Cases of COVID-19 in Toronto Show Different Patterns by Gender and Age and Require Stronger Management Measures

Zihan Jin

2/5/2022

Abstract

COVID-19 case datasets are important tools for rational implementation of medical measures and allocation of medical resources to avoid outbreak caused by transmission existed infections. This report utilizes data on reported probable and confirmed case in Toronto to: (1) analyze the pattern of case outcome by gender and age, (2) discuss sources of infection and their rationality under real-world context, and (3) examine the timeliness of case reporting using time lag between episode date and reported date. General trends can be conclude after discussion that death rate decreases with age and higher in male than in female. Moreover, most proportion of cases are fail to manage with determined infection source and need a reporting time of 3 days, which cause difficulty in mitigating transmission of Coronavirus.

Introduction

The COVID-19 pandemic is now in its third year since the Who declared the novel Coronavirus as a world pandemic on 11th March, 2020 ("World Health Organization" 2020). When it will end, and how it will end, has been the primary concern of the world. According to the latest real-time statistics from WHO, as of 17:30 CET on 4 February, the cumulative number of confirmed COVID-19 cases worldwide has reached 38,6548,962, with 5,705,754 cumulative deaths ("World Health Organization" 2022). Canada reported 3,106,549 confirmed cases and 34,493 deaths as of February 4th afternoon local time according to CTV news ("COVID-19 Canada | Ctv News | Coronavirus" 2022). Start from June 14th 2021, the Ontario government started to conduct a three-step roadmap to reopen safely and cautiously ("Ontario Newsroom" 2022). Until the written date of this report, Toronto is in modified Step 3 of the Province's Roadmap to Reopen from 31th Jan 2022 ("City of Toronto" 2022): all City-operated recreation facilities, including pools, indoor arenas and gyms, are reopened at 50 per cent capacity for drop-in programs. To guarantee the reopen conduct successfully, Toronto Public Health(TPH) required highly effective measurement to manage COVID-19 cases from transmission.

COVID-19 case statistics are an important tool for rational implementation of medical measures and allocation of medical resources to avoid outbreak caused by transmission existed infections.Data from COVID-19 reports contains demographic, geographic, and severity information. Demographic information can used to understand the susceptible population and provide reference for clinical management; geographic information can help track transmission routes for disinfection and immunization; severity information can guide the deployment of medical resources.

Given the importance of COVID-19 case statistics in public health and medical research, it is necessary to understand how COVID-19 related data is reported and how it may be interpreted. For this report, I will use open-access data from the Toronto Police Services. The raw dataset will be cleaned and modified in analyzing purpose, without change to original data. The main content focuses on the pattern of case outcome by groups of gender and age, source of infection, and time lag for case reporting. Analyzing are based on summary table and data visualization using statistical graphes, including proportional, pie, bar, and box plots, as well as the real-world context. The dataset will be processed and analyzed in R primarily using the tidyverse (Wickham et al. 2019) and dplyr (Wickham et al. 2021) packages. Figures and tables will be created with ggplot2 (Wickham 2016), lubridate (Grolemund and Wickham 2011), patchwork (Pedersen 2020), and scales (Wickham and Seidel 2020).

Data Source

This report utilizes data on confirmed or probable COVID-19 cases reported to and managed by Toronto Public Health(TPH) from Case & Contact Management(CCM) System. The TPH reports to the Board of Health and is responsible for the health and well-being of all 2.9 million residents. The reported COVID-19 cases dataset analyzed in this report was obtained in csv format from the City of Toronto Open Data Portal using the R package opendatatoronto (Gelfand 2020). The dataset was last updated on Feb 2nd, 2022.

Methodology and Data Collection

The dataset contains information on all COVID-19 cases that were reported to the Toronto Public Health(TPH) from the years 2020 to 2022. Cases are reported to TPH through the provincial Case & Contact Management System (CCM), a central data repository for COVID-19 case and contact management reporting in Ontario, by local public health unit (PHU). Based on regulations, the PHU must enter contact details into CCM within 24 hours. As a result, the data extracted from CCM may differ from previous or subsequent reports, since it represent a snapshot at the time of data extraction.

