Investigation of Factors that Influence the Fatal Rate of COVID-19 and the Regional Medical Resources Equity in Canada

Ziyue Xu

2020/12/22

The code used for data cleaning and data analysis in this paper is available at:

https://github.com/xuziyue/covid-19-analysis.git

Abstract

Coronavirus disease 2019 (COVID-19) is highly contagious and variable, and it has claimed millions of lives globally. In this paper, first, a logistic regression model is fitted to investigate the factors that influence the fatal rate of COVID-19, and the result indicates that age, gender, hospitalization status, and residence region factors have a significant relationship with the COVID-19 fatal rate. Second, the propensity score matching method is utilized to study if there is a causal link between the living regions of COVID-19 infected individuals and the outcomes, and the result of this causal inference indicates that patients living or getting treated in the developed regions are less likely to have fatal outcomes. Thus, regional medical resources inequality exists in Canada. These conclusions are valuable and instrumental, as they can remind the groups of people with high fatal risks to prevent infection and suggest the Canadian government allocating more medical resources to the less developed regions.

Keywords: Fatal rate, Medical resources equity, Logistic regression, Propensity Score, Causal Inference

Introduction

The global pandemic caused by the Coronavirus disease 2019 (COVID-19) is extremely devastating and causing substantial life loss. According to the cases reported to the World Health Organization (WHO), more than 1.68 million people have lost their lives due to this pandemic as of December 2020, and the number of fatalities continues to increase (World Health Organization, 2020). While the highly contagious COVID-19 virus could infect anyone, infected people with different attributes tend to have different outcomes (Public Health Agency of Canada, 2020). Some of them heal quickly, but some of them, unfortunately, received fatal consequences. The major suspected risk factors include age, gender, whether asymptomatic, etc. Therefore, studying the factors related to the COVID-19 fatal rate by statistical analysis is critical and valuable to reveal what groups of people are inclined to have a high fatal rate. More importantly, the result can remind high-risk groups to take extra precautions to prevent getting infected. Based on a dataset of COVID-19 cases in Canada, the logistic regression model will be utilized to analyze how and to what extent various factors influence the patients' fatal rate.

During this pandemic, another question arises and worth investigation. Do the patients with similar conditions but in different areas in Canada have the same fatal rate? The medical resources, such as hospitals, doctors, and equipment, significantly influence the mortality rate (Chai, Zhang, & Chang, 2020). It is common that medical resources are unequally distributed in different regions in many countries, and the more developed regions might have more medical resources (Chai, Zhang, & Chang, 2020). By conducting a causal inference on the dataset of COVID-19 cases in Canada, the existence of a causal link between where a patient gets treated and whether or not a patient gets a fatal outcome will be examined. Consequently, the medical resources distribution equity in different regions in Canada can be reflected.

Logistic regression is commonly used to model the relationship between predictor variables and a binary response variable (Wu & Thompson, 2020). In this report, logistic regression is appropriate to study the influence of factors including age group, gender, whether asymptomatic, hospitalization status, and region on the patients' fatal rate, since the response variable **Fatal outcome** is binary-valued (either 0 or 1).

In order to investigate whether there is a causal link between where a patient gets treated and whether or not a patient gets a fatal outcome, the propensity score matching method can be used (Rosenbaum & Rubin, 1982). The reason for using propensity score matchig is there are many variables for each case in the dataset, and an advantage of propensity score matching is that it is capable to take many variables into account at once (Alexander, 2020). In this report, a logistic model is first fitted to compute each case's propensity score of living (get hospitalized) in the developed regions. Then a new data containing matched pairs (one in treatment group, one in comparison group) is created. Last, use the new data to fit a logistic regression model, and the differences can be attributed to the treatment.

A cleaned dataset consists of observed COVID-19 cases in Canada is described in the Data section below. In the Model and Methodology section, the predictor variables, response variable, a logistic regression model, and the models used for propensity score matching are shown mathematically and explained. Results of the logistic regression and the propensity score analysis are presented in the Results section. Finally, conclusions drawn from the results, weaknesses, and next steps are discussed and analyzed in the Discussion section.

Data

Data source

The dataset used in this paper consists of detailed information about the COVID-19 cases in Canada by December 10, 2020. It is a published dataset called "Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada" on the Statistics Canada website (Statistics Canada, 2020).

