Visualizing risk factors of COVID-19 infections within Toronto*

Justin Abando

3 February 2023

Data related to COVID-19 cases sourced from OpenDataToronto was observed to draw conclusions of the impact of the COVID-19 pandemic in Toronto. It was found that the children and young adults are more often infected with COVID-19, however the older demographic proportionally suffers more despite having fewer recorded cases due to the observation of more deaths. Younger age groups contract the virus prominently in household settings, whereas older age groups contract it through healthcare institution outbreaks. In addition, it was found that a larger proportion of the adults and older demographics likely transmit the disease to their children. Overall, this data demonstrates the epidemiological impact that COVID-19 has brought to the Toronto community over the course of the past few years.

Introduction

Pandemics have occurred throughout history, and the most recent pandemic that took the entire world by storm was caused by the severe acute respiratory syndrome coronavirus 2, or SARS-CoV-2. Those infected with this virus would then be affected by COVID-19, which was first documented in China in December of 2019, but rapidly spread and developed throughout the world. In March of 2020, the World Health Organization classified the spread of COVID-19 as a global pandemic (Cucinotta and Vanelli 2020). The number of cases from then onward demonstrated the weaknesses in many national and local health infrastructures, as the number of confirmed cases reached as far as hundreds of millions across the world. Even after three years in 2023, the effects of the pandemic still loom with SARS-CoV-2 variants emerging and infecting populations as well.

*Code and data are available at: https://github.com/justabando/COVID-19-Cases-in-Toronto

Controlling the COVID-19 pandemic is an ongoing effort that has exerted its effect on public health over time in the past few years. As a result of the pandemic, COVID-19 has been subject to many epidemiological studies that asserted its impact in both small and large scale contexts. There are several risk factors at play that contribute to the growth of a pandemic such as COVID-19. Individual lifestyles and community environments play a role assessing the spread of the pandemic within Toronto communities. Established relationships exist between COVID-19 disease transmission and prominent risk factors such as social environment and age. Despite quarantining efforts, the pandemic managed to infect many people which requires some analysis to uncover the underlying reasons why COVID-19 was still able to spread to such a degree.

To better understand how to manage the COVID-19 pandemic's effect, analysis of its spread within large communities such as Toronto is useful. Larger cities are primed with a diverse set of circumstances, with many possible situations to potentially transmit or contract the disease. The data observations within this report were intended to communicate the spread of COVID-19 within specific settings, and its overall impact within the population.

It was found that children and younger adults were more susceptible in contracting COVID-19, whereas older adults and the elderly were more susceptible to death as a complication of COVID-19 despite having fewer cases relative to the younger demographic. Observation of the social settings in which COVID-19 could be transmissible demonstrated that children and young adults contract COVID-19 more through household settings and close contact, and elderly populations contracting it through healthcare institution outbreaks. Adults were shown to vary in their setting of disease contraction, and were spread relatively evenly through household contact, community, and healthcare outbreaks.

Data

To investigate and visualize the impact of the possible risk factors, the tidyverse package (Wickham et al. 2019) and dplyr package (Wickham et al. 2023) were utilized through R (R Core Team 2022) to organize and display relevant observations from the OpenDataToronto data set. OpenDataToronto is a public access database with municipal records consisting of topics which include, but are not limited to education, healthcare, crime, and the environment. To import the data set directly into R, the opendatatoronto package was used (Gelfand 2022). The data set was cleaned to be easier to work with in R (R Core Team 2022) using the janitor package (Firke 2021) and was visualized using the ggplot (Wickham 2016) and ggmosaic (Jeppson, Hofmann, and Cook 2021) packages.

Data Collection

The data set containing information of COVID-19 cases is freely available through OpenData-Toronto, with the most recent update of the data set being on January 25th, 2023.

The pertinent areas in the data set that were categorically measured and compared were:

- Age groups
- Classification
- Outcome
- Sources of infection

The age groups were divided in intervals of 10 years, with the exception of those that are 19 and younger and those that are 90 and older. Classification refers to whether or not the person was either likely or confirmed to be infected with COVID-19. Outcome refers to whether or not the person was able to recover from COVID-19, or died while infected by it. Sources of infection refer to the environmental settings that the person was present in when the disease was likely contracted to them.

During the cleaning process there were observations that were missing information on age groups and sources of infection, and were excluded to create uniform visualizations without missing data to prevent confusion.

