Elderly, Male, Sporadic Population is More Likely to Get Severe COVID-19*

A estimation of the probability of developing severe COVID-19 using logistic model

Zihan Zhang

27 April 2022

Abstract

COVID-19 is a respiratory disease first discovered in 2019 and then spread across the world. In this report, a logistic model was built using the programming language R to estimate the probability of severe illness of COVID. The results show that elderly, male population is at high risk for developing severe COVID-19 symptoms. This article illustrates the high-risk population for COVID-19 and suggests a new direction for studying the causes of severe COVID-19 that requires hospitalization.

Keywords: logistic regression, severe illness of COVID-19, open data toronto

Contents

*Code and data are available at: https://github.com/zihanjasmine/Estimation-of-the-probability-of-severe-illness-of-COVID-19

1 Introduction

COVID-19 is an infection disease first discovered in Wuhan, China, in 2019, which then spread across the world. The disease primarily spreads through the direct contact of small infected respiratory droplets with open mucus membranes, such as mouth, nose, or eyes. The use of face masks, vaccinations, and social distancing are effective preventative measures for COVID-19. Evidence shows that people who are unvaccinated have a higher death rate than people who are fully vaccinated (Nick Andrews 2022). Moreover, washing your hands often and avoid touching your mouth, nose, and eyes also reduces the likelihood of being infected.

COVID-19 is highly contagious, but the disease has a low fatality rate of less than 6% in the 20 most affected countries. Most infected people will develop symptoms such as fever, cough, tiredness, etc. These mild symptoms mainly affect the respiratory system, and those affected can recover in 10 days without hospitalization. If you experience symptoms like shortness of breath, chest pain, etc., you might need to seek emergency medical attention. However, some patients experience lingering health problems even when they have recovered from the acute phase of the illness. Some common post-COVID syndromes include loss or distortion of the sense of smell and taste, fatigue, chest pain, etc.

In this paper, I will focus on factors that increase the probability of developing severe COVID-19 symptoms. Based on COVID-19 cases data obtained from Open Data Toronto, a logistic model was built to predict the probability of COVID-19 patients getting severe symptoms. The evidence suggests that aged, male, and sporadic patients are more likely to develop severe symptoms that require hospitalization. Sporadic patients means the patients' virus infection is not associated with healthcare institutions. If age and outbreak associated cases remains the same, the odds of males requiring hospitalization are 1.5 times larger than the odds for females. Since this report suggests that male and aged people have a higher risk of severe illness from COVID-19, more efforts should be made to step up advocacy and raise awareness of the importance of COVID-19 preventative measures. I urge people to get vaccinated, wear face masks, and minimize going to crowded places.

The article is structured as follows. This paper first provides an introduction to the background knowledge about COVID-19. Then data collection, data processing and data charactistic are included in data section. The model section is consists of model assumption, model construction and model validation. Model selection and interpretation are shown in the result section. Finally, discussion is comprised of three parts: summary, findings, and limitation & future work.

2 Data

2.1 Data collection

In attempts to generate an analysis of severe symptom of COVID-19, I manipulated "COVID-19 Cases in Toronto" dataset, managed by Toronto Public Health. The data is retrieved from Open data Toronto. The dataset was gathered in 2020 and it is being refreshed and overwritten every Wednesday. It contains demographic and geographic information about every confirmed COVID-19 cases in Toronto.

2.2 Data processing

All analyses were done with a statistic programming language R (R Core Team 2020). I first used R packages "opendatatoronto"(Gelfand 2020) and "tidyverse" (Wickham et al. 2019)to to import data. Then I cleaned the names of variables and removed missing values for age. The data was done cleaning. R packages "ggplot2" (Wickham 2016) and "KableExtra" (Zhu 2021) were used for data visualization.

The raw dataset contains 32000 observations in total. I randomly selected 30000 observations to complete an exploratory analysis and modeling. I also set the seed as 1 so that the same result can be obtained every time I run the code.

The data was divided into two parts- training and test. Usually, the number of observations in the training dataset is larger than the number of observations in the test dataset, so I randomly selected 70% of data

without replacement (21000 observations) as training data. The training data was used for model building and model selection. Then I fitted the optimum logistic model in the test dataset to see whether the optimum model can be validated. Table 1 and Table 2 show a summary of important variables for both the training and test dataset. The different levels of each variable account for roughly the same proportion of the two datasets, so I believe that there is no significant difference between the training and test dataset. For example, 50.1% of the sample is female in the test dataset, and 49.7% of the sample is female in the training dataset.

2.3 Data characteristics

The dataset contains demographic (age group, gender) and geographic (patient's neighborhood name and postal code), and severity (ever hospitalized, ever in ICU, outcome, etc.) information. This information is sensitive, and it is possible to identify individuals either directly or indirectly. For example, if you see a case of a transgender patient in your community in the data, combined with age and FSA (first three characters of postal code), it is highly likely that you can identify that person.

Table 3 tells that most variables are categorical variables as suggested by an extract of our dataset.