Data collection and processing approach

The confirmed COVID-19 cases and detailed information are collected by the provincial or territorial public health authority in Canada. There are various ways for the local public health authorities in Canada to collect case data, including hospital reports, test locations reports, telephone and email contacts from residents. Then each government of provinces and territories in Canada use a common Case Report Form (CRF) to record detailed information of confirmed cases, and reported it to the Public Health Agency of Canada (PHAC). Statistics Canada collaborates with PHAC to create this update-to-date dataset and provides it to researchers. The variables included in this dataset are the information considered to be both important and of high quality. Besides, PHAC computed and provided some "derived variables" based on the information in the case report forms to Statistics Canada. When the data was processed by Statistics Canada, the confidentiality of patients is protected by following approaches:

1. Some categories of the original responses have been grouped together, such as provinces and territories, age groups.
2. Some categories of the original responses have been reclassified, such as occupations.
3. Some derived variables have been created, such as dates.
4. The same case will have a different case identifier number every time this dataset is updated and released.

The population, the frame, and the sample

- The Target Population: All confirmed COVID-19 cases in Canada
- Sampling Frame: The information of all confirmed COVID-19 cases reported to the Public Health Agency of Canada (PHAC) by all of the provincial or territorial public health authorities in Canada using the common Case Report Form (CRF).

- Sample: All confirmed COVID-19 cases reported to the Public Health Agency of Canada (PHAC) by all of the provincial or territorial public health authorities in Canada, and included in this dataset by Statistics Canada.

Find respondents and handle non-response

The respondents are the people with confirmed COVID-19 infection. Once a case is confirmed, the hospital, test location, or other agency will ask the patient questions, collect detailed information, and report it to the local public health authority. All confirmed cases should be recorded and reported to the Public Health Agency of Canada (PHAC). Under the circumstances that there are some questions the respondents prefer not to answer or don't know, the response "Not Stated/Unknown" will be recorded.

Data strengths and weaknesses

- The strengths of this dataset are:
  1. It respects the confidentiality of the individuals whose information on COVID-19 is reported.
  2. This dataset is updated regularly, and each version released will provide up-to-date detailed information reported by the provinces and territories. Therefore, this dataset is always fresh and time-sensitive. The researchers can conduct real-time analysis, and the analysis conclusion derived by using this dataset can reflect Canada's current condition.
  3. Almost all variables has a value indicates that the respondents prefer not to answer this question or don't know. This is achieved by adding a "Valid Skip/Unknown/Not Stated" option to each question. Therefore, the respondents can have a response to each question in any case, and it makes data cleaning more convenient.
- The weaknesses of this dataset are:
  1. This dataset may not match the total cases reported by the provincials and territorials, which are updated on a daily basis. The inconsistency is due to the delays in reporting or variability in the reporting time of different provincials and territorials.
  2. The comparisons between provinces and territories using the region information in this data may be biased by the differences in testing criteria between provinces and territories.
  3. This dataset has some data quality concerns indicated in its User Guide. Some parts of the common Case Report Form (CRF) were filled in inconsistently, and there are many missing values.

Data visualization and summary

Use the visdat function in the visdat package (Tierney & Hughes, 2019) to visualize the raw dataset "Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada" retrieved from the Statistics Canada. The raw dataset contains 360236 observations. The information of each confirmed case comprises case identifier number, this individual's gender, age group, occupation, living region, time of the episode, whether or not this individual is asymptomatic, what kinds of symptom is shown, hospitalization status, whether or not recovered, time of the recovery, whether or not died while infected by COVID-19, and the location where exposure occurred. There are 30 variables in total to record these detailed information.

Also, it is noticeable that all the variables are numeric value variables. The reason is that the discrete integers are used to represent the answer categories in this dataset. For example, the Region variable represents the Province or Territory where the case resides, and the value 1 is for Atlantic; 2 for Quebec; 3 for Ontario and Nunavut; 4 for Prairies and the Northwest Territories; 5 for British Columbia and Yukon. This answer categories encoding reduced the dataset size, and facilitate the data cleaning process. Since the raw data contains a large number of variables and uses number encoding to represent the categories of each variable, it is inconvenient to do analysis directly on it. The data cleaning is needed.

The raw dataset is cleaned by selecting important variables for the logistic regression model and causal inference in the Model section, including regrouped age group, gender, asymptomatic, hospital status, region,

