Recovery from COVID-19 in Toronto & the Role of One's Demographics and Environment

Jennifer Do

06 January 2021

Abstract

The city of Toronto continues to observe increasing rates of mortality and morbidity due to the COVID-19 pandemic. Via contact tracing reportings collected by Toronto Public Health, this paper aims to determine the relationships between one's demographics, environment, and the probabilities to recover from COVID-19 infection using a logistic regression model. Age, gender, sources of infection, hospitalization and ICU admittance were all observed to have a significant impact on one's likelihood for recovery. These relationships further highlight the importance of apt municipal and institutional policies implemented for prevention and safety.

Keywords: COVID-19, Logistic Regression, Toronto

This report, code and data is hosted on https://github.com/dojennifer/sta304-ps5

Introduction

The COVID-19 pandemic has resulted in drastic spikes in mortality and morbidity rates among the public. As of November 23, 2020, the city of Toronto has implemented municipal policy for a "Grey–Lockdown" period to help reduce the spread of COVID-19 (City of Toronto, 2020).To record information relating to COVID-19 infection and outbreak occurrences, contact tracing has become an important tool for determining various epidemiological vectors of disease. These vectors may include host, environment and transmission of the virus (agent) based on the epidemiologic triad model (CDC, 2012). Contact tracing not only records information relating to the various vectors, but also enables for the identification, monitoring and educating of those who have been in contact with positively-infected individuals to determine potential cases and limit infectious spread (Public Health Ontario, 2020).

Based on recorded information from contact tracing, it is unclear whether one's demographics and environmental factors (such as the impact of location on viral transmission and the consequences of situating in certain locations) may influence one's recovery from COVID-19 infection. For example, impoverished neighbourhoods have 69% higher hospitalization rates, over double the ICU admission rates, and 52% higher death rates related to COVID-19 than well-off neighbourhoods (Public Health Ontario, 2020). Perhaps there exists individuals who may not be able to afford an adequate amount of masks or may be less educated on the risks of not wearing masks. Perhaps the surrounding hospitals have fewer physicians and resources to deal with the pandemic. The potential factors influencing COVID-19 transmission and recovery are numerous – these listed are but a few.

In this report, I aim to use logistic regression modeling to explore and determine whether variable factors including: age, gender, source of infection, hospitalization, and ICU admittance, may influence one's "resolved" status from COVID-19 in Toronto. According to Toronto Public Health (TPH), "Resolved" status reported indicates either recovery, no symptoms of COVID-19 shown after 14 days of initial symptoms and are not currently hospitalized. By exploring TPH's collected data on COVID-19 to observe how Toronto-based variables may influence "resolved" cases, this may enable for further understandings towards recovery,

and relationships between the epidemiological vectors for further safety, preventative and policy measures to be undertaken.

Data

The data set obtained for this study comes from Toronto's Open Data Portal : COVID-19 Cases in Toronto. As the data set used for this paper was last updated December 30, 2020, it had recorded all confirmed and probable cases starting from January 23 to December 28, 2020, as reported to TPH. The dataset contains details relating to demographics (i.e. age, gender), location (i.e. neighbourhood), source of infections, hospitalization information, severity information (i.e. ICU admittance and intubation) and outcomes for each case. The data was extracted and combined from the provincial communicable disease reporting system (iPHIS) along with Toronto's custom COVID-19 case management system (CORES). Data collection and extraction occurred on a weekly basis, with extraction occurring at 3pm on Mondays, and data postings by Wednesdays in any given week. These weekly updates enabled for better accurate representations of the recorded information.

To note, "Source of Infection" as a variable was recorded dependent on 8 potential acquisition sources including: travel, close contact, institutional settings, health care settings, community, pending source, unknown/missing, or N/A (i.e. Outbreak-associated cases). "Outbreak associated" cases were reported to contain institutional settings including but not limited to: long term care homes, retirement homes, hospitals homeless shelters and other congregate settings. In contrast, healthcare as a potential source of infection including clinical settings outside of hospitals such as: family physicians, dentists, opthalmologists, etc. Travel as a potential source was reported to record for infection due to travel outside of Ontario.

It must also be noted that "Outcome" as a variable was defined to be either "Fatal", "Resolved", or "Active". "Fatal", as suggested records cases with a fatal outcome (i.e. deceased). "Resolved" cases were reported to include cases who have "recovered" or were reported to be more than 14 days from symptom onset and are not currently hospitalized. "Active" cases reported all other cases.