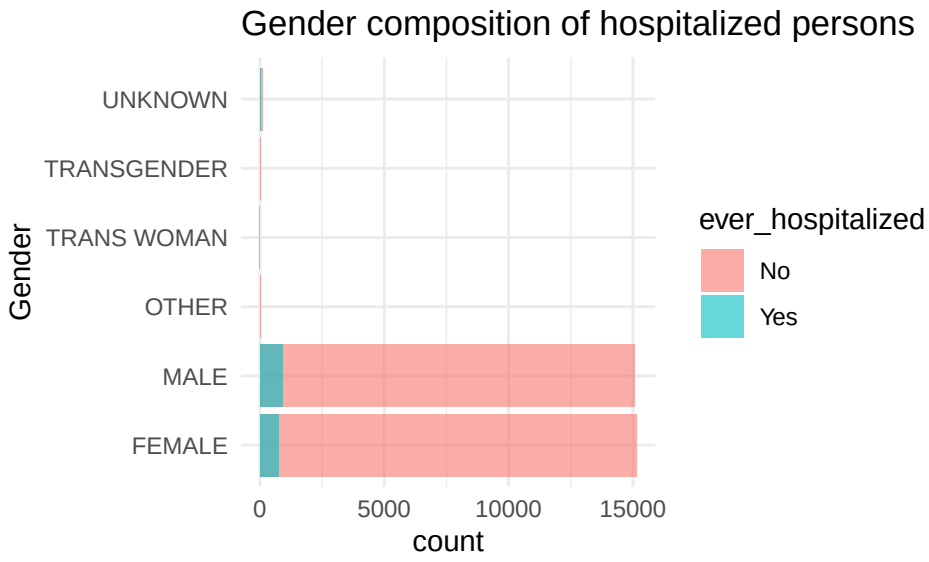

Figure 1: Gender composition of hospitalized persons

Figure 1 indicates that the number of people receiving hospitalization is about the same for women and men. After careful consideration, gender is kept as a binary variable for males and females only for modeling. There are only three observations for transgender patients, which might lead to high instability in our model.

From figure 2, we can see that young and middle-aged patients only account for a quarter of severe illnesses of COVID-19. Most patients with severe symptoms are over 60 years old.

From figure 3, roughly 40% cases that were hospitalized are fatal, which indicates there are no effective treatments for patients with severe symptoms of COVID-19.

From figure 4, the source of infection in half of the cases is unknown and cannot be speculated. Household contact and close contact with covid positive are also important causes of COVID-19 infection. 10% of COVID patients are infected in healthcare institutions. Healthcare Institutions include hospitals, retirement homes, etc. Usually, people who go to the healthcare institutions for help are mainly the aged population. These people are in poor health conditions and have a weaker immunity, making them more susceptible to viral infections.

Table 1: Summary of categorical variables for training dataset

		N	%
outbreak_associated	Outbreak Associated	3395	16.2
	Sporadic	17605	83.8
age_group	19 and younger	3097	14.7
	20 to 29 Years	4174	19.9
	30 to 39 Years	3461	16.5
	40 to 49 Years	3014	14.4
	50 to 59 Years	3041	14.5
	60 to 69 Years	1990	9.5
	70 to 79 Years	948	4.5
	80 to 89 Years	793	3.8
	90 and older	482	2.3
source_of_infection	Close Contact	1172	5.6
	Community	2479	11.8
	Household Contact	3492	16.6
	No Information	10185	48.5
	Outbreaks, Congregate Settings	174	0.8
	Outbreaks, Healthcare Institutions	2057	9.8
	Outbreaks, Other Settings	1330	6.3
	Travel	111	0.5
classification	CONFIRMED	20654	98.4
	PROBABLE	346	1.6
client_gender	FEMALE	10436	49.7
	MALE	10564	50.3
outcome	FATAL	438	2.1
	RESOLVED	20562	97.9
	No	21000	100.0
	No	21000	100.0
	No	21000	100.0
ever_hospitalized	No	19864	94.6
	Yes	1136	5.4
ever_in_icu	No	20790	99.0
	Yes	210	1.0
ever_intubated	No	20878	99.4
	Yes	122	0.6

Table 2: Summary of categorical variables for test dataset

		N	%
outbreak_associated	Outbreak Associated	1479	16.4
	Sporadic	7521	83.6
age_group	19 and younger	1341	14.9
	20 to 29 Years	1863	20.7
	30 to 39 Years	1464	16.3
	40 to 49 Years	1293	14.4
	50 to 59 Years	1284	14.3
	60 to 69 Years	800	8.9
	70 to 79 Years	422	4.7
	80 to 89 Years	315	3.5
	90 and older	218	2.4
source_of_infection	Close Contact	530	5.9
	Community	1033	11.5
	Household Contact	1465	16.3
	No Information	4385	48.7
	Outbreaks, Congregate Settings	63	0.7
	Outbreaks, Healthcare Institutions	858	9.5
	Outbreaks, Other Settings	623	6.9
	Pending	1	0.0
	Travel	42	0.5
classification	CONFIRMED	8846	98.3
	PROBABLE	154	1.7
client_gender	FEMALE	4509	50.1
	MALE	4491	49.9
outcome	FATAL	196	2.2
	RESOLVED	8804	97.8
	No	9000	100.0
	No	9000	100.0
	No	9000	100.0
ever_hospitalized	No	8565	95.2
	Yes	435	4.8
ever_in_icu	No	8909	99.0
	Yes	91	1.0
ever_intubated	No	8955	99.5
	Yes	45	0.5

Table 3: A glimpse of important varaibles

Reported date	Age group	Source of infection	Neighbourhood name	Ever hospitalized	outcome
2020-10-18	20 to 29 Years	No Information	West Hill	No	RESOLVED
2020-10-14	30 to 39 Years	No Information	Willowdale East	No	RESOLVED
2020-10-15	60 to 69 Years	No Information	North Riverdale	No	RESOLVED
2020-10-18	40 to 49 Years	No Information	Bendale	No	RESOLVED
2020-10-18	30 to 39 Years	Outbreaks, Other Settings	Malvern	No	RESOLVED
2020-10-13	30 to 39 Years	No Information	Rosedale-Moore Park	No	RESOLVED
2020-10-15	50 to 59 Years	Travel	Don Valley Village	No	RESOLVED
2020-10-13	20 to 29 Years	No Information	Elms-Old Rexdale	No	RESOLVED

