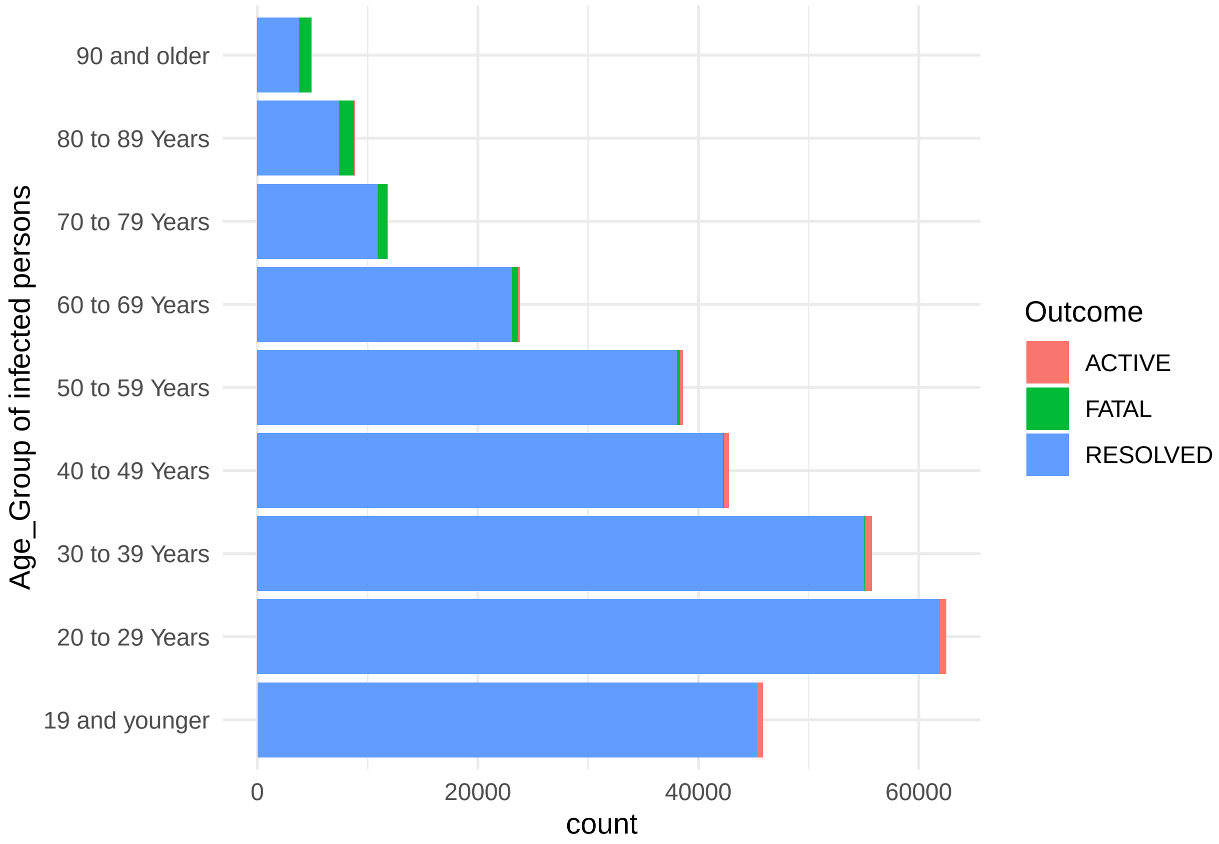

Figure 2: outcome for infected people in different age groups

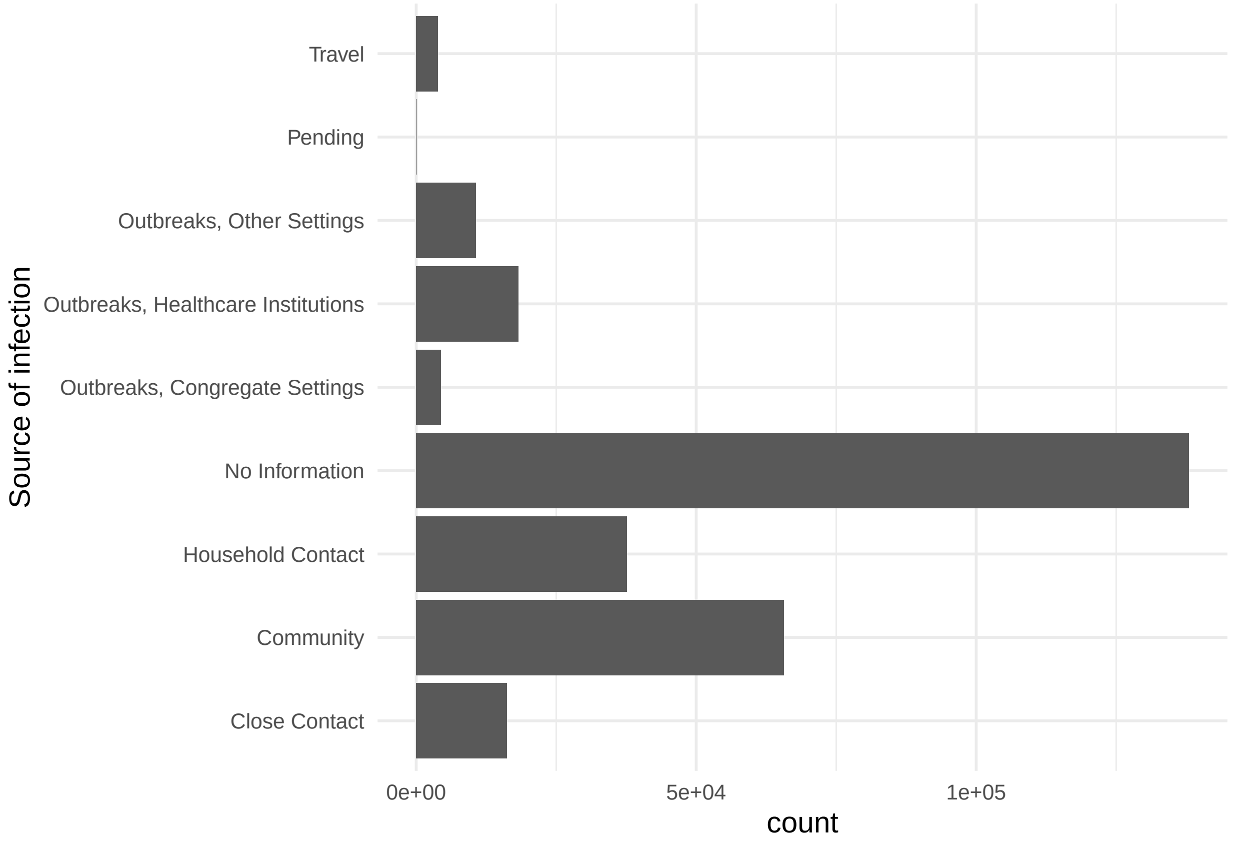

Figure 3: sources of infection distribution

According to Figure 3, patients can acquire covid-19 through a variety of routes, including travel, close contact, and community. Apart from those who withhold information concerning the source of illness, community, household contact, and outbreaks, three key sources are regarded to be health care facilities.

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

From Figure 4, we found that 86.75% of cases are sporadic, and 13.25% are outbreak-associated. This represents that sporadic cases are nearly seven times that of outbreak-associated 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