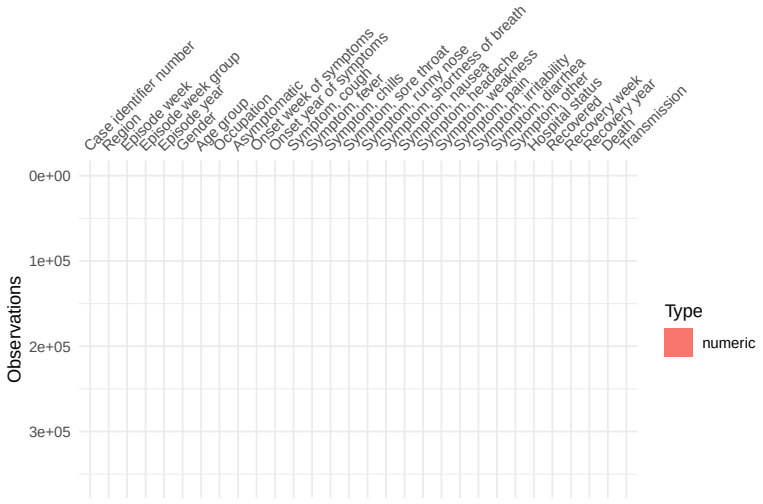

Figure 1: Visualization of the raw data

and fatal outcome. The cases with "Not Stated/Unknown" responses are also removed, since they are not very useful for analysis and causal inference.

Use the visdat function in the visdat package (Tierney & Hughes, 2019) to visualize the cleaned dataset. The cleaned dataset contains 62648 observations and 7 variables. The variables **Age group**, **Gender**, **Asymptomatic**, **Hospital status**, **Region**, **Fatal outcome**, **Developed regions** are selected. The **Age group**, **Gender**, **Asymptomatic**, **Hospital status**, **Region** are categorical variables. The **Fatal outcome**, **Developed regions** are processed to take binary values because these two variables will be used as the response variable to fit the logistic regression models in the Model section. Also, these variables were selected because they are important in the following analysis. More detailed reasons are discussed in the Model section.

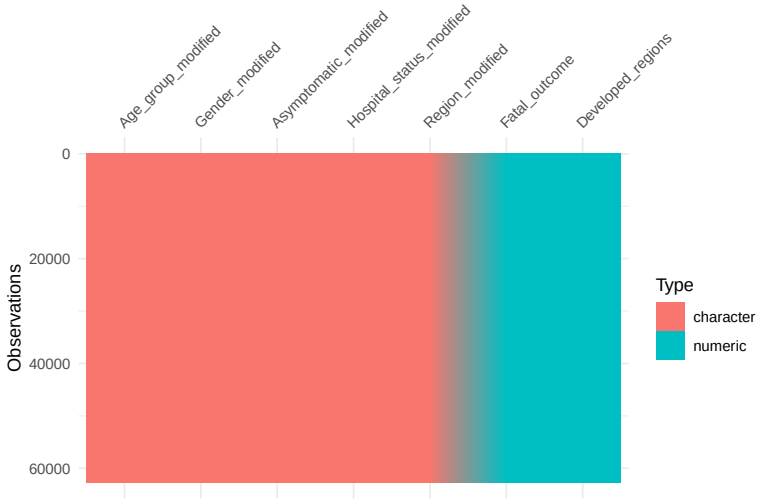

Figure 2: Visualization of the cleaned data

Then visualize what categories does each variable has in the cleaned dataset by plotting pie charts:

From above pie charts (Figure 3), the **Age group** variable has 4 categories: age 0 - 19, age 20 - 39, age 40 - 59, age >= 60. Each age group span two decades. The **Gender** variable has 3 categories: Female, Male, Note stated/Other. The **Asymptomatic** variable has 2 categories: No, Yes. These two values represent whether a COVID-19 infected person is asymptomatic. The **Hospital status** variable has