While this dataset contains information on all reported COVID-19 cases in Toronto, it is not an accurate representation of actual infection rates in the city. Several studies shows that large amount of people shows no symptoms or have no awareness during their infection of COVID-19. Accumulating evidence, for instance, indicating that a substantial fraction of SARS-CoV-2 infected individuals are asymptomatic (Mizumoto et al. 2020). Approximately one half of residents in Seattle & King Counties with SARS-CoV-2 infection in initial symptom screening are failed to identified (Anne Kimball 2020). 42.5% (95% CI: 31.5–54.6%) of the confirmed SARS-CoV-2 infections had no symptoms while doing swab testing and did not develop symptoms afterwards according to the collected information in Padua, Italy (Lavezzo 2020). These studies indicate that there is non-response bias in data collection, as the infections fail to be tested positive and thus report their infection to the public health institutions. Combining the changes in the availability of testing, driven by increasing COVID-19 cases related to Omicron variant, "cases counts in CCM data reports are underestimate of true number of individuals with COVID-19 in Ontario" ("Ontario Covid-19 Data Tool" 2022). In other words, the estimated infection rate is lower than the real population infection rate, resulted from the bias dataset.

In all other aspects, the dataset still have the advantages of high timeliness and wide area coverage. This is benefit from the strict regulation of uploading cases to the CCM system within 24 hours and the bottom-up reporting of public health management.

Data Characteristic

The dataset contains aggregated data of all confirmed or probable COVID-19 cases reported to TPH bewteen the year 2020 to 2022 ("Ontario Covid-19 Data Tool" 2022). There were 277473 observations in the orginal dataset and 18 attributes: id, Assigned ID, Outbreak Associated, Age Group, Neighborhood Name, FSA (Forward sortation area), Source of Infection, Classification, Episode Date (when the disease was acquired), Reported Date, Client Gender, Outcome, Currently Hospitalized, Currently in ICU, Currently Intubated, Ever Hospitalized, Ever in ICU, Ever Intubated.

For cleaning purpose, the original dataset was modified by 5 steps in a new copy. To begin with, spaces in attribute names are replaced with underscores to facilitate subsequent processing with R language. Then, the first two attributes, id and Assigned ID, were removed as they are numerical identifier for Open Data database and meaningless to the COVID-19 cases description. Missing values are removed as well. Moreover, an additional attribute was created to estimate the time lag in days between the disease acquired data and the date reported to Toronto Public Health, by subtract the Episode Date from Reported Date. The time lag should be zero or positive days. So, the observations with negative time lag were removed, as they contain unreliable information in either attribute of Episode Date or that of Reported Date.

Finally, the dataset contains 277244 observations and 17 attributes.

Outcome

According to information extracted from CCM system, there are three outcomes of the COVID-19 cases: Fatal, Resolved, and Active. Specifically, Fatal are those cases with a fatal outcome reported. Resolved are those not reported as deceased, and who are either reported as "recovered" or where the report date is more than 14 days from symptom onset & the case is not currently hospitalized. Active, in the end, are all other cases. Figure 1 displays the number of three types of outcome from 2020 to 2022.