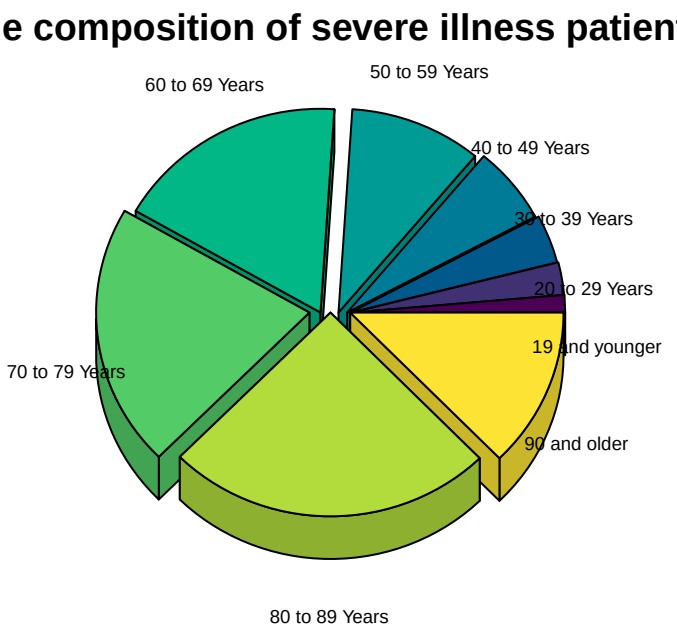

Figure 2: Age composition of severe illness patients

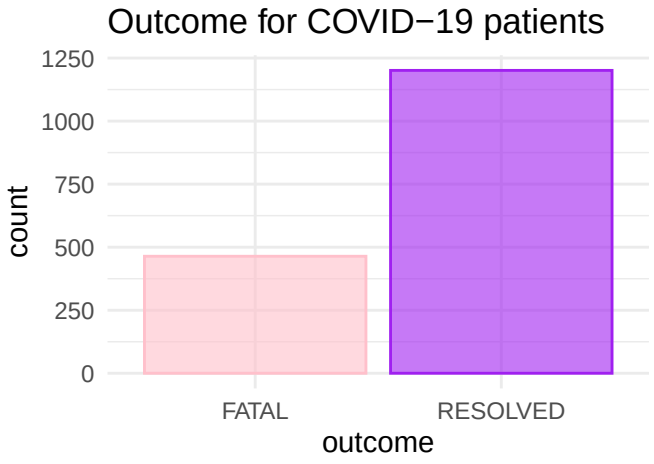

Figure 3: Outcome for COVID-19 patients

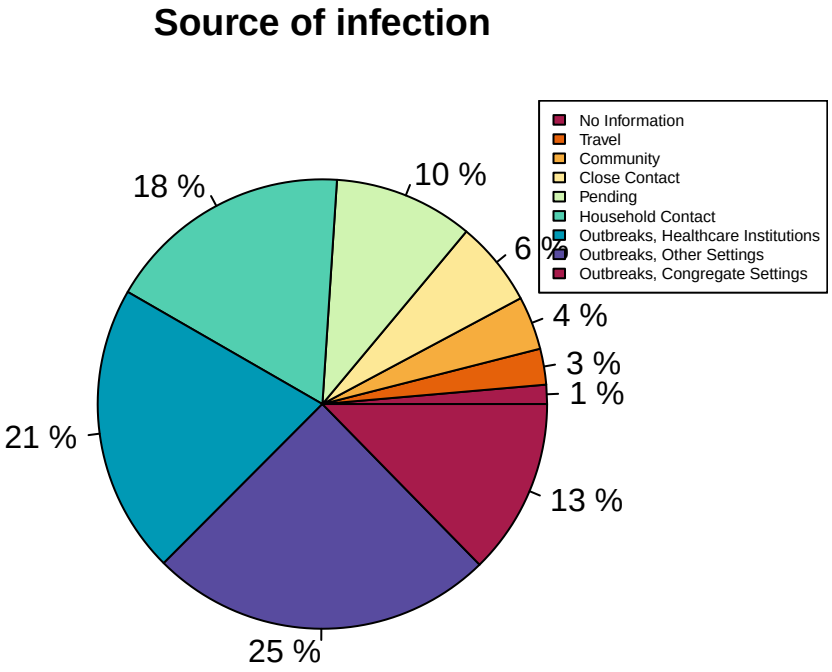

Figure 4: Source of infection

3 Model

To find out what factors can make you seriously ill after getting covid, I performed logistic regression using R (R Core Team 2020). Through the logistic regression model, I am unable to define causality, but rather explore the relationship between different variables. Since patients with mild symptoms can be recovered in a few days without hospitalization, I consider those who are hospitalized as severe illness of COVID-19. I set "ever hospitalized" as the outcome variable. The variable "ever hospitalized" is a binary variable with two levels: Yes and No, which is an appropriate structure of the outcome variable. I recoded the "ever hospitalized" and set the outcome variable to be 1 if the patient received hospitalization, and 0 if vice versa. The logistic model is constructed using the R package "stats" (R Core Team 2021).

The general form of logistic model is (see Equation (1)) :

$$log(\frac{\hat{p}}{1-\hat{p}}) = X_i\beta \tag{1}$$

- $\hat{p}$ is the probability of hospitalization given X_i
- X_i are covariates for observation i
- β is a vector of regression coefficients

I will compare 3 logistic regression models, each with different independent variables, and ultimately choose the optimum model.