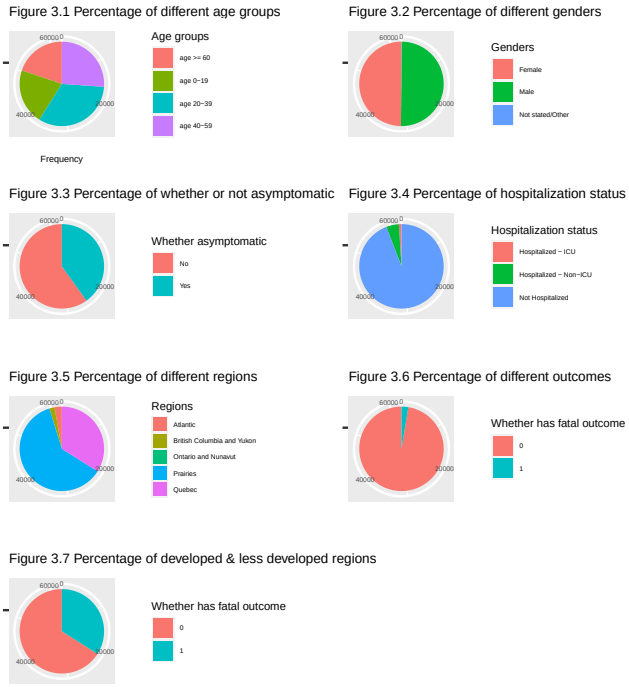

Figure 3: Visualization of the variable categories in cleaned dataset

3 categories: Hospitalized - ICU, Hospitalized - Non-ICU, Not Hospitalized. These three values represent the hospitalization status of a patient. The **Region** variable has 5 categories: Atlantic, Quebec, Ontario and Nunavut, Prairies, British Columbia and Yukon. These five values represent where a patient lives (grouped by geographical regions). The **Fatal outcome** variable has 2 values: 0, 1. 0 represents an individual did not get a fatal outcome, 1 represents an individual got a fatal outcome. The **Developed regions** variable is constructed from the **Region** variable. It has 2 values: 0, 1. 0 represents an individual lives in the less developed regions, 1 represents an individual lives in the developed regions. It is used to make a causal inference. How it is constructed and why it is constructed will be discussed in detail in the Model section.

Model and Methodology

Logistic regression model for the first research question

The first question to investigate is how do various factors influence the fatal rate of COVID-19 infected individuals. Since the fatal rate is computed based on whether an individual died while infected by COVID-19, the response variable is the **Fatal Outcome**. The **logistic regression** model should be utilized since the response variable takes a binary value, which uses 1 to represent that the patient got a fatal outcome due to COVID-19 and 0 to represent not. As described in the Data section above, the predictor variables are **Age group**, **Gender**, **Asymptomatic**, **Hospital status**, and **Region**. The reason for including **Age group** as a predictor is that the Public Health Agency of Canada published an article claiming that older adults have a higher risk of more severe disease or outcomes, and the risk increases with each decade, especially over 60 years (Public Health Agency of Canada, 2020). Therefore, the analysis result derived by using age groups instead of continuous valued ages might be more significant and informative. The **Gender** predictor is also included because some sex-disaggregated data reveal that more males are dying from COVID-19 than females (Canadian Institutes of Health Research, 2020). It is worthwhile to test whether the gender factor influences the fatal rate of COVID-19 infected individuals. The predictor **Asymptomatic**, which represents whether a patient is asymptomatic (infected but no symptom), is included because Public Health Agency also claimed that if the symptoms of difficulty breathing, chest pain or pressure, and difficulty waking up are shown, the patient is at greater risk (Public Health Agency of Canada, 2020). The **Hospital status** predictor, which represents the hospitalization status of a patient, is also included. The reason is that the

patients with severe condition are always hospitalized or even in ICU. So, this hospitalization status factor might have significant influence on the infected people's fatal rate. The **Region** predictor is taken into consideration as well because patients in different regions might have different medical resources, such as the number of doctors and equipment, and there is a study indicates that medical resources have a significant negative correlation with the mortality rate (Chai, Zhang, & Chang, 2020). As a result, the Province or Territory where the case resides may influence the fatal rate. All these predictors are categorical variables and using "dummy" variable coding for modeling.

The logistic regression model to predict how likely a COVID-19 infected individual will get a fatal outcome (death) is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{age\_group} + \beta_2 x_{gender} + \beta_3 x_{asymptomatic} + \beta_4 x_{hospital\_status} + \beta_5 x_{region} + \epsilon \quad (1)$$

where p represents the probability that a COVID-19 infected individual will get a fatal outcome. $\log\left(\frac{p}{1-p}\right)$ is the log odds that a COVID-19 infected individual will get a fatal outcome. The odd represents the proportion of cases get fatal outcomes against the proportion of cases do not get fatal outcomes.

The β_0 is the intercepts of this model, which is the log odds that a COVID-19 infected individual will get a fatal outcome if this individual is female, over 60 years old, has symptoms, hospitalized in ICU, and resides in Atlantic (New Brunswick, Nova Scotia, Prince Edward Island, Newfoundland and Labrador). The x_{age_group} is a categorical variable with 4 categories and the β_1 is its coefficient, which represents the difference between age group over 60 and other age groups' influence to the log odds that a COVID-19 infected individual will get a fatal outcome (here should be 3 coefficients, one for each age group except the age group over 60, use β_1 for simplicity). The x_{gender} is a categorical variable with 3 categories, and the β_2 is its coefficient which represents the difference between gender female and other genders' influence to the log odds that a COVID-19 infected individual will get a fatal outcome (here should be 2 coefficients, but use β_3 for simplicity). The categorical variable $x_{asymptomatic}$ with 2 categories, and the β_3 is its coefficient which represents the difference between an asymptomatic patient and a symptomatic patient of getting a fatal outcome. The $x_{hospital\_status}$ is a categorical variable with 3 categories, and the β_4 is its coefficient which represents the difference between hospitalized in ICU and other hospitalization status's influence to the log odds that a COVID-19 infected individual will get a fatal outcome (here should be 2 coefficients, but use β_4 for simplicity). Last, the x_{region} is a categorical variable with 5 categories, and the β_5 is its coefficient which represents the difference between living in Atlantic region and other regions' influence to the log odds that a COVID-19 infected individual will get a fatal outcome (here should be 4 coefficients, but use β_5 for simplicity). The ϵ is the error term.