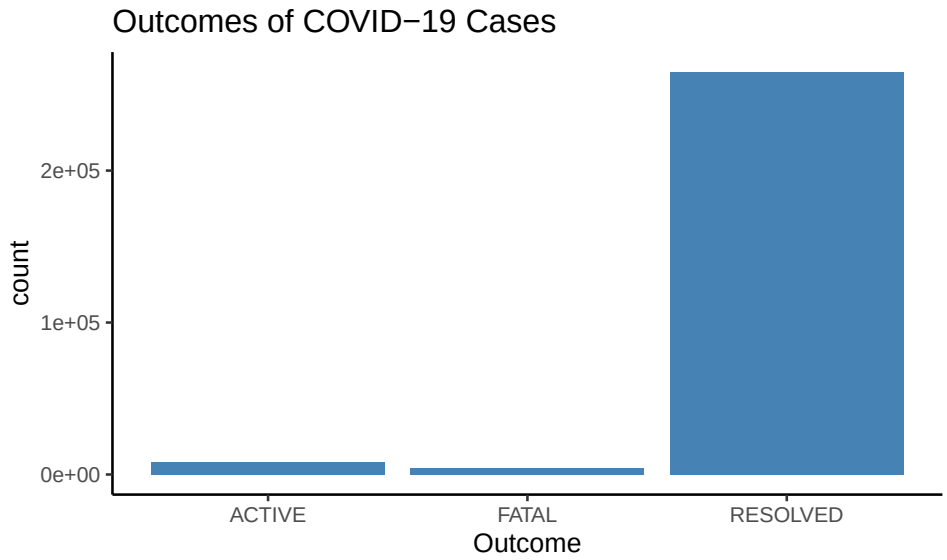

Figure 1: Outcomes of COVID-19 Cases

Based on the figure, we can see that most of the COVID-19 cases are resolved. The proportion of fatal cases is low. According to the definition of Resolved outcome, however, people get infected after two weeks without hospitalized and not reported as deceased are automatically considered as Resolved, which underestimates long-term infections. Therefore, there should be more observations with Outcome of Active and less with that of Resolved, though the proportion shows on the graph would change little in general.

Outcome By Age Group

The outcomes of COVID-19 cases should be analyzed with the discussion of age group, as the proportion of death rate may various in age. In this dataset, the age of patients was selected at time of illness and divided into 9 groups every 10 years old. Figure 2 displays the proportion of each type of outcomes in age groups and Figure 3 displays the proportion of each age groups in different outcomes.

Table 1: Number of COVID-19 Cases in 9 Age Groups

Age_Group	Count
19 and younger	42277
20 to 29 Years	59439
30 to 39 Years	52326
40 to 49 Years	40001
50 to 59 Years	36673
60 to 69 Years	22546
70 to 79 Years	11103
80 to 89 Years	8191
90 and older	4492

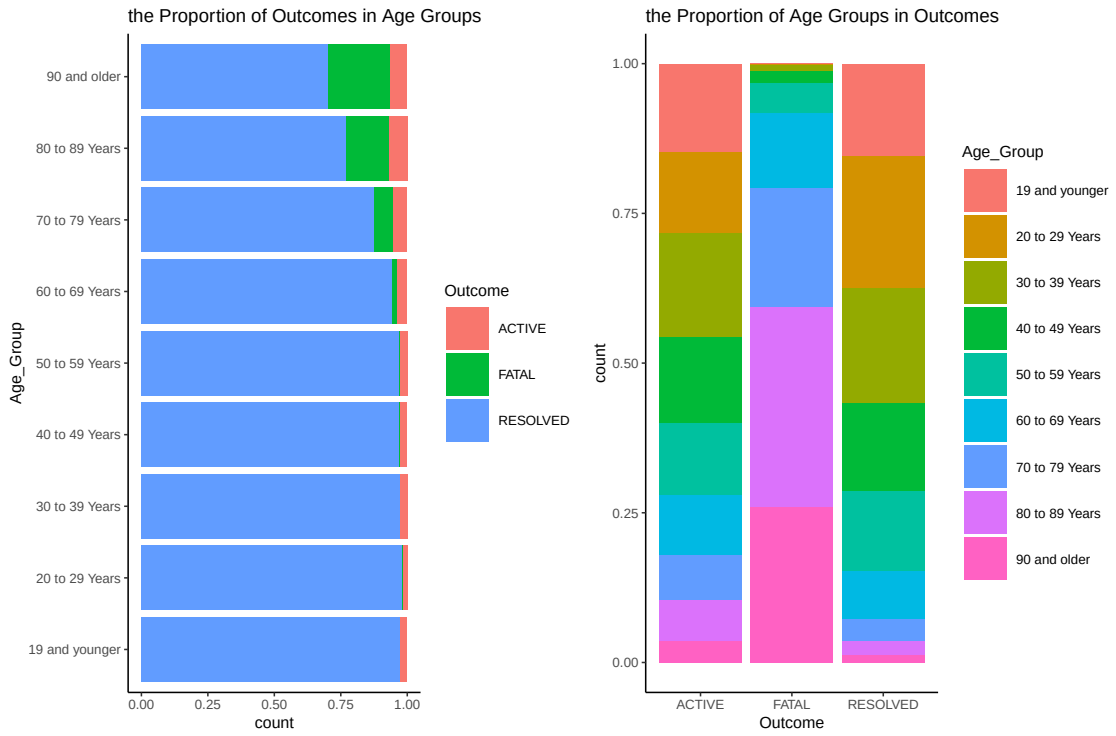

Figure 2: COVID-19 Cases Outcomes By Age

1) Fatal cases:

According to figure 2, death rate climbs to the highest in infections of 90 years old and older among nine age groups, which is about 25 percent. It gradually falls to about 3% in the 60 to 69 years age range and tends to zero in the 20 to 29 years age range. In general, the death rate decreases with age and infection can cause a great threat to their life for elderly, especially those over the age of 60. The conclusion is confirmed by figure 3, which indicates that more than fifty percent of fatal cases are among people older than 80 years old. Fewer cases in group of 90 years and older than in 80 and 89 years old, which is due to smaller population size according to table 1.