Figure 1

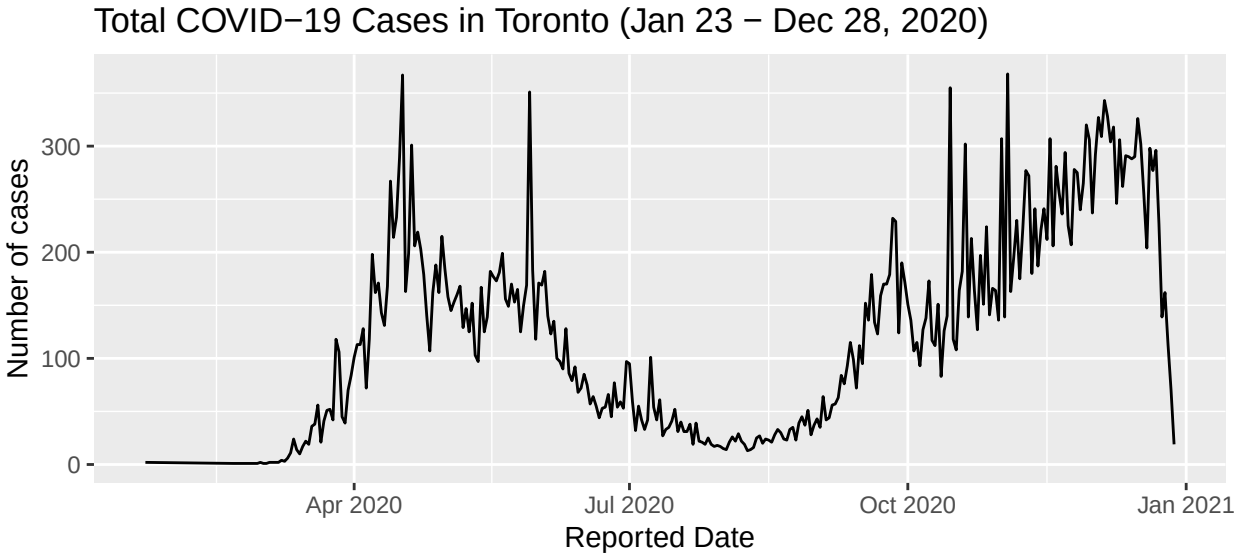


Figure 1 shows the total number of COVID-19 cases occurring in Toronto with the earliest case reported on January 23, 2020 and the last reported case on December 28, 2020. During this time period, TPH reported a cumulative 60,333 cases of COVID-19. The first wave of outbreaks is observed to have occurred

approximately between April to Mid July, while the second wave of COVID-19 starting from approximately September, has continued to rise in total number of cases.

To compare between other variables, figures modeling and exploring the dataset will be presented and discussed later on in the results section for evident visual contrast and comparison (Figures 2,4 and 6 in the Results section).

Model

To model whether demographic and environmental variables may influence recovery (i.e. obtain a "resolved" status) from COVID-19 infection, a logistic regression was performed on the TPH's data set: COVID-19 cases in Toronto. A binary outcome of 1 was designated for "resolved"/recovered cases and 0 for "non-resolved"/non-recovered cases as the response variable. Given that the paper aims to analyze for "resolved" and "recovered" outcomes given the variables, the "Fatal" and "Active" cases were excluded from modeling. This enabled for the regression output to interpret the demographic and environmental variables as probabilities for recovery. Here, the analysis of multiple variables and its outcome results would render meaningless in a linear regression model due to outcomes potentially falling outside a range of 0 -1 and provide limited context for interpretation of recovery. The model used was:

$$P(Resolved) = logit^{-1}(\beta_0 + \beta_1 age_i + \beta_2 gender_i + \beta_3 sourceofinfection_i + \beta_4 everhospitalized_i + \beta_5 everinICU_i)$$