3.1 model assumption

Before model construction, I will check whether the assumptions of the model are satisfied. It is important to check the assumptions of the model because if the assumptions of the model are not satisfied, it will lead to inaccurate predictions.

There are four assumptions for a logistic model.

- Assumption of appropriate outcome type

Logistic regression assumes the outcome variable to be binary with two levels. The response variable should follow binomial or bernoulli distribution.

- Assumption of independence of observations

Logistic regression assumes that the observations must be independent of each other. Each observation should represent one patient, so no repeated data is measured. In this paper, the residual plot will be used to check the independence assumption. If the residuals in the plot are randomly scattered around 0, the independence assumption is satisfied.

- Assumption of absence of multicollinearity

Multicollinearity refers to a strong linear relationship between predictors, which might result in non-significant predictors, weakened statistical power of our logistic model, etc. Multicollinearity is accessed through VIF value, a predictor with a VIF value that exceeds 5 indicates strong multicollinearity, which suggests the predictor might need to be removed.

- Assumption of large sample size

Logistic model assumes there are sufficient observations for each variable to avoid overfitting.

3.2 Model construction

3.2.1 Model 1

First, I would like to examine the effect of gender and age on the probability of hospitalization.

The first model is (see Equation (2)):

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Client\ gender + \beta_2 Age\ group \tag{2}$$

The probability of hospitalization is (see Equation (3)):

$$\hat{p} = \frac{e^{\hat{\beta_0}+\hat{\beta_1} Client\ gender+\hat{\beta_2} Age\ group}}{e^{\beta_0+\hat{\beta_1} Client\ gender+\hat{\beta_2} Age\ group}+1} \tag{3}$$

3.2.2 Model 2

On the basis of model 1, I want to study the effect of the variable "outbreak associated". This variable tells us whether the case is associated with healthcare institutions or not. If the infection is related to healthcare institutions, it is outbreak associated and if the infection is not hospital related, it is sporadic.

The second model is (see Equation (4)) :

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 Outbreak\ associated + \beta_2 Client\ gender + \beta_3 Age\ group \tag{4}$$

The probability of hospitalization is (see Equation (5)):

$$\hat{p} = \frac{e^{\hat{\beta_0}+\hat{\beta_1} Outbreak\ associated+\hat{\beta_2} Client\ gender+\hat{\beta_3} Age\ group}}{e^{\hat{\beta_0}+\hat{\beta_1} Outbreak\ associated+\hat{\beta_2} Client\ gender+\hat{\beta_3} Age\ group}+1} \tag{5}$$

3.2.3 Model 3

Based on model 2, I am also interested in whether the interaction term of outbreak associated and gender affect the probability of developing severe COVID-19 that requires hospitalization.

The third model is (see Equation (6)) :

$$log(\frac{p}{1-p}) = \beta_0+\beta_1 Outbreak\ associated+\beta_2 Client\ gender+\beta_3 Age\ group+\beta_4 Client\ gender : Outbreak\ associated \tag{6}$$

The probability of hospitalization is :

$$\hat{p} = \frac{e^{\hat{\beta_0}+\hat{\beta_1} Outbreak\ associated+\hat{\beta_2} Client\ gender+\hat{\beta_3} Age\ group+\hat{\beta_4} Client\ gender:Outbreak\ associated}}{e^{\hat{\beta_0}+\hat{\beta_1} Outbreak\ associated+\hat{\beta_2} Client\ gender+\hat{\beta_3} Age\ group+\hat{\beta_4} Client\ gender:Outbreak\ associated}+1} \tag{7}$$

After model construction, I will select the best model in the next step. The preferred model will be tested with test data to see if it can be validated.

3.3 Model validation

Data validation is an important process to be carried out after model selection, it is a way to evaluate the effectiveness of our optimum model. The COVID-19 cases in Toronto dataset was separated into training and test dataset in the beginning, the training dataset is used for model building and model selection, and the test dataset is used for model validation.

The test dataset will be used to fit the optimum model. If the following conditions are satisfied, it is safe to say the optimum model is validated. First, the difference of model coefficients of the test dataset is within the standard error of corresponding coefficients of the training dataset. Second, the predictors of the optimum model appear significant in both training and test datasets. Thirdly, when fitting the optimum model using test dataset, the test dataset must not have more serious assumption violations.

4 Result

After plugging in the data, we obtained the regression coefficients for the three models.

The first model is :

$$log(\frac{\hat{p}}{1-\hat{p}}) = -5.5164 + 0.434Client\ gender_{Male} + 0.5701Age\ group_{20-29\ years} + 0.8237Age\ group_{30-39\ years} + 1.5476Age\ group_{40-49\ years} + 2.0775Age\ group_{50-59\ years} + 3.0685Age\ group_{60-69\ years} + 4.2582Age\ group_{70-79\ years} + 4.6742Age\ group_{80-89\ years} + 4.4333Age\ group_{90+} \quad (8)$$

Table 4 shows a summary of model 1 (see Equation (8)). Most of the predictors are statistically significant at a 0.05 significant level.

