# Analysis of Covid-19 cases in Toronto, Canada (2020-2024)

Jim Zhang

The main aim of the current report is to identify the factors which influenced the trends in outbreak associated cases in Toronto, Canada. The findings from initial trend analysis suggests that the overall resolved cases in Toronto were more than fatal cases. Regression analysis confirms that the likelihood of outbreak associated cases are higher among individuals in higher age groups as compared to the young people. Also, household, and other settings are the prominent sources of infection when the cases are mostly associated with an outbreak. However, the main limitation of the analysis is lack of discussion of the results in contrast with other cities of Canada to present more robust findings.

## Table of contents

# 1. Introduction

The Covid-19 pandemic cases in Toronto, Canada, surged because of under-reporting of the cases which led to unwarranted growth in the infection cases. The results indicate that overall reported community cases were 18 and most of individuals were those who sought early preventive care (Desta et al. 2023). Other studies highlighted that most of the affected population in Toronto comprised professional individuals whereby the professions did not allowed opportunities of remote work. Thus, the labor market in Toronto was severely affected by the lockdown and social distancing measures. In addition, many care workers were exposed to transmission and mortality risk because of close contact with the patients (Rao et al. 2021). (McKenzie 2021) attempted to explore socio-economic disparities in Toronto following an increase in the number of cases. However, the findings suggested a lack of information and uniformity pertaining to the disclosure of confirmed cases by the individuals.

In other words, the disclosure of confirmed cases in Toronto remained uneven as many people failed to provide accurate information on the source of infection. Due to these discrepancies, many unreported cases went unnoticed which led to the outbreak and fast spread of the virus. These issues raised several questions regarding the credibility of reported cases in Toronto. Therefore, the current project aims to identify whether the identified cases across different age groups, source of infection, gender, and previous medical care were associated with the likelihood of outbreak or sporadic spread of the pandemic in Toronto, Canada.

The current project structures by initially outlining the information about the data and the process of data cleaning/transformation. The next step discusses the key findings obtained from the analysis using summary statistics and logistic regression analysis to answer the key research question. The next segment of the project critically discusses the findings from the analysis followed by a brief restatement of the main findings in the conclusion segment.

# 2. Data

## 2.1 Software and Packages

The current project uses R software (2023) for the preparation, cleaning, analysis, plotting, and modelling of the data associated with Covid-19 cases in Toronto, Canada. To obtain the dataset, ‘opendatatoronto’ (Gelfand 2022) package is used whereby the data from January 2020 to February 2023 is downloaded. dplyr and tidyverse packages are used for the cleaning and tidying up of the data for ease in data analysis. fastDummies (Kaplan 2023) package is used to create dummy variables of the main variables such as gender, outbreak, classification, source of infection, outcome, and status of medical assistance history. ggplot2 (Wickham 2016) package is used for the generation of relevant graphs, knitr (Xie 2023) package is used for making tables, and broom (Robinson, Hayes, and Couch 2023) package is used for cleaning tidy tibbles of the models.

### 2.2 Data Source

The data is collected from Open Data Catalogue provided by Toronto Public Health. The data set records information regarding Covid-19 cases in Toronto, Canada from January 2020 to February 2023(“Open Data Dataset — Open.toronto.ca”) (“Open Data Dataset — Open.toronto.ca”).[1]

### 2.3 Data Cleaning

The data obtained for analysis contains 32000 observations and 15 variables associated with the location, severity, status of infection, source of infection and outcome of the disease. The dataset does not contain any incorrect information and missing values. However, the variable name of ‘ever incubated’ is misspelled as “ever intubated” which is corrected. In addition, the dummy variables associated with “source of infection”, “classification”, “client gender”,“outcome”,“ever hospitalized”,“ever in ICU”, and “ever incubated” are generated for ease in data analysis. After the creation of dummy variables, the total number of variables increased from 15 to 41 in the data set. The dummy variables follow binary coding whereby 1 indicates the presence of a certain variable classification while 0 suggest otherwise.