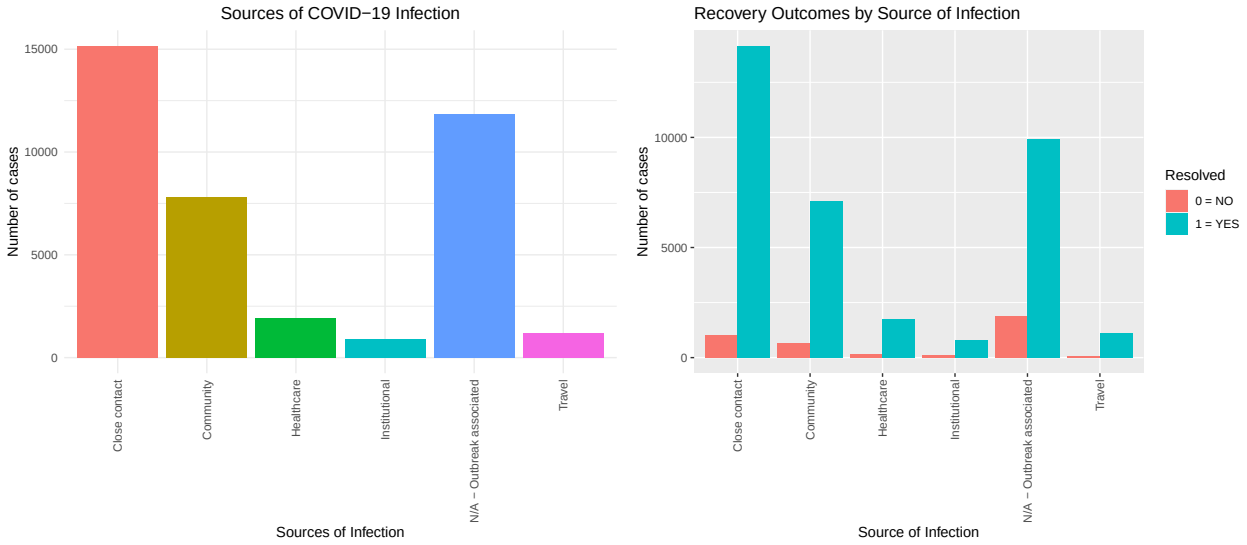
From TPH's data set, the explanatory variables chosen were: age, gender, source of infection, if a person was hospitalized, and if a person was ever admitted to the ICU. Age and gender variables enabled for further understanding relating to demographics, while source of infection, hospitalization and ICU admittance enabled for understanding relating to external environment. The data set had further variables including: current hospitalization, current ICU admittance, intubation, and neighbourhood. However, these variables were excluded from the model. As the variables for a case being ever hospitalized and/or ever in ICU contains cases that are currently hospitalized and currently admitted to the ICU, the exclusion of the current (hospitalized/in ICU) variables was undertaken to avoid oversampling. Data relating to intubation was also excluded as it is assumed that cases that have been intubated are both hospitalized and/or are in the ICU due to the severity of health decline. Neighbourhood as a variable was also excluded as the observation for the determination of acquisition of infection was included in chosen variable "Source of Infection", with "community" being a potential source.

Among the variable "Source of infection", cases classified as "Pending", or "Unknown/Missing" were removed due to unclear source of infection. As this paper aimed to analyze whether environmental factors may play a role onto recovery, the unclear determination of a source of infection would not provide for significant analysis. In the dataset, the classification for cases were either confirmed (i.e. patient had laboratory confirmation of the SARS-CoV-2 infection or had SARS-CoV-2 antibodies via a validated assay) or were probable of infection (i.e. patient did not have laboratory diagnosis, traveled to or from an infected area 14 days prior to symptom onset, had close contact with confirmed case, lived or worked in a facility known to have an outbreak, or have symptoms compatible with COVID-19 and had inconclusive laboratory results) (Ontario Ministry of Health, 2020). To ensure for accurate modeling related to recovery, probable cases were removed due to its uncertainty in COVID-19 infection.

Overall, from the original data set of 59,030 cases reported, after cleaning of the data set to match the criteria needed, 20,254 cases were removed for a total of 38,776 cases analyzed.