Table 4: A summary of model 1

term	estimate	std.error	statistic	p.value
(Intercept)	-5.5164	0.2542	-21.6975	0.0000
client_genderMALE	0.4340	0.0677	6.4076	0.0000
age_group20 to 29 Years	0.5701	0.2989	1.9074	0.0565
age_group30 to 39 Years	0.8237	0.2968	2.7758	0.0055
age_group40 to 49 Years	1.5476	0.2782	5.5623	0.0000
age_group50 to 59 Years	2.0775	0.2676	7.7638	0.0000
age_group60 to 69 Years	3.0685	0.2615	11.7335	0.0000
age_group70 to 79 Years	4.2582	0.2612	16.3004	0.0000
age_group80 to 89 Years	4.6742	0.2620	17.8426	0.0000
age_group90 and older	4.4333	0.2712	16.3471	0.0000

The second model is :

$$log(\frac{\hat{p}}{1-\hat{p}}) = -5.7373 + 0.279Outbreak\ associated_{Sporadic} + 0.4189Client\ gender_{Male} + 0.537Age\ group_{20-29\ years} + 0.7959Age\ group_{30-39\ years} + 1.5287Age\ group_{40-49\ years} + 2.0612Age\ group_{50-59\ years} + 3.0585Age\ group_{60-69\ years} + 4.2752Age\ group_{70-79\ years} + 4.7766Age\ group_{80-89\ years} + 4.61334Age\ group_{90+} \quad (9)$$

Table 5 shows a summary of model 2 (see Equation (9)).

Table 5: A summary of model 2

term	estimate	std.error	statistic	p.value
(Intercept)	-5.7373	0.2640	-21.7357	0.0000
outbreak_associatedSporadic	0.2790	0.0874	3.1925	0.0014
client_genderMALE	0.4189	0.0679	6.1683	0.0000
age_group20 to 29 Years	0.5370	0.2990	1.7958	0.0725
age_group30 to 39 Years	0.7959	0.2969	2.6810	0.0073
age_group40 to 49 Years	1.5287	0.2783	5.4934	0.0000
age_group50 to 59 Years	2.0612	0.2676	7.7017	0.0000

Table 7: Comparison of model 1 and model 2 using the likelihood ratio test

#Df	LogLik	Df	Chisq	Pr(>Chisq)
10	-3332.750	NA	NA	NA
11	-3327.525	1	10.45149	0.0012255

term	estimate	std.error	statistic	p.value
age_group60 to 69 Years	3.0585	0.2615	11.6945	0.0000
age_group70 to 79 Years	4.2753	0.2613	16.3600	0.0000
age_group80 to 89 Years	4.7766	0.2641	18.0885	0.0000
age_group90 and older	4.6133	0.2773	16.6380	0.0000

The third model is :

$$log(\frac{\hat{p}}{1-\hat{p}}) = -5.8792 + 0.4807Outbreak\ associated_{Sporadic} + 0.707Client\ gender_{Male} + 0.5474Age\ group_{20-29\ years}$$
$$+0.8055Age\ group_{30-39\ years} + 1.5366Age\ group_{40-49\ years} + 2.0678Age\ group_{50-59\ years} + 3.0633Age\ group_{60-69\ years}$$
$$+4.2713Age\ group_{70-79\ years} + 4.7898Age\ group_{80-89\ years} + 4.6534Age\ group_{90+}$$
$$-0.4029Outbreak\ associated_{Sporadic} : Client\ gender_{Male} \quad (10)$$

Table 6 shows a summary of model 3 (see Equation (10)).

Table 6: A summary of model 3

term	estimate	std.error	statistic	p.value
(Intercept)	-5.8792	0.2697	-21.7974	0.0000
outbreak_associatedSporadic	0.4807	0.1158	4.1497	0.0000
client_genderMALE	0.7070	0.1263	5.5986	0.0000
age_group20 to 29 Years	0.5474	0.2990	1.8304	0.0672
age_group30 to 39 Years	0.8055	0.2969	2.7130	0.0067
age_group40 to 49 Years	1.5366	0.2783	5.5212	0.0000
age_group50 to 59 Years	2.0678	0.2676	7.7259	0.0000
age_group60 to 69 Years	3.0633	0.2615	11.7129	0.0000
age_group70 to 79 Years	4.2713	0.2613	16.3471	0.0000
age_group80 to 89 Years	4.7898	0.2640	18.1418	0.0000
age_group90 and older	4.6534	0.2777	16.7567	0.0000
outbreak_associatedSporadic:client_genderMALE	-0.4029	0.1489	-2.7054	0.0068

In order to figure out the optimum model, the likelihood ratio test is performed. The likelihood ratio test helps us access the goodness of fits by maximizing the entire parameter space, which is difficult and time-consuming to calculate by hand. I used the R function "lrtest( )" from R package "lmtest" (Zeileis and Hothorn 2002) to run the test. The two models to be compared must be nested.

In this case, model 1 is a subset of model 2, so the two models are nested. Table 7 shows a likelihood ratio test to compare model 1 and model 2. From table 7, the p-value is 0.0012, which is smaller than 0.05, so I have strong evidence against the null hypothesis that the reduced model (mode 1) better explains the COVID-19 cases data. In conclusion,compared to model 1, model 2 (full model) better explains the data.

Table 8: Comparison of model 2 and model 3 using the likelihood ratio test

#Df	LogLik	Df	Chisq	Pr(>Chisq)
12	-3323.869	NA	NA	NA
11	-3327.525	-1	7.310733	0.0068544

Model 2 is a subset of model 3, so I performed a likelihood ratio test to compare mode 2 and model 3. (Table 8) From table 8, the p-value is 0.068, which is greater than 0.05, so I failed to reject the null hypothesis (H_0) that model 2 better explains the COVID-19 cases data. Compared to model 3, model 2 better explains the data. In conclusion, model 2 (see Equation (9)) is the optimum model.

4.1 Model interpretation

The optimum model is :

$$log(\frac{\hat{p}}{1-\hat{p}}) = -5.7373 + 0.279Outbreak\ associated_{Sporadic} + 0.4189Client\ gender_{Male} + 0.537Age\ group_{20-29\ years} + 0.7959Age\ group_{30-39\ years} + 1.5287Age\ group_{40-49\ years} + 2.0612Age\ group_{50-59\ years} + 3.0585Age\ group_{60-69\ years} + 4.2752Age\ group_{70-79\ years} + 4.7766Age\ group_{80-89\ years} + 4.61334Age\ group_{90+} \quad (11)$$

- Outbreak associated

When all other predictors except outbreak associated remain the same, the odds of hospitalization for an outbreak-associated patient is 1.3218 ($e^{0.279}$) times larger than the odds of hospitalization for a sporadic patient.