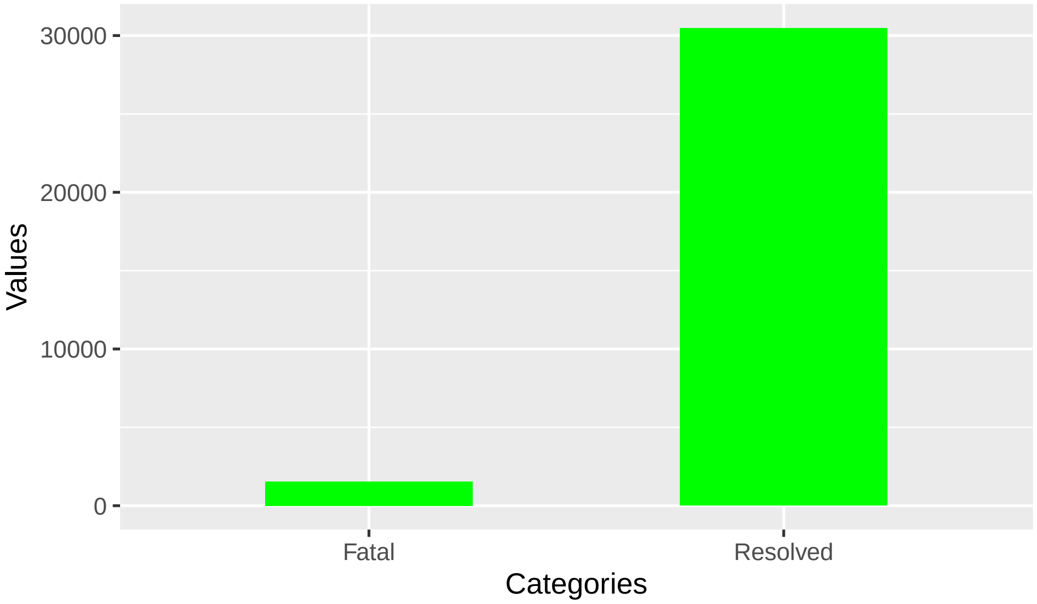
## 3. Findings

### 3.1 Trends in Covid-19 Cases Outcomes

Figure 1 indicates that the growth of Fatal outcomes (1534) is less than the cases with resolved (30466) outcomes.

[1] Code can be found at: https://github.com/JimZhang23/Outbreak_Toronto_2024.git

Figure 1: Bar Plot of Covid−19 Cases Outcome



## 3.2 Summary Statistics

Table 1 indicates the summary statistics of key variables of interest such as if the case is associated with an outbreak, age group of the respondents, source of the infection, classification of the case, gender of the client, main outcome of case after treatment, and the status of the client (hospitalized, ICU, and incubated).

Table 1: Summary Statistics

| Key Variable | Frequency |
|---|---|
| Outbreak Associated | |
| • Yes | 8039 |
| • No | 23961 |
| Age Group | |
| • 19 and Younger | 3196 |
| • 20 to 29 Years | 6288 |
| • 30 to 39 Years | 5163 |
| • 40 to 49 Years | 4401 |
| • 50 to 59 Years | 4698 |
| • 60 to 69 Years | 3019 |
| • Others | 5235 |
| Source of Infection | |
| • No Information | 8314 |

| Key Variable | Frequency |
|---|---|
| • Household Contact | 6736 |
| • Outbreaks, Healthcare Institutions | 6063 |
| • Community | 4834 |
| • Close Contact | 3121 |
| • Outbreaks, Other Settings | 1356 |
| • Other | 1576 |
| Classification | |
| • Confirmed | 30404 |
| • Probable | 1596 |
| Gender | |
| • Female | 16609 |
| • Male | 15241 |
| • Non-Binary | 1 |
| • Other | 7 |
| • Transgender | 6 |
| • Unknown | 136 |
| Outcome | |
| • Fatal | 1534 |
| • Resolved | 30466 |
| Hospitalized | |
| • Yes | 2712 |
| • No | 29288 |
| ICU | |
| • Yes | 557 |
| • No | 31443 |
| Incubated | |
| • Yes | 358 |
| • No | 31642 |

The findings suggests that there are 23961 cases of no outbreak and 8039 cases of outbreak, 30404 confirmed incubated cases, and 1596 probable cases. There are 16609 females in the sample data set, 15231 males, 1 non-binary, 6 transgenders, and 7 individuals from other genders. Among 32000 cases, 1534 fatal outcome cases are there while 30466 cases are resolved. Also, there are 299288 cases with no hospitalization record while 2712 cases are the cases of hospitalization. Around 31443 cases reported were not admitted in the ICU while 557 were admitted to the ICU. Almost 31642 cases of no incubation were reported while 358 incubation cases were reported. The findings suggest that the source of infection remained unknown for 8314 reported cases while household contact was the most prominent reason of infection source in 6736 cases. In terms of the age group, most of the cases identified in the age group of 20 years to 29 years (6288) and 30 to 39 years (5163), excluding “Others” category.