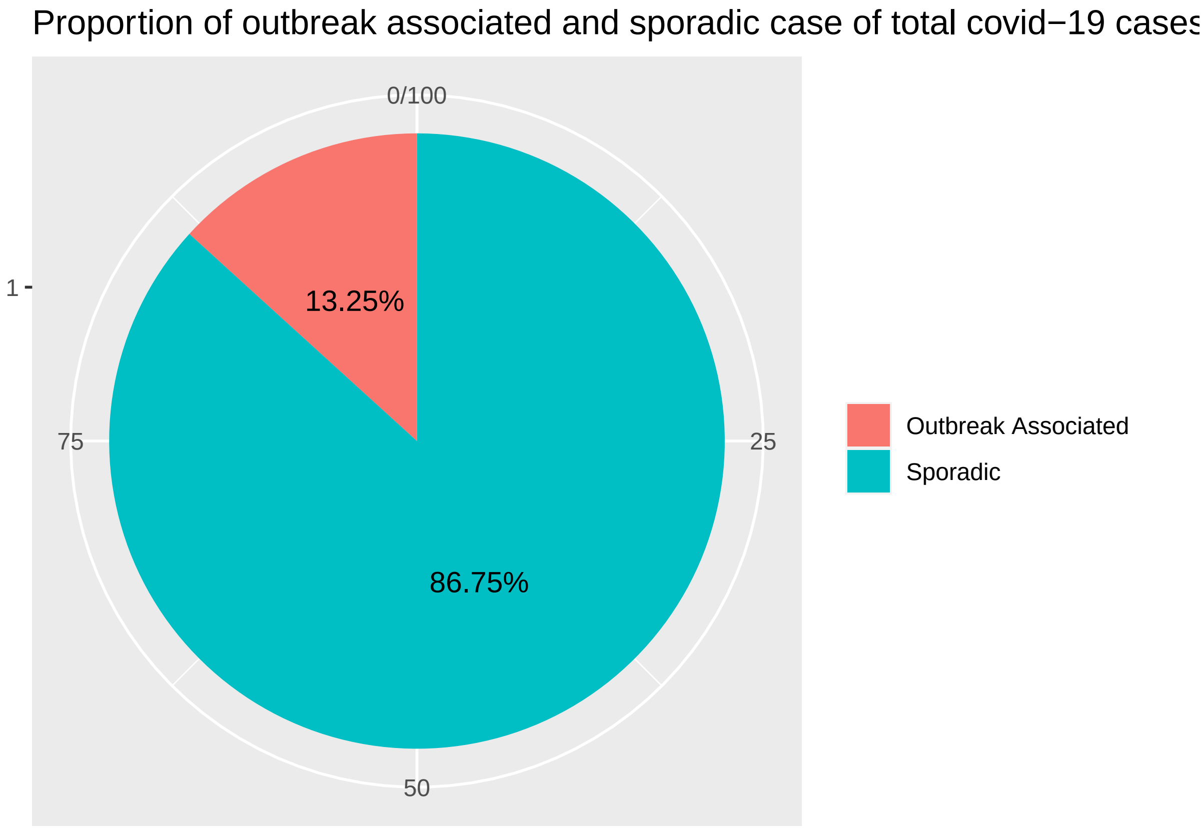

Figure 4: proportion of outbreak associated and sporadic cases

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

The tables 3 to 5 detail the number of persons who have ever or currently been in an intensive care unit, who have ever or currently been hospitalized, and who have ever or currently been intubated in various age categories. According to tables 1 and 3, the proportion of infected individuals in each age group who accept ICU or intubated therapy is rather low, less than 5%. However, I can see that the fraction of hospitalized patients increases as the patient's age increases. The table2 demonstrates that while the proportion of people in the age groups "19 and younger" and "20 to 29" who are hospitalized is less than 1%, the proportion of people in the age groups "70 to 79," "80 to 89," and "90 and older" who are hospitalized is 23 percent, 30%, and 24.7 percent, respectively.

4 Model

Because the response is binomial, it is preferable to use logistic regression to examine the effect of age, sex, and source of infection on the chance of being admitted to the intensive care unit. According to the model summary (Table 6), I infer that community-infected males aged 70-89 are more likely to accept ICU care. Additionally, I discover that the 70-79 and 80-89 age groups had a 62.8 and 78.4 times increased likelihood of being in the intensive care unit, respectively, compared to the under 19 age group. Males have a 1.77-fold greater chance of being in the intensive care unit than females. In comparison to those infected through close contact, those infected in the community had a 1.3-fold increased risk of being in the ICU, whereas those infected through health care have a reduced risk (0.355).

5 Discussion

5.1 Conclusion of Covid infection model model

In 2019, the covid-19 epidemic had a huge impact on worldwide economic systems and public health. This virus is extremely infectious and is mostly spread when humans breathe air polluted with virus-containing droplets and tiny airborne particles. Prior to April 16, 2022, over 503 million confirmed cases have been reported worldwide, with over 6.196 million individuals dying as a result. The group of patients infected with Covid-19 may experience fever, coughing, difficulty breathing, and loss of smell and taste, while others may experience dyspnea, hypoxia, shock, or multiorgan failure. After doing an investigation, we discovered that confirmed instances surpass 40000 in January and February in Toronto, posing a major threat to people's health.

I utilized data taken from the Case Contact Management System to determine the characteristics of individuals who are more likely to contract a virus for this study. Additionally, I examined whether individuals may experience severe symptoms, such as being admitted to the intensive care unit. My target demographic is all of Toronto's confirmed or likely afflicted residents. Toronto Public Health receives and manages all patient information.

In this research, I compiled the number of verified cases in Toronto each month from 2020 to examine the virus's spread. Additionally, I discussed the number of persons infected in Toronto by various sources since 2020 and the possible association between gender, age, and ever receiving ICU treatment. Finally, the rate of resolution and mortality are stated.

In conclusion, the overall number of confirmed and probable cases has decreased significantly during the last two months. Males are aged 70-89 years who are infected with the virus in the community have a greater chance of being admitted to the intensive care unit.

5.2 Global vision

As a consequence of my investigation, I've concluded that Covid-19 is very infectious and spreads rapidly, particularly in early 2021 and January and February 2022. As a result, I argue that a lockdown strategy is critical for virus management and that mandated vaccination is a great method of reducing the danger of infection.

Although the Covid-19 virus is easily distributed, the resolved rate is far greater than the active and deadly rates, indicating that the virus will have little effect on human health. However, I've discovered that older adults, particularly those aged 60 and above, are at a larger risk of having severe symptoms. Additionally, the number of elderly individuals admitted to the hospital or intensive care unit to accept treatment is significantly higher than the number of younger patients. To lower the number of diseased older adults, the government should encourage seniors to increase their physical activity and maintain a healthy diet in order to boost their immunity.