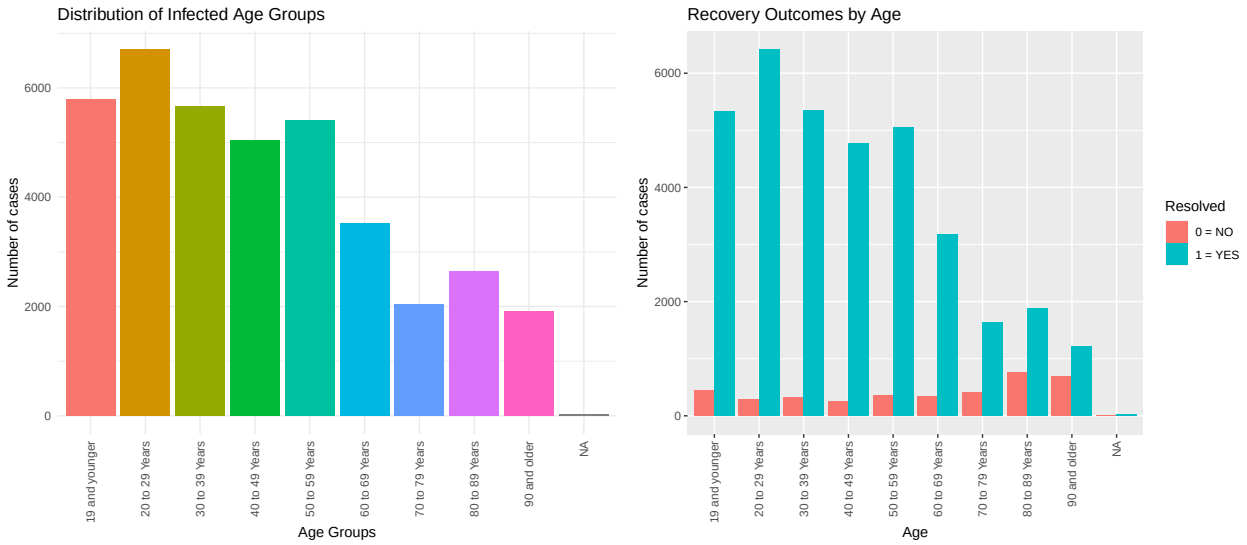
Results

Figures 2 & 3



Both Figures 2 & 3 showcases the total number of cases resulting from various sources of infection. Despite the majority of cases arising from close contact (i.e.close contact with a positively-infected person), as observed in Figure 2, it is also observed to have the highest recovery outcomes (Figure 3). Despite a large number of cases also sourced in unknown infections(i.e. N/A- Outbreak associated), the high rates of recovery are also observed. However, as the unknown source cannot be traced, this may explain this category as having the highest amount of cases for non-recovery.

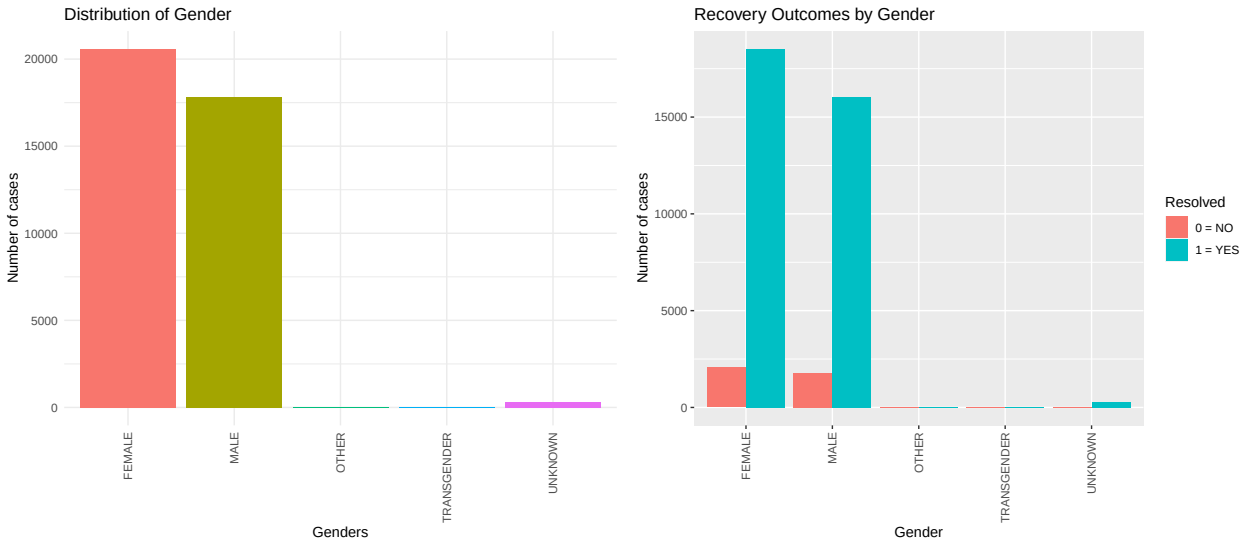
Figures 4 & 5



Both Figures 4 & 5 showcase the distribution of various age groups and the associated infected cases. In Figure 4, it is observed that persons aged 20-29 despite having the highest rate in infected cases (Figure 4), have the highest amount and proportion of cases recovered (Figure 5). In contrast to an older demographic (i.e. persons aged 90 and older), despite having the lowest number of infection cases (Figure 4), have