## 3.3 Logistic Regression

In this project, logistic regression modelling is conducted whereby the dependent variable is "whether the reported case is outbreak or sporadic" which takes 0 if the registered case is sporadic and 1 if the case is an outbreak. Logistic regression model is suitable to model the relationship between variables whereby the target variable is binary (Sperandei 2014). In this case, the main predictors are age group of the respondents, classification, source of infection, gender, outcome, and the status of medical admission history of an individual.

The findings suggest that the likelihood of an outbreak case among individuals with 20 to 29 years of age is 0.37 times higher than the individuals in other age groups. However, the likelihood of a case being associated with outbreak is 0.89 times higher among individuals with 70 to 79 years of age as compared to the other age groups. In terms of the p-value approach, it is observed that the outbreak cases are significantly associated when the source of infection is ether household or other settings. In other words, the findings suggest that the likelihood of outbreak associated cases are 0.39 times higher when the source of infection is household, compared to no information of the source. Also, the likelihood of outbreak associated cases are 2216.29 times higher when the source of infection is other settings, compared to no information of the source. As compared to confirmed cases, the likelihood of outbreak associated cases is 1.52 times higher when the classification of cases is probable. As compared to females, the likelihood of outbreak cases is 1.11 times higher when the patient is male. In case of the outcome of the cases, the likelihood of resolved cases being outbreak associated is 4.57 times higher than the fatal cases. The likelihood of outbreak cases is 1 time higher in case the individual is hospitalized as compared to the individual who is never hospitalized. The likelihood of outbreak associated cases is 0.25 times higher if the individual has been to ICU as compared to the individual who are never in ICU. At last, the findings strongly suggest that the likelihood of outbreak associated cases is 19 times higher among individuals who are incubated as compared to those who are never incubated.

Table 2: Logistic Regression Model

| Term | Estimate | Std. Error | Statistic | P.value |
|---|---|---|---|---|
| (Intercept) | -6.3493 | 2.1812 | -2.9109 | 0.0036 |
| Age: 19 and younger | 0.4294 | 1.9367 | 0.2217 | 0.8245 |
| Age: 20 to 29 Years | -0.9905 | 1.9352 | -0.5118 | 0.6088 |
| Age: 30 to 39 Years | -1.5951 | 1.9356 | -0.8241 | 0.4099 |
| Age: 40 to 49 Years | -1.3691 | 1.9373 | -0.7067 | 0.4798 |
| Age: 50 to 59 Years | -1.8503 | 1.9359 | -0.9558 | 0.3392 |
| Age: 60 to 69 Years | -1.7225 | 1.9463 | -0.8850 | 0.3761 |
| Age: 70 to 79 Years | -0.1245 | 2.0406 | -0.0610 | 0.9513 |
| Age: 80 to 89 Years | -0.9673 | 2.1453 | -0.4509 | 0.6521 |
| Age: 90 and older | -2.2172 | 2.1660 | -1.0236 | 0.3060 |

| Term | Estimate | Std. Error | Statistic | P.value |
|---|---|---|---|---|
| Source of Infection: Community | -16.7251 | 671.9821 | -0.0249 | 0.9801 |
| Source of Infection: Household Contact | -0.9417 | 0.4138 | -2.2755 | 0.0229 |
| Source of Infection: No Information | -16.8034 | 512.6703 | -0.0328 | 0.9739 |
| Source of Infection: Outbreaks, Congregate Settings | 30.9247 | 1631.8870 | 0.0190 | 0.9849 |
| Source of Infection: Outbreaks, Healthcare Institutions | 29.6997 | 595.9290 | 0.0498 | 0.9603 |
| Source of Infection: Outbreaks, Other Settings | 7.7036 | 0.3331 | 23.1236 | 0.0000 |
| Source of Infection: Travel | -0.2157 | 0.7834 | -0.2753 | 0.7831 |
| Classification: PROBABLE | 0.4165 | 0.3795 | 1.0976 | 0.2724 |
| Gender: MALE | 0.1000 | 0.1433 | 0.6980 | 0.4852 |
| Gender: NON-BINARY | -0.7085 | 48199.83 | 0.0000 | 1.0000 |
| Gender: OTHER | -9.0655 | 817.2964 | -0.0111 | 0.9912 |
| Gender: TRANSGENDER | -14.8174 | 17754.46 | -0.0008 | 0.9993 |
| Gender: UNKNOWN | 1.1037 | 1.1437 | 0.9650 | 0.3346 |
| Outcome: RESOLVED | 1.5193 | 0.9664 | 1.5721 | 0.1159 |
| Ever Hospitalized: Yes | 0.0068 | 0.5534 | 0.0123 | 0.9901 |
| Ever in ICU: Yes | -1.3533 | 1.0923 | -1.2389 | 0.2154 |
| Ever Incubated: Yes | 2.9718 | 1.3403 | 2.2172 | 0.0266 |