Propensity score matching method for the second research question

The second question to investigate is medical resources equity in Canada. Are the medical resources equally distributed in different regions in Canada? There is a study shows that the more developed areas tend to have more medical resources (hospitals, doctors, equipment) allocated and have lower mortality rates than the less developed areas (Chai, Zhang, & Chang, 2020). Therefore, the medical resources equity within Canada can be reflected by investigating the COVID-19 fatal rate in the developed regions vs. in the less developed regions in Canada.

First, we need to define which regions are the developed areas, and which regions are the less developed areas in Canada. Based on the Real Gross Domestic Product (GDP) of Canada provinces in 2019 (Duffin, 2020), the top two developed provinces in Canada are Ontario and Quebec. Since the **Region** variable of the dataset has 5 categories: *Atlantic (New Brunswick, Nova Scotia, Prince Edward Island, Newfoundland and Labrador)*; *Quebec*; *Ontario and Nunavut*; *Prairies (Manitoba, Saskatchewan, Alberta) and the Northwest Territories*; *British Columbia and Yukon*, let's define the categories *Ontario and Nunavut* and *Quebec* to be the developed regions in Canada, and the remaining 3 categories to be the less developed regions in Canada. As a result, the treatment is the COVID-19 infected individual lives (gets hospitalized) in the developed regions (Ontario and Nunavut, Quebec).

The most naive approach is performing an experiment by assigning the COVID-19 infected individuals into two groups (live or hospitalized in developed regions vs. live or hospitalized in less developed regions)

and determine whether there is a causal link between the living regions and the COVID-19 fatal rate by analyzing the result. However, this approach is unethical, infeasible, and impractical.

Since we have observational data containing the geographical information and patients' outcomes, the propensity score matching method can be used to "assign" observations to different groups and make causal statements between whether the patient resides in developed regions and whether the patient gets a fatal outcome. The treatment is the COVID-19 infected individual lives (gets hospitalized) in the developed regions. The outcome of interest is whether the patient gets a fatal result.

The steps for doing propensity score matching in this paper are:

Step 1. Estimate propensity score:

Fit a logistic regression where whether a patient lives (gets hospitalized) in the developed regions is the outcome of interest, and the other variables except the region are the predictors. The logistic regression model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 x_{age\_group} + \beta_2 x_{gender} + \beta_3 x_{asymptomatic} + \beta_4 x_{hospital\_status} + \epsilon \tag{2}$$

very similar to the logistic regression equation (1), p represents the probability that a COVID-19 infected individual lives (gets hospitalized) in the developed regions. $\log\left(\frac{p}{1-p}\right)$ is the log odds that a COVID-19 infected individual lives (gets hospitalized) in the developed regions. The odd represents the proportion of cases live (get hospitalized) in the developed regions against the proportion of cases do not live (get hospitalized) in the developed regions.

The variables x_{age_group}, x_{gender}, $x_{asymptomatic}$, $x_{hospital\_status}$ are exactly the same variables in the logistic regression equation (1). Also, the coefficients β_0, β_1, β_2, β_3, β_4 have same meaning as in equation (1). The β_0 is the intercept, β_1, β_2, β_3, β_4 are the coefficients for variables x_{age_group}, x_{gender}, $x_{asymptomatic}$, $x_{hospital\_status}$ respectively. The ϵ is the error term.

Based on the **Age group**, **Gender**, **Asymptomatic**, **Hospital status** information, use the above fitted logistic regression model to compute each case's propensity of living (getting hospitalized) in the developed regions regardless of whether this patient lives (gets hospitalized) in the developed regions.

Step 2. Match:

Use the matching function in the arm package (Gelman et al., 2020) to find pairs (one is treated and the other is untreated) of similar propensity scores. In this paper, for each case that lives (gets hospitalized) in the developed regions, we want to find another case with similar propensity score of living (getting hospitalized) in the developed regions but does not live (get hospitalized) in the developed regions in reality. Then create a new dataset consisting of all the matched pairs.