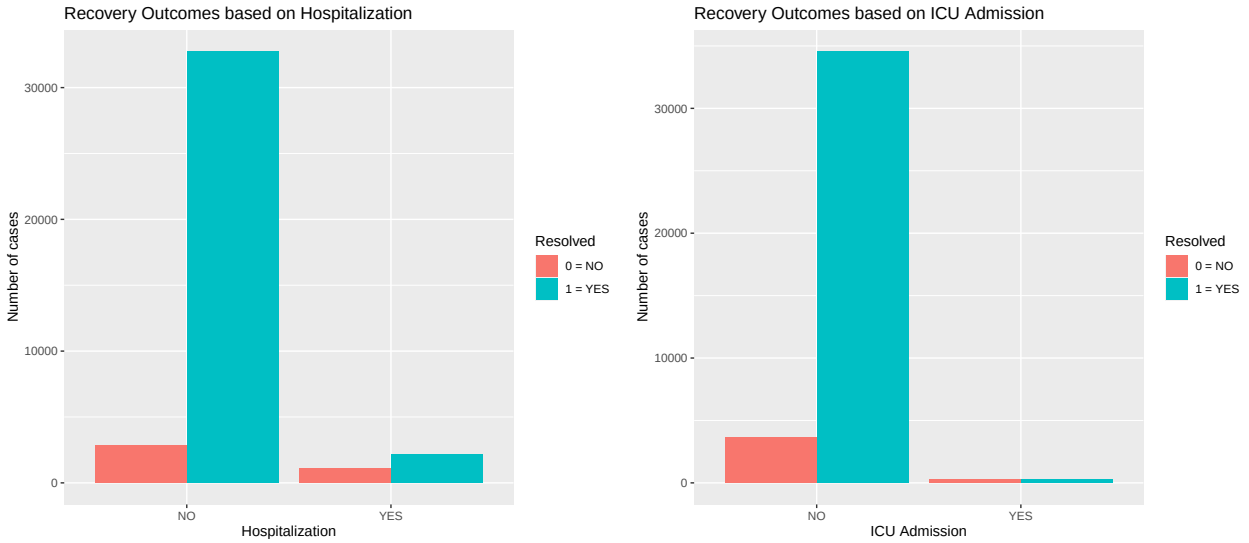
the lowest recovery outcomes (Figure 5). As populations grow older, it is observed that full recovery is diminished. This may be due to a growing number of co-morbidities contributing health decline as one ages.

Figures 6 & 7



Both Figures 6 & 7 showcase the distribution of positive cases based on gender. It is observed that though there are more females than males who are infected (Figure 6), there also exists a higher rate of recovered cases for females than males (Figure 7). As gender was self-reported, there exists other categories (i.e. "Other", "Transgender" and "Unknown"), however due to the limited cases reported in these categories, the results are seen to be statistically insignificant as discussed later.

Figures 8 & 9



In Figure 8, it is observed that cases that have been not been hospitalized are significantly more likely to recover from COVID-19 as opposed to cases that have been hospitalized. However it must be noted that a majority of cases have not been hospitalized. Among those that have been hospitalized, full recovery is seen to account for more than 50% of hospitalized cases.

Figure 9 showcases cases that have been admitted to intensive care unit (ICU) due to COVID-19 infection. It is observed that the overwhelming majority of cases that have not been admitted to ICU have full recovery. However, those that are admitted have significantly lower rates of recovery. This may be due to ICU admission being dependent on severe infection of COVID-19 and rapid declining health.

Table 1

Observations	36895 (3 missing obs. deleted)
Dependent variable	Outcome
Type	Generalized linear model
Family	binomial
Link	logit

$\chi^2(19)$	3342.33
Pseudo-R² (Cragg-Uhler)	0.18
Pseudo-R² (McFadden)	0.14
AIC	21137.17
BIC	21307.48

	Est.	S.E.	z val.	p
(Intercept)	2.60	0.06	44.07	0.00
‘Age Group‘20 to 29 Years	0.61	0.08	7.56	0.00
‘Age Group‘30 to 39 Years	0.38	0.08	4.76	0.00
‘Age Group‘40 to 49 Years	0.53	0.08	6.35	0.00
‘Age Group‘50 to 59 Years	0.33	0.08	4.22	0.00
‘Age Group‘60 to 69 Years	0.10	0.08	1.22	0.22
‘Age Group‘70 to 79 Years	-0.61	0.08	-7.37	0.00
‘Age Group‘80 to 89 Years	-1.16	0.08	-15.11	0.00
‘Age Group‘90 and older	-1.62	0.08	-20.18	0.00
‘Client Gender‘MALE	-0.13	0.04	-3.31	0.00
‘Client Gender‘OTHER	-0.32	1.07	-0.30	0.76
‘Client Gender‘TRANSGENDER	-1.30	1.10	-1.18	0.24
‘Client Gender‘UNKNOWN	-0.17	0.18	-0.93	0.35
‘Source of Infection‘Community	-0.18	0.06	-3.26	0.00
‘Source of Infection‘Healthcare	0.38	0.10	3.79	0.00
‘Source of Infection‘Institutional	-0.43	0.11	-3.78	0.00
‘Source of Infection‘N/A - Outbreak associated	-0.16	0.05	-3.04	0.00
‘Source of Infection‘Travel	0.88	0.17	5.31	0.00
‘Ever Hospitalized‘	-1.01	0.05	-18.68	0.00
‘Ever in ICU‘	-1.26	0.10	-12.65	0.00

Standard errors: MLE

After running the logistic regression model, this table displays the outcome values including estimates, standard errors, z-values and p values. Considering p-values that are less than 0.05 to be statistically significant, it is observed that most variables are significant. The exception for this includes variables for Gender(“Other”, “Transgender”, and “Unknown”) as well the variable "Age Group for 60-69 Years.