Additionally, I discovered that community and home contact are two of the key causes of infection in the available data. As a result, I infer that a person has a greater chance of infecting his roommate or other members of the same community. It demonstrates how rapidly Covid-19 spreads and how easily it is transferred through close contact. According to this discovery, I propose that the government impose a seven-day home quarantine on persons who have intimate contact with sick people and monitor their temperature throughout this time. Additionally, if a person becomes infected with the virus, all residents in the same neighbourhood should be notified and encouraged to self-test.

Finally, the logistic regression model indicates that males are more likely to be admitted to the intensive care unit than females when they have ever had Covid-19 infections. Vaccines are critical instruments in our pandemic-fighting arsenal. Additionally, I recommend that all guys minimize their smoking and drinking habits, as smokers and drinkers are more likely to be susceptible to Covid-19 and may already have lungs. Males, on average, may have greater financial responsibilities and be preoccupied with their employment, leading to erratic lives. Numerous guys work overtime or through the night, which disrupts their biological clock and weakens their immune system. I strongly advise them to alter their lives, such as sleeping before 11 a.m. and strengthening their immune systems.

5.3 Weaknesses

My research has various constraints that may cause the data to be wrong, and I need to better in the future. To begin, the data I utilized may change as public health investigations into reported instances and quality improvement efforts continue. Each week, it is totally renewed, extracted on Tuesday, and uploaded on Wednesday. Because the data come from a variety of sources, they may differ from those previously published.

Moreover, the data population is sufficiently big to encompass all confirmed and probable cases submitted to the government. Nonetheless, some citizens may conceal their infection. As a result, our data is still incomplete, but the foundation has existed. To resolve this issue, the government's sole metric is that strengthens the covid-19 regulating campaign and penalize individuals who conceal their true health status.

Certain asymptomatic individuals may have an effect on the precision of the data and provide some context. Asymptomatic infection refers to those who are infected but do not manifest symptoms during the illness phase. As a result, they are capable of transmitting the virus to others, although they are unaware of this. Our findings omit these asymptomatic instances, as asymptomatic patients may not be able to test for infection. To eliminate the inaccuracies, I propose that residents in Toronto self-test weekly and that the government give sufficient iHealth tests.

5.4 Future Works

In sum, the paper notes that males with Covid-19 infections are more likely to be admitted to the intensive care unit than females, although we are unable to explain this finding. Further study is essential in order to give more effective proposals and techniques for controlling the virus's spread. To be more precise, I make several assumptions and attempt to demonstrate them. Is Covid-19 associated with a smoking habit or with a particular type of male blood cell? Is Covid-19 connected to the disjunction between work and life? Another area in which we might gain knowledge is the source of illness. According to the data I utilized in the article, there are several situations where no source of infection information is provided, which may affect our diagnosis and prevention of Covid-19 spread. In the future, I should gather as much pertinent information 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)?*
    - HaoCheng Xu (author) 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 residents of Toronto with a positive self-covid-19 test will be required to voluntarily report to Toronto Public Health. Additionally, the Toronto Public Health system will automatically record covid-19 test results for individuals who visit walk-in clinics or other comparable medical facilities.
2. *How many instances are there in total (of each type, if appropriate)?*

- There are a total of 295104 observations and 17 variables. Gender was classified into six categories; age group was classified into nine categories; source of infection was classified into nine categories; and other variables such as "Currently Hospitalized", "Currently in ICU", "Currently Intubated", "Ever Hospitalized", "Ever in ICU", and "Ever intubated" were classified into two 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 comprises all potential cases rather than a subset, thus all infected individuals in Toronto should be included in the 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, all confirmed and probable cases submitted to Toronto Public Health include demographic, geographic, and severity information.

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 statistic depicts the proportion of women with a secondary education in Turkey in 1998. Because individual cases are secret, we do not know if any information from particular occurrences is missing.
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 instance, several people infected with Covid-19 may share a home.

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 accessible on the University of Toronto's website, "open.toronto.ca." Weekly, the database will be updated and rewritten. The website is only accessible to individuals with a unique login and password.

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.*

    - Subpopulations are identified in the dataset. It simply keeps track of infected and possibly infected individuals.

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". I didn't find any error in the data, but I found some missing data 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 report contains statistics on affected individuals in Toronto prior to 2022-03-23. However, this data is updated weekly, so you may obtain the most recent version from the 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 the government records all information on sick individuals and notifies them prior to 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?*
   - Additionally, the dataset may be used to estimate the number of infected persons in Toronto and to study patient characteristics. By examining the source of infection, the government can devise a strategy for controlling 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:
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?*
   - HaoCheng Xu
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - haocheng.xu@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 daily 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. Weekly, the data will be totally updated and rewritten, retrieved at 8:30 AM on Tuesdays and uploaded on Wednesdays. Please keep in mind that these figures may differ from those published elsewhere, as data is extracted at various points in time and from a variety of 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.*
   - Pull from github
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.