- Gender

When all other predictors except gender remain the same, the odds of hospitalization for a male is 1.5203 times larger than the odds for a female being admitted to a hospital.

- Age group

When all other predictors except age group remain the same, the odds of hospitalization for a patient aged 20-29 is 1.711 times larger than the odds for a patient aged below 20. The odds of hospitalization for a patient aged 30-39 is 2.2164 ($e^{0.7959}$) times larger than the odds for a patient aged below 20. The odds of hospitalization for a patient aged 40-49 is 4.6122 ($e^{1.5287}$) times larger than the odds for a patient aged below 20. The odds of hospitalization for a patient aged 50-59 is 7.8554 ($e^{2.0612}$) times larger than the odds for a patient aged below 20. The odds of hospitalization for a patient aged 60-69 is 21.2656 ($e^{3.0585}$) times larger than the odds for a patient aged below 20. The odds of hospitalization for a patient aged 70-79 is 71.9017 ($e^{4.2753}$) times larger than the odds for a patient aged below 20. The odds of hospitalization for a patient aged 80-89 is 118.7 ($e^{4.7766}$) times larger than the odds for a patient aged below 20. The odds of hospitalization for a patient aged over 90 is 104.9412 ($e^{4.6534}$) times larger than the odds for a patient aged below 20.

- Interaction of gender and outbreak

When all other predictors remain the same, the odds of hospitalization for an outbreak-associated male patient is 0.6684 times larger than the odds of hospitalization for a sporadic female patient.

4.2 Model assumption

For the optimum model (see Equation (11)), the outcome variable "ever hospitalized" has two levels, Yes and No, which means the assumption of appropriate outcome type is met. From figure 5, the residuals are roughly randomly centralized around 0, so the independence assumption holds. It is difficult and time-consuming to

Table 9: VIF values for different predictors

	GVIF	Df	GVIF^(1/(2*Df))
outbreak_associated	1.426547	1	1.194381
client_gender	1.039718	1	1.019665
age_group	1.442878	8	1.023180

calculate VIF value by hand, so R package "car" (Fox and Weisberg 2019) was used to calculate VIF values for model 2. From table 9, since no VIF values exceed 5, it is safe to say the absence of multicollinearity assumption is satisfied. The dataset contains over 30000 observations, which suggests the assumption of large sample size holds.

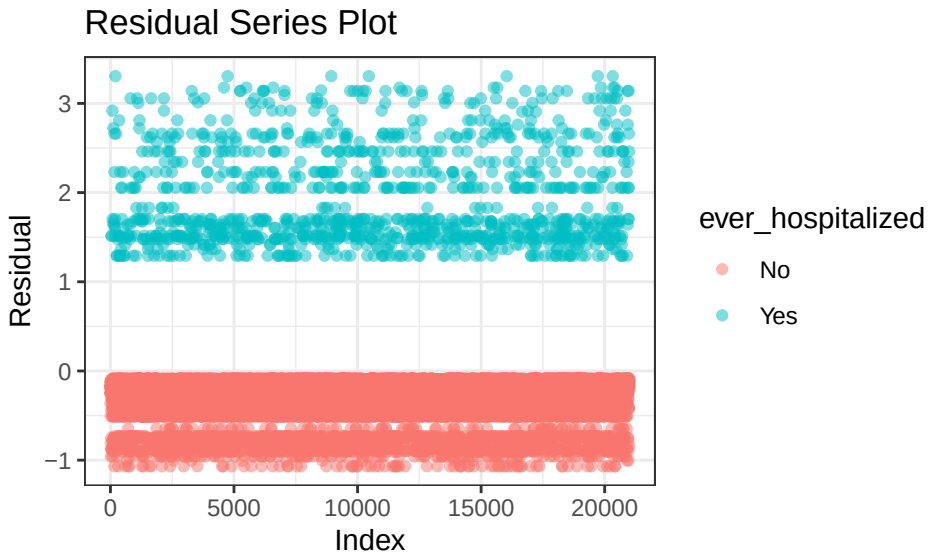

Figure 5: Residual Series Plot

4.3 Model validation

I fitted the preferred model using the test dataset, and table 10 shows a summary of the preferred in the test dataset. Comparing table 5 and table 10, the same predictors are statistically significant in both training and test datasets. What's more, the response variable is a binary variable, which indicates the assumption of appropriate outcome type holds. Figure 6 shows the assumption of independence assumption holds. From table 11, the VIF values for predictors are smaller than 5, indicating the assumption of absence of multicollinearity is satisfied. Overall, compared to model assumptions using the training datset, the model assumptions are not more seriously violated in the test dataset.

As for the model coefficients, the difference of coefficients of $client\ gender_{male}, outbreak\ associated_{Sporadic}$ are bigger than the standard error of the corresponding coefficient using the training dataset. For example, the difference of coefficient for age group between 20 and 29 is 1.7706, which is larger than 0.2969 (the standard error for age group between 20 and 29 in the training dataset). The difference of coefficients of other terms is within the standard error of corresponding coefficients of the training dataset. Overall, most conditions are satisfied, so the optimum model is validated.