Step 3. Evaluate quality of matching:

Graphically compare the age, gender, asymptomatic, hospital status of the cases in the treatment group and in the comparison group. If the distributions are similar in the treatment group and in the comparison group, we have a good matching.

Step 4. Evaluate outcomes:

Fit a logistic regression where whether a patient gets a fatal outcome is the response variable, and **Age group**, **Gender**, **Asymptomatic**, **Hospital status**, **Developed regions** are predictors. This can show the effect of the treatment of living or getting hospitalized in the developed regions. The rationale is: we divided the data into groups similar before the treatment, and hence any differences should be caused by the treatment. The logistic regression model is:

$$\log\left(\frac{p}{1-p}\right) = \beta_0+\beta_1 x_{age\_group}+\beta_2 x_{gender}+\beta_3 x_{asymptomatic}+\beta_4 x_{hospital\_status}+\beta_5 x_{developed\_regions}+\epsilon \tag{3}$$

very similar to the logistic regression equation (1), p represents the probability that a COVID-19 infected individual will get a fatal outcome. $\log\left(\frac{p}{1-p}\right)$ is the log odds that a COVID-19 infected individual will get a fatal outcome. The odd represents the proportion of cases get fatal outcomes against the proportion of cases do not get fatal outcomes.

The variables x_{age_group}, x_{gender}, $x_{asymptomatic}$, $x_{hospital\_status}$ are exactly the same variables in the logistic regression equation (1). Also, the coefficients β_0, β_1, β_2, β_3, β_4 have same meaning as in equation (1). The β_0 is the intercept, β_1, β_2, β_3, β_4 are the coefficients for variables x_{age_group}, x_{gender}, $x_{asymptomatic}$, $x_{hospital\_status}$ respectively. The new variable $x_{developed\_regions}$ is a categorical variable with 2 categories. β_5 is its coefficient, which represents the difference of whether or not living in a developed region's influence to the log odds that a COVID-19 infected individual will get a fatal outcome. The ϵ is the error term.

Results

The logistic regression model (1) in the Model section is fitted to model how likely a confirmed COVID-19 individual gets a fatal outcome. The fitted logistic regression model (1) can be used to predict the probability that a COVID-19 infected individual will get a fatal outcome based on the predictors **Age group**, **Gender**, **Asymptomatic**, **Hospital status**, and **Region**. The model summary is shown in the Table 1 below:

Table 1: Model Summary for the logistie regression model (1)

	Estimate	Std. Error	z value	Pr(>|z|)
(Intercept)	-0.3169788	0.1771081	-1.7897476	0.0734945
Age 0-19	-19.1526907	252.8392076	-0.0757505	0.9396176
Age 20-39	-19.2209765	202.7186378	-0.0948160	0.9244610
Age 40-59	-4.1273915	0.1922107	-21.4732634	0.0000000
Gender - Male	0.0723431	0.0558116	1.2962023	0.1949058
Gender - Not stated/Other	1.5447618	0.5648846	2.7346505	0.0062447
Asymptomatic - Yes	-0.0139057	0.0835873	-0.1663617	0.8678723
Hospital status - Hospitalized - Non-ICU	-0.5758630	0.1090455	-5.2809428	0.0000001
Hospital status - Not Hospitalized	-2.1769059	0.1066597	-20.4098325	0.0000000
Region - British Columbia and Yukon	-0.3525936	0.2040921	-1.7276200	0.0840564
Region - Ontario and Nunavut	-1.7583681	1.0497387	-1.6750531	0.0939238
Region - Prairies	0.1235493	0.1498276	0.8246094	0.4095934
Region - Quebec	-0.0963468	0.1655572	-0.5819549	0.5605971

The step 3 of the propensity score matching method is evaluating the quality of matching. Since for each case, the values of the age, gender, asymptomatic, hospital status are categorical, the distribution of cases in the treatment group and comparison group can be visualized by the pie charts:

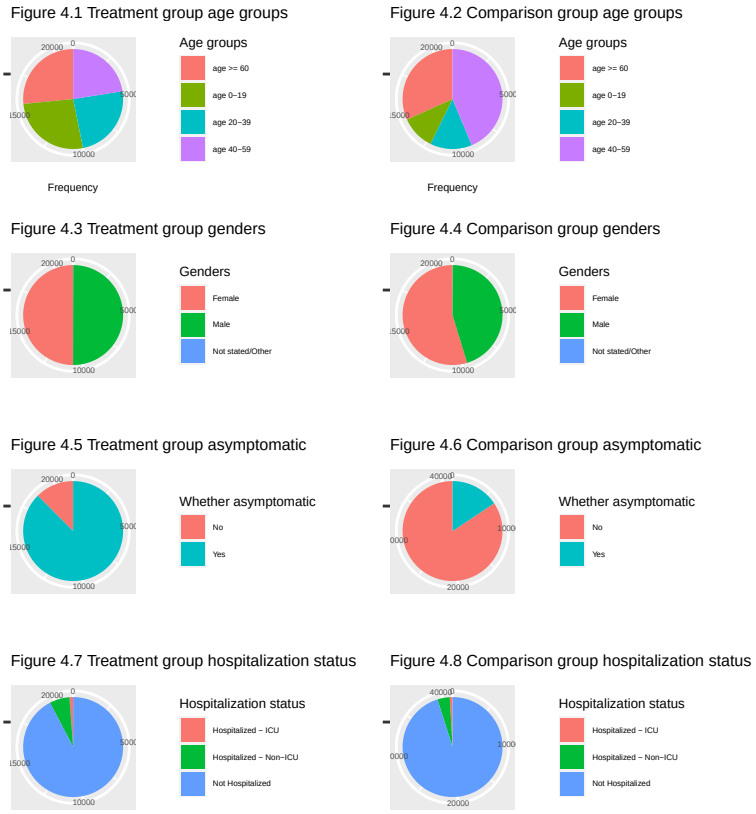

Figure 4: Treatment group and Comparison group cases distribution

Altogether, the distribution of cases in the treatment group and the comparison group are not very similar, which implies that the treatment group and the comparison group "assigned" by the propensity score matching method are not very identical before the treatment. In this case, the conclusion drawn from the propensity score matching method may not be very accurate.

Table 2: Model Summary for the logistie regression model (3)

	names	model1
		(1)
1	(Intercept)	-0.297329418890301 **
2		(0.105296271190076)
3	Age_group_modifiedage 0-19	-18.1045860909544
4		(197.446307412119)
5	Age_group_modifiedage 20-39	-18.2232643318659
6		(194.75135537968)
7	Age_group_modifiedage 40-59	-4.01533086474804 ***
8		(0.195424695251114)
9	Gender_modifiedMale	0.0775771492112437
10		(0.0557633859756372)
11	Gender_modifiedNot stated/Other	1.66509866338849
12		(1.27358411569472)
13	Asymptomatic_modifiedYes	0.0155319521487685

	names	model1
14		(0.0830730856217027)
15	Hospital_status_modifiedHospitalized - Non-ICU	-0.530549235513452 ***
16		(0.107961599005682)
17	Hospital_status_modifiedNot Hospitalized	-2.11099396324454 ***
18		(0.104779234345049)
19	Developed_regions	-0.203180125775451 *
20		(0.0829249196991848)
1.1	N	42590
2.1	logLik	-4521.97878993819
3.1	AIC	9063.95757987639
.1	*** p < 0.001; ** p < 0.01; * p < 0.05.	

The Table 2 generated using the huxtable package (Hugh-Jones, 2020) shows that whether a COVID-19 infected individual lives (gets hospitalized) in the developed regions (Ontario and Nunavut, Quebec) in Canada is significant in influencing whether a COVID-19 infected individual gets a fatal outcome. This implies that the COVID-19 infected individual lives (gets hospitalized) in the developed regions (Ontario and Nunavut, Quebec).

Discussion

Result Explanation

The summary information of the logistic regression model shown in Table 1 indicates how likely a confirmed COVID-19 infected individual gets a fatal outcome based on the age, gender, whether is Asymptomatic, hospitalization status, and living region factors. The intercept and many predictor are statistically significant (p-value smaller than 0.05), which implies that these factors influence the probability of getting a fatal outcome for a COVID-19 infected individual. The estimated coefficients of each dummy variable reflect its influence to the log odds of death due to COVID-19 compared to its reference category. In specific, the people older than 60 tends to have a higher probability of getting a fatal outcome than the people in other age groups, and people between 20 to 39 years old tends to have the lowest probability of getting a fatal outcome among all four age groups. This is consistent with the article, published by the Public Health Agency of Canada, that older adults have a higher risk of more severe disease or outcomes, and the risk increases with each decade, especially over 60 years (Public Health Agency of Canada, 2020). Similarly, the female tends to have the lowest probability of getting a fatal outcome, the male has a little bit higher probability of getting a fatal outcome, and the other gender has the highest probability of getting a fatal outcome. This is consistent with the result of a study from Canadian Institutes of Health Research that more males are dying from COVID-19 than females (Canadian Institutes of Health Research, 2020). The coefficient of the dummy variable Asymptomatic indicates that asymptomatic patients tend to have a lower probability of getting a fatal outcome, although it is not statistically significant that whether an individual is asymptomatic will influence the probability of getting a fatal outcome. The patients hospitalized in ICU have the highest probability of getting a fatal outcome, the patients hospitalized not in ICU have the lower probability, and the not hospitalized people have the lowest probability. This is consistent with the fact that patients with severe condition are always hospitalized or even in ICU. Last, the cases reside or get treated in Prairies tend to have the highest probability of getting a fatal outcome when compared with other regions. This might be a piece of evidence that the medical resources are unequally distributed within Canada, and this question is discussed in more detail in the paragraph below. In general, the estimated coefficients of variables indicate that the male and other gender people older than 60, who are symptomatic, get hospitalized in ICU, resides in Prairies regions have the highest risk of getting fatal outcome.