Figure 10

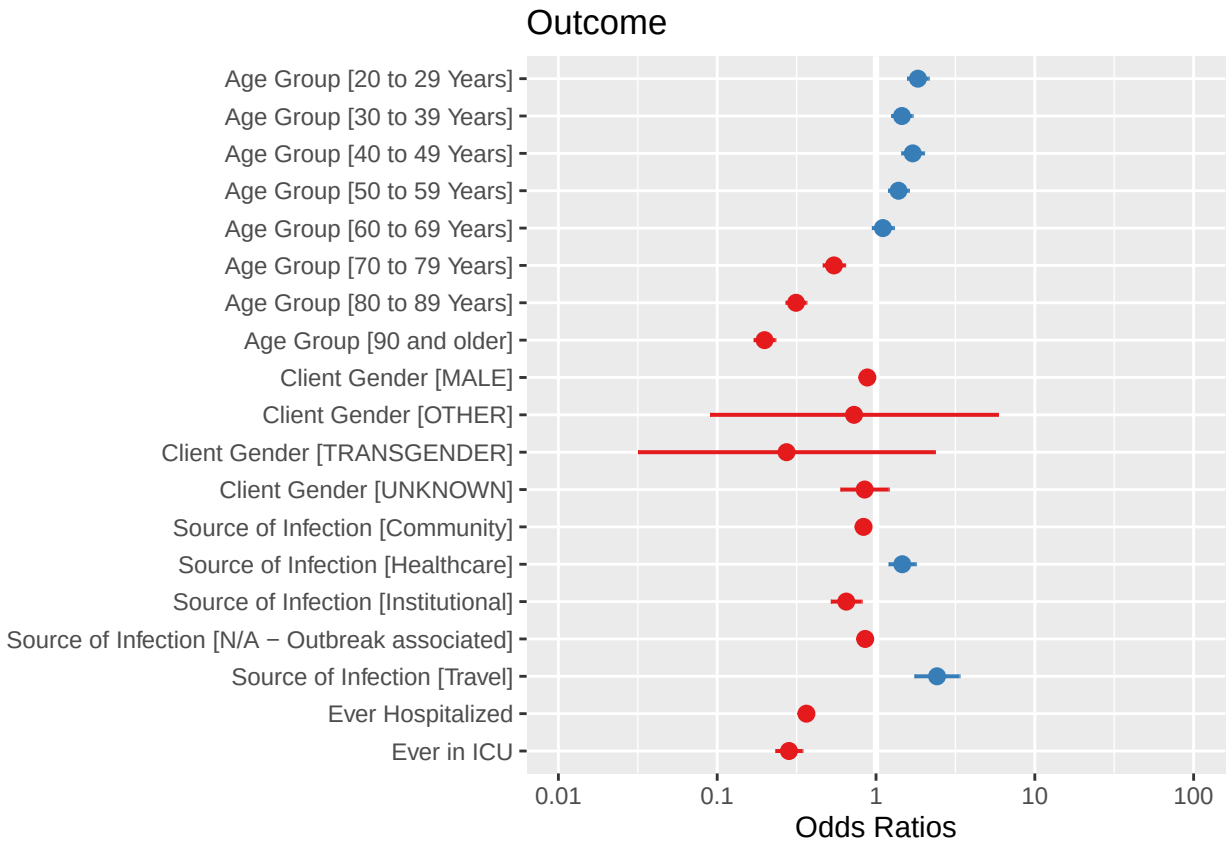


Figure 10 illustrates the Odds Ratios as determined from our logistic regression model. As odds ratios can help determine whether particular exposures are risk factors for outcomes, in this case, we are able to determine the magnitude and risk association of our explanatory variables (i.e Age, Gender, Source of Infection, Hospitalization, ICU Admittance) onto our response variable (i.e. Recovery). Given that odds ratios which are greater than 1 are associated with higher odds of outcome (recovery), it is observed that cases between the ages of 20 - 59, and cases which had been infected via healthcare or travel may have higher odds for recovery. In contrast, cases aged 70 or older, and cases were hospitalized or admitted to the ICU had lower odds for recovery due to its odds ratios being less than 1.