2) Active cases:

Of all age groups, active cases had the highest proportion in age of 80 to 89 years old as shown in figure 2, which is slightly higher than in group of 90 and older age. Then, the proportion of active case basically showed a decreasing trend with the decrease of age, until it showed a slight increase in the age group 19 and younger. However, an opposite trend is basically shown in figure 3. This is due to differences in the population size of each age group. Still, we can make estimation that the difficulty of recovery increases with age, but adolescents are also susceptible groups.

3) Resolved cases:

As shown in figure 2, the majority of infected people, in all age groups, eventually recover, and the proportion decreases with age. The largest proportion of resolved cases so far are in the 20-29 age group, which is due to the highest number of people becoming infected in that age group based on table 1.

Outcome By Client Gender

To further understand the attributes of COVID-19 infection outcome, clients' gender is taken into consideration. The gender recorded in the dataset is based on self-reported information of the clients. It is defined as a system that operates in a social context and generally classifies people based on their assigned biological sex.

There are 9 groups of clients' gender in this dataset, which are Female, Male, Non-Binary, Not Listed, Other, Trans Man, Trans Woman, Trans Gender, and Unknown.

Table 2: Number of COVID-19 Cases in 9 Client Genders

Client_Gender	Count
FEMALE	141524
MALE	133204
NON-BINARY	125
NOT LISTED, PLEASE SPECIFY	2
OTHER	15
TRANS MAN	19
TRANS WOMAN	11
TRANSGENDER	22
UNKNOWN	2126

According to the table 2, there are more female cases(51.5%) than male cases(48.5%), which is basically consistent with the gender distribution in Toronto, slightly more females (52%) than males (48%), according to the population demograph in 2016 (Population demographics - toronto 2016). Therefore, we can conclude that the men and women are about equally likely to be infected, in a broad definition of gender.

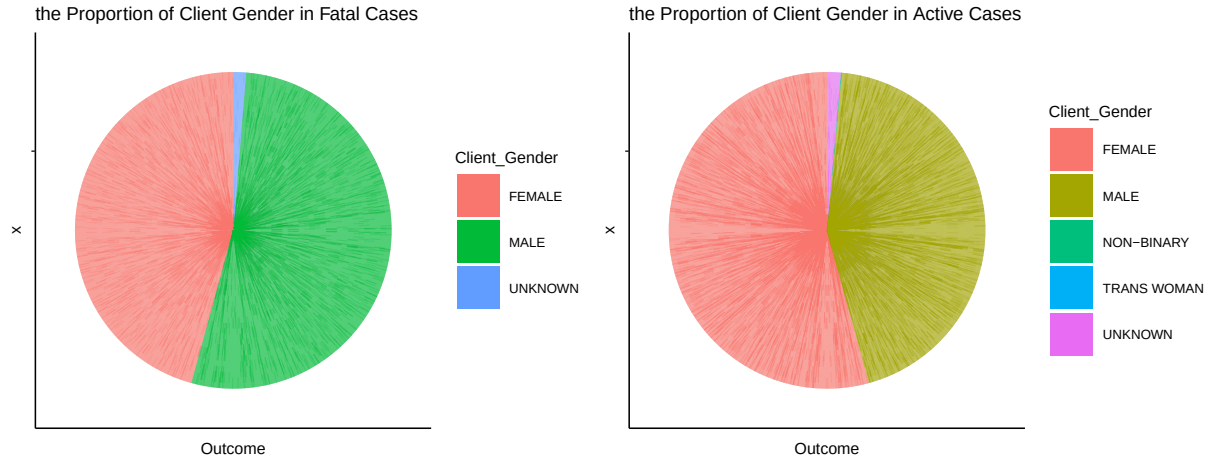

Figure 3: COVID-19 Fatal and Active Cases by Client Gender

To make a further description, however, there are observable differences in proportion of gender in various COVID-19 cases outcomes. As shown in figure 3, death rate is higher in male, even if all the clients with unknown gender are attribute to males. It indicates that male may suffer more severe consequences from infection of CPVID-19 than female. On the other hand, female account for the largest proportion of active cases until Feb 2nd, 2022. If not because of occasional outbreaks, this could indicate that the strains currently circulating are more likely to infect female, or that female are less likely than male to recover from COVID-19 infection.