Table 11: VIF values for different predictors using test dataset

	GVIF	Df	GVIF^(1/(2*Df))
outbreak_associated	1.471568	1	1.213082
client_gender	1.067905	1	1.033395
age_group	1.517738	8	1.026419

Table 10: A summary of model 2 using test dataset

term	estimate	std.error	statistic	p.value
(Intercept)	-5.9561	0.4033	-14.7684	0.0000
outbreak_associatedSporadic	0.4365	0.1414	3.0866	0.0020
client_genderMALE	0.5834	0.1122	5.2012	0.0000
age_group20 to 29 Years	-1.2336	0.6913	-1.7845	0.0743
age_group30 to 39 Years	0.8155	0.4474	1.8226	0.0684
age_group40 to 49 Years	1.5099	0.4219	3.5785	0.0003
age_group50 to 59 Years	1.6662	0.4162	4.0029	0.0001
age_group60 to 69 Years	2.9996	0.3978	7.5405	0.0000
age_group70 to 79 Years	3.9176	0.3988	9.8241	0.0000
age_group80 to 89 Years	5.0033	0.4028	12.4221	0.0000
age_group90 and older	4.8578	0.4184	11.6097	0.0000

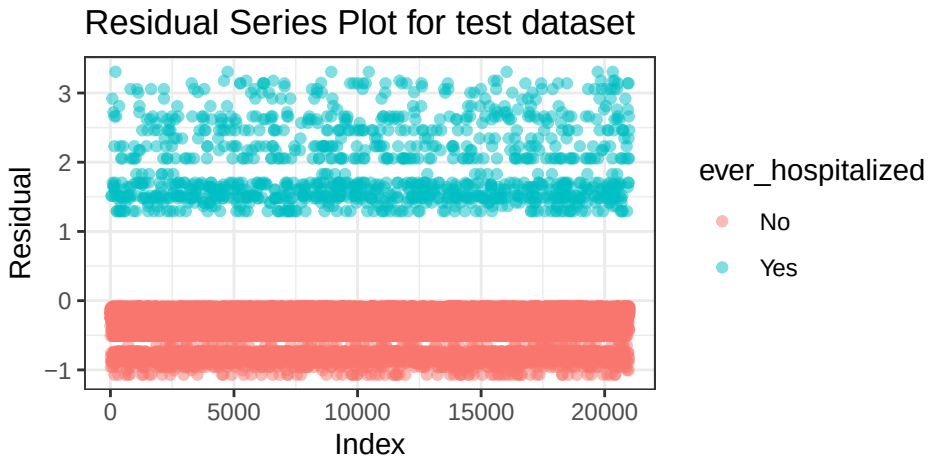

Figure 6: Residual Series Plot for test dataset

5 Discussion

5.1 Summary

COVID-19 is not only a global pandemic but also causes economic devastation. Many people may be facing an increased level of anxiety and stress due to the outbreak of COVDI-19. The speed of the spread of COVID-19 is much faster than the pace that we learn about the virus. Almost as soon as the virus started to break out, scientists around the world began researching diagnostic methods, vaccines, and treatments. Now there are five major variants of COVID-19: Alpha, Beta, Gamma, Delta, and Omicron ("COVID-19 Variants of Concern" 2022).

In this report, I used the COVID-19 Cases in Toronto dataset from Open Data Toronto to study various factors that increase the probability of getting severe illness due to COVID-19. The target population is all confirmed COVID-19 positive cases reported to Toronto Public Health since 2020. I focused on the effect of age group, gender, and whether the infection occurs in healthcare institutions on the probability of developing severe symptoms due to the virus. In this paper, logistic regression is built to estimate the likelihood of hospitalization after getting COVID using the training dataset. The model is then validated using the test dataset.

5.2 Findings

In this report, I used a logistic regression built to predict the likelihood of hospitalization after getting COVID. The assumptions of appropriate outcome type, independence assumption, and assumption of absence of multicollinearity are checked before fitting the logistic model. No violations of model assumptions are found so it is safe to say our optimum model is reliable. We found out that there is a linear relationship between age, gender, whether the infection occurs in healthcare institutions and the probability of hospitalization. Aged, male population is more likely to get severe symptoms of COVID-19. Aged population over 80 years old are roughly 100 times more likely to develop symptoms severe enough to require hospitalization than teenagers. This data is very alarming, and one possible guess is that older people have weaker immune systems and are more vulnerable to viral infections. In addition, older people may have other geriatric conditions such as high blood pressure, heart disease, diabetes, etc. People infected in healthcare institutions like hospitals, and retirement homes are 1.3 times more likely to develop symptoms severe enough to require hospitalization than people getting COVID-19 from other places. A possible speculation is that retirement homes are filled with elderly people, and they are vulnerable to virus infection. People who go to medical centers to seek help also are not in good health conditions and their immune system is depressed. What's more, males are roughly 1.5 times more likely to develop symptoms severe enough to require hospitalization than females. I removed transgender, trans women, and unknown responses for gender, because the observations for transgender, trans women are too small, which might lead to extremely high variability in the model.

Based on the above information obtained from the preferred logistic model (see Equation (11)), we can suggest some useful measurements to reduce infections in high-risk groups. First, if you feel sick and go to a medical facility for help, be sure to watch out for nosocomial infections. Never take your mask off in the hospital, and try to wear disposable gloves if conditioning allows. When you are in poor health conditions, you are more vulnerable to various virus infections. Secondly, the elderly should try to avoid going to crowded places. For example, they can ask their children or neighbors to help them order vegetables and other foods online so that they don't have to go to places like supermarkets where there is no air circulation and a lot of people.