Data Observations

An observation to be made is in regards to the frequency of Toronto residents that have caught, or likely caught COVID-19. More importantly, the specific age groups of those that caught the disease are of interest, as clues to its potential spread and impact can be visualized. A general trend can be drawn from this and is demonstrated in Figure 1.

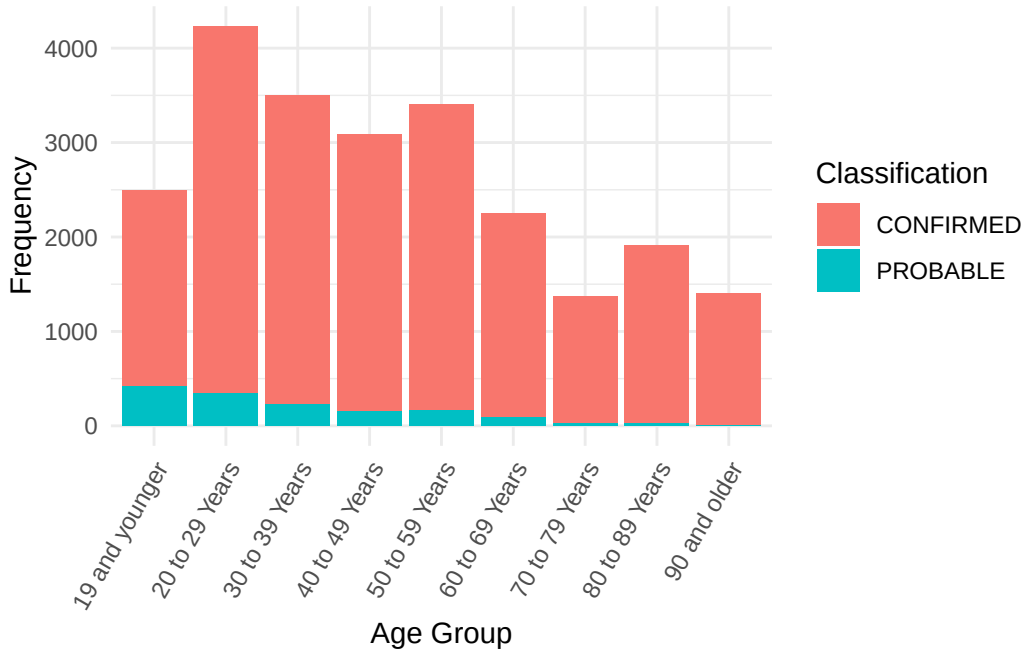

Figure 1: Frequency of confirmed and probable COVID-19 cases based on age grouping.

The data leads to the suggestion that younger age groups and adults tend to exhibit more confirmed and probable cases of COVID-19 in comparison to older groups (Figure 1). A sensible reason for this observation is that the younger groups have social obligations to attend to, whether it would be work or school. As such, they require to be outside in social settings more. If this data set was analyzed during quarantine, a more uniform distribution of cases would be visualized instead. However, this data set accounts for all currently recorded cases, so as schools and workplaces began to shift back into requiring physical presence, the frequency of cases are skewed as such.

In terms of the outcomes of those that were infected with COVID-19, the data suggests that older age groups are more susceptible to its fatal effects (Figure 2).