Source of Infection

The Source of Infection variable contains information on the most likely way that cases acquired their COVID-19 infection is determined by examining several data fields including public health investigator's assessments, confirmed COVID-19 outbreaks associations, and reported risk factors. The dataset lists eight main sources of infection: travel, outbreaks (other settings), outbreaks (health institutions), outbreaks

(congregate settings), household contact, community, close contact, and no information. Besides, some of the information is still pending for closer investigation.

There is hierarchy to infer source of acquisition, if the public health investigator's assessment is absent ("Ontario Covid-19 Data Tool" 2022). Cases with episode dates before April 1 2020 treat travel as the most possible infection source rather than outbreak, and followed with household contact, close contact, community, and no information. However, for cases with episode dates on or after April 1 2020, travel is removed from the hierarchy with the following stay the same. This hierarchy increases the reliability of inference, as COVID-19 has not spread widely around the world until April 2020. If infected, it is likely to have been caused by travel to areas where the outbreak was concentrated. However, since April 2020, due to the rise of local epidemic, countries have adopted lockdown policies and circuit breakers, making tourism no longer a key route of transmission.

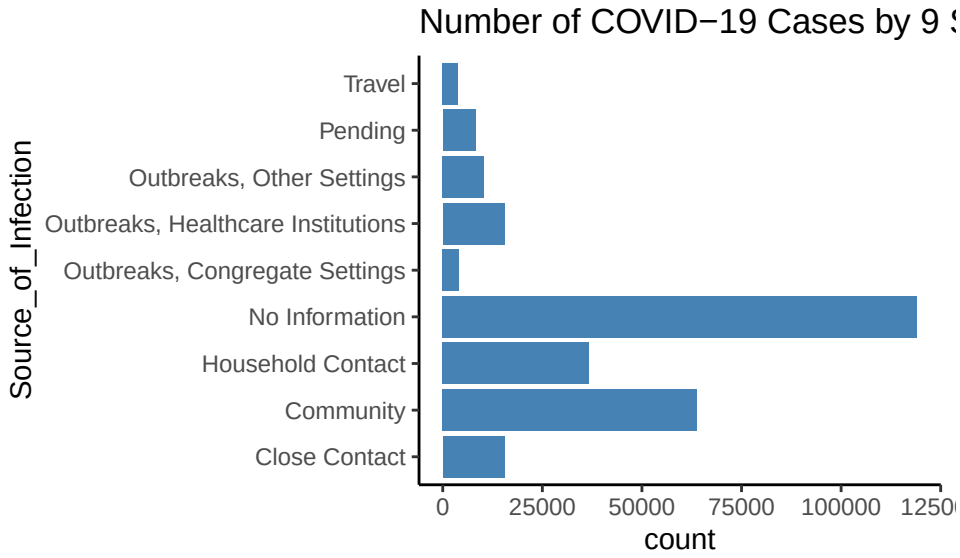

Figure 4: Bar Plot of COVID-19 Cases by Source of Infection

According to figure 4, most cases cannot trace its source of infection, which can partly explain the difficulty of preventing COVID-19 transmission. Community is the second most frequently sources of infection, which may caused by its definition — "Cases who did not travel outside of Ontario, did not identify being a close contact with a COVID-19 case, and were not part of a known confirmed COVID-19 outbreak" ("Ontario Covid-19 Data Tool" 2022) — that leave a huge space for unidentified source of cases and attributes them to the source of community. The proportions of outbreaks in total and household contact are relatively equal. An intuitive explanation is that clients get infected from outbreaks simultaneously passes the virus onto their households.

Time Lag

The 'Time Lag' variable in the dataset refers to the time gap between episode time and reported time, which represents the COVID-19 information timeliness and effectiveness of collection. It is calculated by subtracting episode date from report date in unit of days.

Since the outliers are extremely large compare to the median and 1st & 3rd quantile, figure 5 takes a log of the scale to improve the readability of the data.