Based on the logistic model, most of the variables are observed to be significant. Thus, we can model the variables in this following equation:

$$\log(\frac{\widehat{p}}{1-\widehat{p}}) = 2.60 + 0.60x_{age20-29} + 0.38x_{age30-39} + 0.53x_{age40-49} + 0.33x_{age50-59} - 0.61x_{age70-79}$$

$$-1.16x_{age80-89} - 1.62x_{age90+} - 0.13x_{male} - 0.18x_{community} - 0.38x_{healthcare}$$

$$-0.43x_{inst.} - 0.16x_{outbreakasc} - 0.80x_{travel} - 1.01x_{hospitalized} - 1.26x_{ICU}$$

Discussion

Based on Figures 2-9, the trends of distributions in demographics and environment are similarly seen among recovery rates. From our logistic regression model, we are able to observe that age, gender, source of infection, hospitalization and ICU admittance have a significant influence on recovery from infection, with most cases able to recover. Among the various sources of infection, close contact led to the highest number of cases, yet also has the highest proportion of recovered cases (Figures 2 & 3). It appears that females are infected slightly more often than males, although their recovery rates are higher (Figures 6 & 7). Young adults aged 20-29 are infected most often but have the highest recovery rates compared to other age brackets. Younger demographics (i.e. persons aged 69 and younger) have higher odds for recovery as opposed to an older demographic (i.e. 70 and older). This may be expected as older demographics often have comorbidities, leading to higher likelihoods for health decline due to weakened immunity (as observed in Figure 10). It is interesting to note that the Age Group 60-69 years was considered insignificant with a p-value of 0.22. Though this age group has one of the lowest amounts of infected cases (followed only by the age group 90 and older) (Figure 4), it is surprising to note that the model did not consider this group statistically significant. As well, cases not involving hospitalization or ICU admission had drastically higher recovery rates than more severe cases requiring medical intervention. A likely factor is the median age of 70 among hospitalized cases (b). Overall, the largest observable trend within the data is worsening prognosis as age increases.

From the variable of gender, it is also observed that Client Gender of "Other", "Transgender", and "Unknown", were reported to be not statistically significant (Table 1). This may be attributed to the notion of gender being a social construct, self-reported, and the limited number of cases present for these 3 categories (as observed in Figures 6 & 7).

Based on the modeled distributions of demographics and environment, from our modeled equation, one can determine the various scenarios and probabilities of recovery from a COVID-19 infection. To model common scenarios, the following two examples are provided as case studies:

Example 1

A 22 year old male, who had traveled to Alberta and returned to Toronto was infected with COVID-19. Once in Toronto, the male was admitted to the hospital for further treatment.

$$\begin{aligned}
\log(\frac{\hat{p}}{1-\hat{p}}) &= 2.60 + 0.60(1) + 0.38(0) + 0.53(0) + 0.33(0) - 0.61(0) \\
&\quad - 1.16(0) - 1.62(0) - 0.13(1) - 0.18(0) + 0.38(0) \\
&\quad - 0.43(0) - 0.16(0) + 0.89(1) - 1.01(1) - 1.26(0) \\
&= 2.60 + 0.60 - 0.13 + 0.89 - 1.01 \\
&= 2.95 \\
\therefore \hat{p} &= 0.999
\end{aligned}$$

Applying the logistic regression equation, we are able to determine that the individual will have a 0.999 probability of recovering from the infection. This probability showcases that this case and individuals in a similar scenario, are among those who are most likely to have a "resolved" status (i.e. recovery from COVID-19).

Example 2

A 90 year old male, who was infected with COVID-19 was living in a long term care home. He was then hospitalized and admitted to the ICU.

$$\begin{aligned}
\log(\frac{\hat{p}}{1-\hat{p}}) &= 2.60 + 0.60(0) + 0.38(0) + 0.53(0) + 0.33(0) - 0.61(0) \\
&\quad - 1.16(0) - 1.62(1) - 0.13(1) - 0.18(0) + 0.38(0) \\
&\quad - 0.43(1) - 0.16(0) + 0.89(0) - 1.01(1) - 1.26(1) \\
&= 2.60 - 1.62 - 0.13 - 0.43 - 1.01 - 1.26 \\
&= -1.85 \\
\therefore \hat{p} &= 0.013
\end{aligned}$$

Applying the logistic regression model and equation, it is observed that he would have a 0.013 probability of recovery. The incredibly low probability for recovery showcases that individuals matching this scenario would be among those least likely to recover/have a "resolved" case.

From this modeling, we are able to observe the likelihoods of recovery rates. However, this is not to say that exceptions to cases may also occur and this model solely predicts probability in relation to cases reported to TPH from January to December 2020. These probabilities may help determine future policy measures to be undertaken and highlight the importance to current preventative and safety measures.