## 4. Discussion

The key takeaway from the analysis suggests that outbreak cases are more prominent among individuals in higher age group as compared to younger generations whereby the cases are mostly sporadic and less severe. It is further observed that most of the reported cases of Covid-19 were mainly outbreak associated when the source of infection was either household or other areas. These findings are inconsistent with the current research which asserts that most of the outbreak cases in Toronto were observed among healthcare and front line workers as they are in direct contact with the patients. Therefore, these individuals and health care workers were at higher risk of transmission of the disease and the risk of death (Rao et al. 2021). These findings suggest that under reporting of Covid-19 cases in Toronto was mainly driven by lack of understanding in public to grasp the source of transmission of the infection which led to the unwarranted growth of the pandemic.

The results also suggests that males were at higher risk of infection as compared to females. Also, the resolved cases in Toronto were mostly associated with outbreak as compared to the confirmed cases which indicates efficient response mechanism of the healthcare authorities to contain the pandemic (Desta et al. 2023). The findings suggests that individuals with previous hospital admissions and incubation were most likely to contact outbreak associated cases as compared to those individuals who were neither hospitalized nor incubated.

## 5. Limitation of the Analysis

The main limitation of the analysis is lack of comprehensive evaluation of the state of Covid-19 pandemic and its treatment in Toronto as compared to the other cities of Canada. Another main limitation is lack of identification of the significant effects of individual age and gender on the likelihood of a Covid-19 case to be "outbreak" associated. In order to improve the analysis, data pertaining to pandemic from other cities of Canada can be obtained for a detailed and comprehensive assessment.

## 6. References

Desta, Binyam N, Sylvia Ota, Effie Gournis, Sara M Pires, Amy L Greer, Warren Dodd, and Shannon E Majowicz. 2023. "Estimating the Under-Ascertainment of COVID-19 Cases in Toronto, Ontario, March to May 2020." *Journal of Public Health Research* 12 (2): 227990362311741. https://doi.org/10.1177/22799036231174133.

Gelfand, Sharla. 2022. "Opendatatoronto: Access the City of Toronto Open Data Portal." https://sharlagelfand.github.io/opendatatoronto/.

Kaplan, Jacob. 2023. "fastDummies: Fast Creation of Dummy (Binary) Columns and Rows from Categorical Variables." https://github.com/jacobkap/fastDummies.

McKenzie, Kwame. 2021. "Socio-Demographic Data Collection and Equity in Covid-19 in Toronto." *EClinicalMedicine* 34 (April): 100812. https://doi.org/10.1016/j.eclinm.2021.100812.

"Open Data Dataset — Open.toronto.ca." https://open.toronto.ca/dataset/.

———. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

R Core Team. 2023. "R: A Language and Environment for Statistical Computing." https://www.R-project.org/.

Rao, Amrita, Huiting Ma, Gary Moloney, Jeffrey C. Kwong, Peter Jüni, Beate Sander, Rafal Kustra, Stefan D. Baral, and Sharmistha Mishra. 2021. "A Disproportionate Epidemic: COVID-19 Cases and Deaths Among Essential Workers in Toronto, Canada." *Annals of Epidemiology* 63 (November): 63–67. https://doi.org/10.1016/j.annepidem.2021.07.010.

Robinson, David, Alex Hayes, and Simon Couch. 2023. "Broom: Convert Statistical Objects into Tidy Tibbles." https://broom.tidymodels.org/.

Sperandei, Sandro. 2014. "Understanding Logistic Regression Analysis." *Biochemia Medica*, 12–18. https://doi.org/10.11613/bm.2014.003.

Wickham, Hadley. 2016. "Ggplot2: Elegant Graphics for Data Analysis." https://ggplot2.tidyverse.org.

Xie, Yihui. 2023. "Knitr: A General-Purpose Package for Dynamic Report Generation in r." https://yihui.org/knitr/.