The median is about 3 days to detect a new COVID-19 case by the Toronto Public Health. Senventy-five percent of cases can be reported within five days. There is an average of 365 new cases were reported in Toronto each day from 2020 to 2022 using the 277,473 observations in the dataset divided by 760 days within two years and two months. A 3 days of delay in reporting of cases means more than one thousand of clients

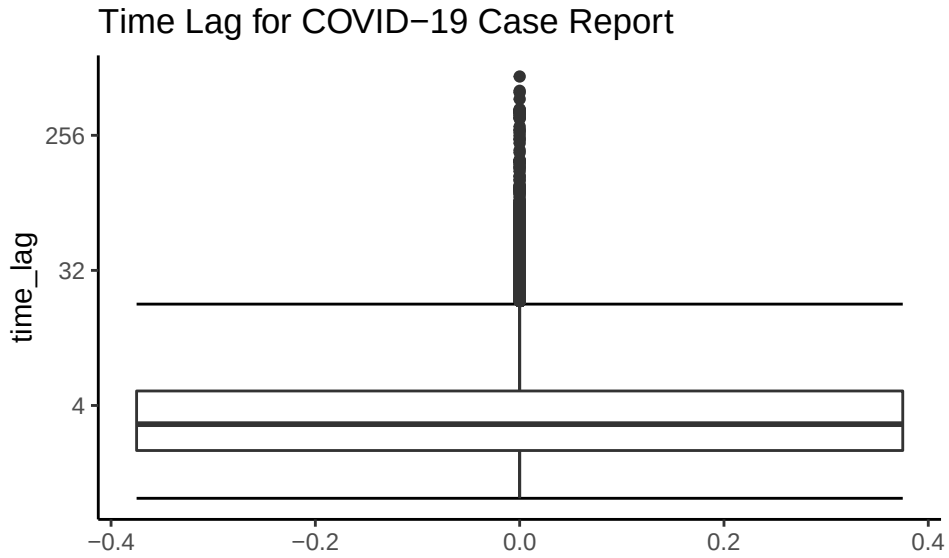

Figure 5: Box Plot of Time Lag for COVID-19 Case Report

cannot be managed systematically to avoid large outbreaks. Besides, there are some extreme outliers with time lag of more than a year, but hard to determine the causes.

Reference

Anne Kimball, MSPH; Melissa Arons, MD; Kelly M. Hatfield. 2020. "Asymptomatic and Presymptomatic Sars-Cov-2 Infections in Residents of a Long-Term Care Skilled Nursing Facility — King County, Washington, March 2020." *MMWR. Morbidity and Mortality Weekly Report* 69. https://www.cdc.gov/mmwr/volumes/69/wr/mm6913e1.htm?s_cid=mm6913e1_w.

"City of Toronto." 2022. https://www.toronto.ca/home/covid-19/.

"COVID-19 Canada | Ctv News | Coronavirus." 2022. https://www.ctvnews.ca/health/coronavirus/usteaserjan2-1.5251301.

Gelfand, Sharla. 2020. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

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

Lavezzo, Franchin, E. 2020. "Suppression of a Sars-Cov-2 Outbreak in the Italian Municipality of Vo'." *MMWR. Morbidity and Mortality Weekly Report*. https://doi.org/https://doi.org/10.1038/s41586-020-2488-1.

Mizumoto, Kenji, Katsushi Kagaya, Alexander Zarebski, and Gerardo Chowell. 2020. "Estimating the Asymptomatic Proportion of Coronavirus Disease 2019 (Covid-19) Cases on Board the Diamond Princess Cruise Ship, Yokohama, Japan, 2020." *Eurosurveillance* 25. https://doi.org/10.2807/1560-7917.es.2020.25.10.2000180.

"Ontario Covid-19 Data Tool." 2022. https://www.publichealthontario.ca/en/data-and-analysis/infectious-disease/covid-19-data-surveillance/covid-19-data-tool.

"Ontario Newsroom." 2022. https://news.ontario.ca/en/backgrounder/1000159/roadmap-to-reopen.

Pedersen, Thomas Lin. 2020. *Patchwork: The Composer of Plots*. https://CRAN.R-project.org/package=patchwork.

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

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

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

Wickham, Hadley, and Dana Seidel. 2020. *Scales: Scale Functions for Visualization*. https://CRAN.R-project.org/package=scales.

"World Health Organization." 2020. https://www.who.int/director-general/speeches/detail/who-director-general-s-opening-remarks-at-the-media-briefing-on-covid-19---11-march-2020.

"World Health Organization." 2022. https://covid19.who.int.