Limitations & Future Work

Given that this data had solely analyzed infection recovery in Toronto, this data must be taken with consideration for analysis in context of other cities. As well, this data set is not representative of the Canadian population at a provincial or federal level and must also be analyzed with careful consideration. Post-stratification modeling based on provincial or federal levels as future steps, may provide further insight to infection within the overall Canadian population. Here, the data set analyzed from a time period between January - December 2020. Given that this is a wide time period for cases to arise and be recorded, further analyses may specify on particular time periods (i.e. first wave, or second wave of infection) to note for particular changes. Additionally, 20,254 cases were removed to match the criteria, with 25 cases classified as "Pending" for infection. These cases though were not confirmed to have COVID-19 infection, would have attributed to sampling bias. Contact tracing as a method to identify and monitor infection outbreak is at risk for sampling bias and must be further carefully analyzed to ensure accurate representation of a population (Whitby A, 2020). Future work may also develop on analyzing for sampling biases as reported within TPH's data set, to provide further insight to the impact of COVID-19 in Toronto.

References

Arnold JB (2019). ggthemes: Extra Themes, Scales and Geoms for 'ggplot2'. R package version 4.2.0. https://CRAN.R-project.org/package=ggthemes

Centres for Disease Control and Prevention. Principles of Epidemiology. (2012, May 18). Retrieved January 06, 2021, from https://www.cdc.gov/csels/dsepd/ss1978/lesson1/section8.html

City of Toronto. (2020, December 31). COVID-19: Lockdown Guide for Toronto Residents. Retrieved January 06, 2021, from https://www.toronto.ca/home/covid-19/covid-19-reopening-recovery-rebuild/covid-19-guide-for-toronto-residents/

City of Toronto Open Data. (2020). *COVID-19 Cases in Toronto* [CSV data file]. https://open.toronto.ca/dataset/covid-19-cases-in-toronto/

Müller K (2020). here: A Simpler Way to Find Your Files. R package version 1.0.1. https://CRAN.R-project.org/package=here

Long JA (2020). *jtools: Analysis and Presentation of Social Scientific Data.* R package version 2.1.0, <URL: https://cran.r-project.org/package=jtools>.

Lüdecke D (2020). *sjlabelled: Labelled Data Utility Functions (Version 1.1.7).* doi: 10.5281/zenodo.1249215 (URL: https://doi.org/10.5281/zenodo.1249215), <URL: https://CRAN.R-project.org/package=sjlabelled>.

Lüdecke D (2018). sjmisc: Data and Variable Transformation Functions. *Journal of Open Source Software*, *3*(26), 754. doi: 10.21105/joss.00754 (URL: https://doi.org/10.21105/joss.00754).

Lüdecke D (2020). *sjPlot: Data Visualization for Statistics in Social Science.* R package version 2.8.6, <URL: https://CRAN.R-project.org/package=sjPlot>.

Public Health Agency of Canada (2020). COVID-19 in Canada. Weekly Epidemiology Update (29 July - 4 August 2020). Retrieved January 06, 2021, from https://www.canada.ca/content/dam/phac-aspc/documents/services/diseases/2019-novel-coronavirus-infection/en-surv-covid19-weekly-epi-update-20200807.pdf

Public Health Ontario (2020). COVID-19 in Ontario - A Focus on Material Deprivation: January 15, 2020 to June 3, 2020. Retrieved January 06, 2021, from https://www.publichealthontario.ca/-/media/documents/ncov/epi/2020/06/covid-19-epi-material-deprivation.pdf?la=en

R Core Team (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Solymos P and Zawadzki Z (2020). pbapply: Adding Progress Bar to '*apply' Functions. R package version 1.4-3. https://CRAN.R-project.org/package=pbapply

Szumilas M. (2010). Explaining odds ratios. Journal of the Canadian Academy of Child and Adolescent Psychiatry = Journal de l'Academie canadienne de psychiatrie de l'enfant et de l'adolescent, 19(3), 227–229.

Wickham H (2016) ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag New York, 2016.

Wickham H, et al. (2019). Welcome to the tidyverse. Journal of Open Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686

Wickham H, et al. (2020) dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr

Whitby A. (2020). Contact tracing can give a biased sample of COVID-19 cases [Blog post]. https://andrewwhitby.com/2020/11/24/contact-tracing-biased/

Xie Y (2020). knitr: A General-Purpose Package for Dynamic Report Generation in R. R package version 1.29.

Xie Y (2019) TinyTeX: A lightweight, cross-platform, and easy-to-maintain LaTeX distribution based on TeX Live. TUGboat 40 (1): 30–32. http://tug.org/TUGboat/Contents/contents40-1.html