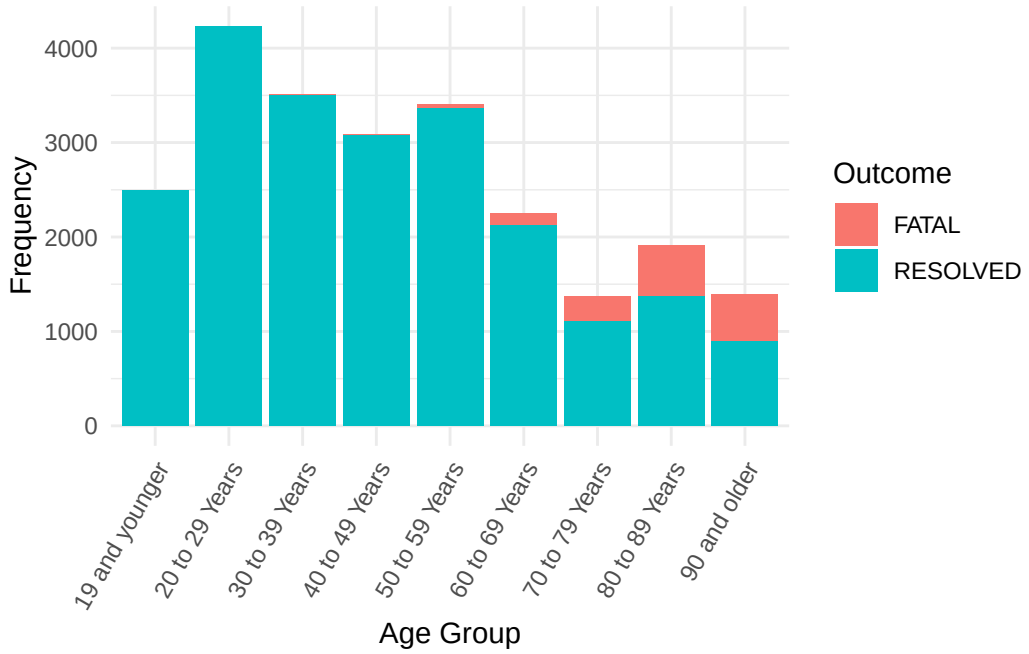

Figure 2: Frequency of resolved or fatal COVID-19 outcomes based on age grouping.

As outlined in Figure 2, the observations are still based on age groups, but instead focus on the outcome of disease progression. It was observed that most cases of COVID-19 are typically resolved within younger age groups and adults, however as the age groups increasingly get older, the prevalence of death begin to show.

The data suggests that the older demographic range (notably from 60 years and onward) are more susceptible to death due to COVID-19. However, it must be taken into account that there are likely other comorbidities present in an older demographic, meaning that there are certainly other health issues that can contribute to death as it becomes natural with age. As a result, the observation that COVID-19 is more lethal to older groups than younger groups would be true only in a vacuum, as there will be other factors that were not taken into account in this data set. For example, idiopathic pulmonary fibrosis (IPF) and chronic obstructive pulmonary disorder (COPD) are respiratory diseases that become increasingly more prevalent with age (Rojas et al. 2015). The presence of such diseases puts those in the aging population at a much higher risk of contracting the COVID-19 as it is also a respiratory virus. Diseases that are linked to aging can be responsible for immunocompromising the elderly, leading to an increased risk serious complications such as death.

To take into account the possible sources of COVID-19 transmission within the population, data regarding the age groups of affected or potentially affected individuals are juxtaposed with the environmental setting in which they contracted it (Figure 3).