5.3 Limitation & Future Work

First, the raw data only collects confirmed cases reported to the Toronto Public Health, which accounts for a fraction of those infected with COVID-19. Data from patients who tested positive using the self-test kit are not collected. In addition to performing extensive screening, data from asymptomatic patients are difficult to collect. A large number of missing data from asymptomatic and self-testing patients can lead to a higher proportion of infected individuals that that aren't admitted to hospitals in the dataset. Asymptomatic and self-testing patients do not need hospitalization. If we consider asymptomatic patients, the percentage of patients who develop symptoms severe enough to require hospitalization will decline sharply. The proportion of patients who are hospitalized versus those who are not is important for fitting the logistic model. If the outcome is rare, even though the sample is large, it is inappropriate to fit a logit model.

Secondly, the dataset contains some ethical issues. The dataset contains sensitive information such as home address. Additionally, the data includes transgender as one option, however, only the option for trans women is included seems problematic. There are only three people are found to be transgender, and only one individual is a trans woman. Combining the neighborhood name and age group of the trans women, it is highly likely to identify the individual. The re-identification risk is a serious violation of people's privacy rights and has many potential risks. In addition to this, data collection is mandatory due to public health

concerns, which is contrary to the principle of voluntariness. The data also contains medical information such as whether you are incubated, whether or not you are admitted to ICU, etc. This information should be kept between doctors and patients.

In this report, I found out that severe symptoms of COVID-19 are related to age, gender, etc. Even though no causality is defined, the paper can warn people of the potential risks. Hopefully, this article will make readers aware of the probability of developing serious symptoms if they are infected with the virus. In the future, more deep research will be done to prevent people from developing symptoms severe enough to require hospitalization. From the results shown above, men are more likely to develop symptoms severe enough to require hospitalization, but the underlying cause is not clear. Is it because of the difference in physiological constitution or the difference in lifestyle habits? This is a new topic worth studying.

6 Appendix

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 enable analysis of the probability of covid-19 patients requiring hospitalization. There was no specific gap that needed 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)?*
   - The dataset is managed by Toronto Public Health.The data are extracted from the provincial Case & Contact Management System.
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.*
   - The data is funded by City of Toronto government in Toronto, Ontario, Canada.
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.*
   - This data contains all confirmed and probable COVID-19 cases reported to Toronto Public Health. Each observation of the dataset is an individual.
2. *How many instances are there in total (of each type, if appropriate)?*
   - There are 31997 observations in the clean data.
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 contains all confirmed and probable COVID-19 cases in Toronto. It can be considered as s subset of COVID-19 cases in Ontario.
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.*
   - Each instance consist of demographic (age group, gender) and geographic (patient's neighbourhood name and postal code), and severity (ever hospitalized, ever in icu, outcome etc.) information.
5. *Is there a label or target associated with each instance? If so, please provide a description.*
   - Yes, every observation has a unique assigned ID.
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.*
   - Age group, neighborhood name and fsa is not available in all cases. If the patient is not a Canadian citizen, but a visitor, information such as neighborhood name and fsa may be 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, through the unique assigned ID.
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. In this paper, 70% of the dataset goes into the training set and 20% of the dataset goes into the testing set.
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 dataset rely on open data Toronto website. The data are subject to change as public health investigations into reported cases are ongoing. The consumer should follow Open Data License. The dataset is available from https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

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.*
    - Yes. The datast contains information whether you are incubated, whether you are admitted to ICU, etc.
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.*
    - Yes. The dataset identifies sub-populations by age, gender, and neighborhood name.
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.*
    - Yes. It is likely to identify someone combined with age group, gender, neighborhood name, and other datasets.
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.*
    - The dataset contains sensitive information, such as neighborhood name, and whether they are incubated or admitted to ICU.
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 ware gathered from the provincial Case & Contact Management System.
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 downloaded from Open Data Toronto.
3. *If the dataset is a sample from a larger set, what was the sampling strategy (for example, deterministic, probabilistic with specific sampling probabilities)?*
    - The dataset is not a sample.
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)?*
    - The dataset was collected by Toronto Public Health.
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 data was collected from January 2020 to now. It is updated on a weekly basis.

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 data was collected by Toronto Public Health.
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.*
   - No.
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).*
    - No.
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.*
   - Yes, cleaning of the data was done. Missing values of age is removed.
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.*
   - Yes. It is available through 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.*
   - No.
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?*
   - The data can be used to analyze the Toronto epidemic.
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 is available from https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.
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 distributed through API.
3. *When will the dataset be distributed?*
   - The dataset was first published in January 2020.
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 under Open Data 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.*
   - None that are known.
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.*
   - None that are known.
7. *Any other comments?*
   - No.

Maintenance

1. *Who will be supporting/hosting/maintaining the dataset?*
   - The dataset is supported by Toronto Public Health.
2. *How can the owner/curator/manager of the dataset be contacted (for example, email address)?*
   - The publisher's email is edau@toronto.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)?*
   - The dataset will be refreshed and overwritten every Wednesday. The dataset is available from https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.
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.*
   - No.
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.The data is just updated.
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.

- No.

8. *Any other comments?*
   - No.

Reference

"COVID-19 Variants of Concern." 2022. Public Health Ontario.

Fox, John, and Sanford Weisberg. 2019. *An R Companion to Applied Regression*. Third. Thousand Oaks CA: Sage. https://socialsciences.mcmaster.ca/jfox/Books/Companion/.

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

Nick Andrews, Elise Tessier. 2022. "Duration of Protection Against Mild and Severe Disease by Covid-19 Vaccines."

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

———. 2021. *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.

Zeileis, Achim, and Torsten Hothorn. 2002. "Diagnostic Checking in Regression Relationships." *R News* 2 (3): 7–10. https://CRAN.R-project.org/doc/Rnews/.

Zhu, Hao. 2021. *KableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.