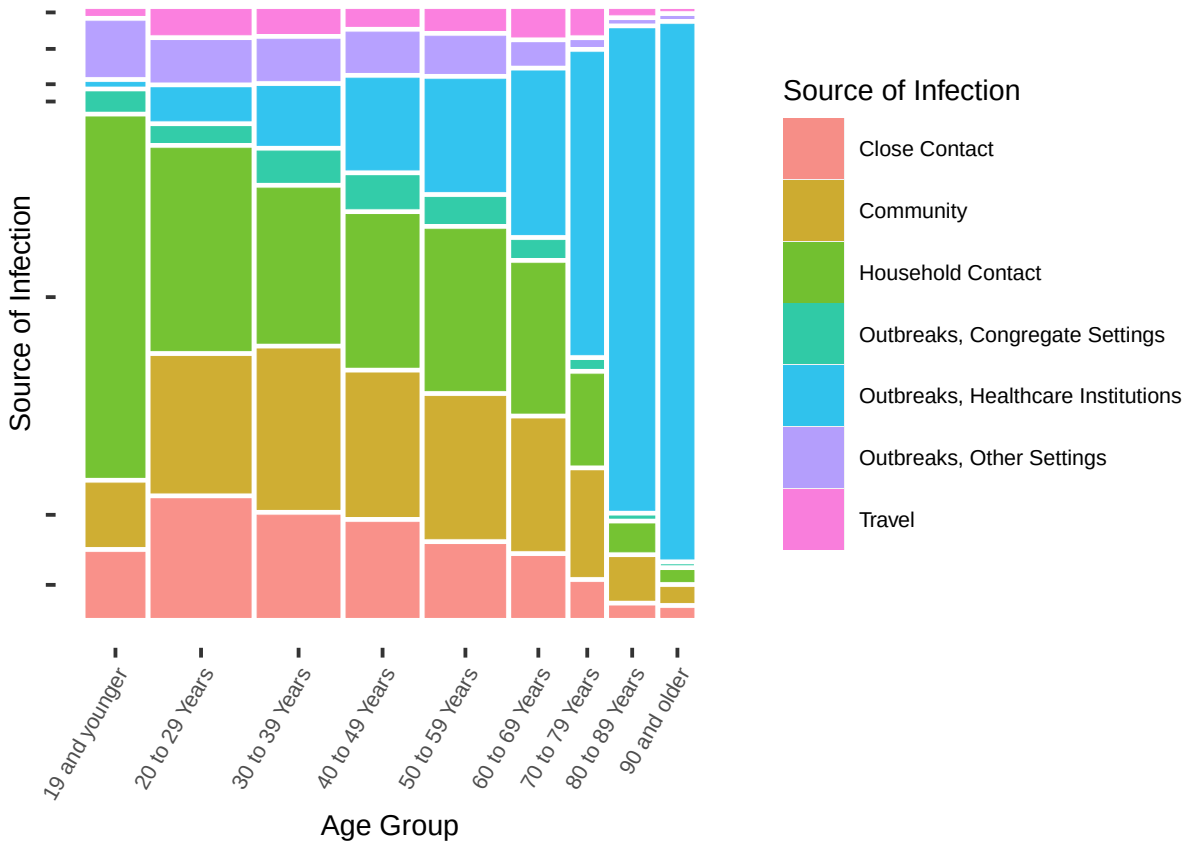

Figure 3: The proportion affected COVID-19 individuals based on the infection source.

Drawing conclusions off Figure 3, there is a dominant proportion of young adults and children that contracted COVID-19 in a household setting. By comparison, adults and elderly were observed to contract the disease more commonly through community settings and through close contacts. The elderly above 80 years old are seen to exhibit a dominant trend of contracting the virus within a healthcare institution outbreak.

An interesting observation can be seen with how children and young adults contract the virus most dominantly through household contact in comparison to any other age group (Figure 3). A reason for this observation can be explained that they contract COVID-19 through their parents, especially those that were considered essential workers during quarantine. This also explains why the adult demographic are infected more from community settings than children and the elderly. In general, there is a trend of cases originating from household contact and community settings becoming less common as the demographic gets older, and is compensated by an increase of cases originating from outbreaks in healthcare institutions.

Limitations

There are limitations and bias that skew the accuracy of the results from the data presented, and as a result may affect some of the conclusions. A large consideration has to be made in the accuracy of the screening processes that determine if an individual has COVID-19. This data set was created when the COVID-19 cases first started to appear in early 2020, and outdated screening protocols were used to assess COVID-19 infection since then. As the pandemic progressed, we learned more about SARS-CoV-2 and the risk factors that lead to a higher susceptibility of being infected with it, and so more robust and accurate screening protocols were used in practice. To how it affects the data set and results, the entries up to certain points will contain infection data at the time when screening protocols were not optimal. As a result, the data will not be completely consistent and there may be some observations that were false negatives or even false positives. Despite this, all of the data was still used to generate visualizations and conclusions were still drawn from the data set to reflect the current information present within it. Regardless, the data still provides a general idea of the epidemiological impact of COVID-19.

In addition, the data set excluded specific entries in the visualizations that were missing information of individuals in their age group or source of infection. This does affect the data as the visualizations are not completely representative of the entire data set, however this was done to maintain consistency in ensuring that every observation had the required information to be able to gather appropriate results.

References

Cucinotta, Domenico, and Maurizio Vanelli. 2020. "WHO Declares COVID-19 a Pandemic." *Acta Biomed.* 91 (1): 157–60.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*.

Jeppson, Haley, Heike Hofmann, and Di Cook. 2021. *Ggmosaic: Mosaic Plots in the 'Ggplot2' Framework*. https://github.com/haleyjeppson/ggmosaic.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Rojas, Mauricio, Ana L Mora, Maria Kapetanaki, Nathaniel Weathington, Mark Gladwin, and Oliver Eickelberg. 2015. "Aging and Lung Disease. Clinical Impact and Cellular and Molecular Pathways." *Ann. Am. Thorac. Soc.* 12 (12): S222–7.

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, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*.