The Figure 4 pie charts shown in the Results section above are used to evaluate the quality of matching. The pie charts compare the distribution of cases in the treatment group and comparison group by graphically showing the percentage of values in both groups. From the pie charts Figure 4.1 and Figure 4.2, the distribution of age groups in the treatment group and the comparison group are not very similar.

In the treatment group, the percentages of age groups are close (around 1/4 for each age group), but in the comparison group, there are more age 40-59 and age >= 60 cases than other cases. From the pie charts Figure 4.3 and Figure 4.4, the distribution of genders in the treatment group and the comparison group are similar. In both treatment group and the comparison group, the percentage of female and male cases are around 1/2, and the percentage of the other gender is close to 0. From the pie charts Figure 4.5 and Figure 4.6, the distribution of whether a patient is asymptomatic in the treatment group and the comparison group are very different. There are more asymptomatic cases in the treatment group, but more symptomatic cases in the comparison group. From the pie charts Figure 4.7 and Figure 4.8, the distribution of hospitalization status in the treatment group and the comparison group are very similar. In both treatment group and the comparison group, the proportion of not hospitalized cases is largest, then follows the proportion of hospitalized but not in ICU cases, and the proportion of hospitalized and in ICU is the smallest. The results reveal that the percentage of ages, and the percentage of asymptomatic patients in the treatment group and comparison group are quite different. Therefore, the treatment group and the comparison group "assigned" by the propensity score matching method are not very identical before the treatment.

The summary information of the logistic regression model shown in Table 2 shows that the p-value of variables age group, hospitalization status, and developed regions are smaller than 0.05, which means they are statistically significant. In particular, whether a COVID-19 infected individual lives (gets hospitalized) in the developed regions (Ontario and Nunavut, Quebec) in Canada is significant in influencing whether a COVID-19 infected individual gets a fatal outcome. Since the coefficient is negative, it implies that living (getting hospitalized) in the developed regions in Canada will lower the probability of getting a fatal outcome. This result reflects that it is very likely the developed region in Canada has more medical resources, and the COVID-19 infected individual tends to have more and better access to medical aids such as hospitals, doctors, and equipment. This conclusion and reasoning might be inaccurate due to the bad quality of matching discussed above.

Summary

In this paper, two questions are analyzed by using a published dataset called "Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada" on the Statistics Canada website (Statistics Canada, 2020), which contains detailed information about each confirmed cases in Canada. First, we investigate how and to what extend various factors influence the fatal rate of COVID-19 cases by using a logistic regression model. The raw dataset is cleaned by removing missing values and selecting the potentially important variables including age, gender, asymptomatic, hospital status, region, and fatal outcome. Then a logistic regression model is fitted by using the cleaned data, and this model can statistically show the influence of each factors to the COVID-19 fatal rate. Second, we conduct a causal inference on this observational data by using propensity score method to study whether there is a causal link between the region of living or getting hospitalized and whether the patients get fatal outcomes. It is suspected that COVID-19 infected individuals reside or get hospitalized in the developed regions in Canada are less likely to get fatal outcomes because the developed regions tend to have more medical resources than the less developed regions. Based on the region information in the dataset and the 2019 GDP of provinces in Canada (Duffin, 2020), the Provinces and Territories in Canada are divided into two groups. The developed regions are Ontario and Nunavut, Quebec, and the less developed regions are the remaining regions in Canada. Then "assign" the cases into treatment group (in developed regions) and comparison group (in less developed regions) by computing the propensity score and matching. Then a logistic regression model is fitted using the matched pairs, and this model can statistically show the influence of living or getting hospitalized in developed areas to the fatal outcome. Thus, the existence of causal link between the region of living or getting hospitalized and whether the patients get fatal outcomes will be revealed.

Conclusions

Based on above statistical analysis and results, two conclusion can be drawn. First, the factors of age, gender, hospitalization status, and residence region significantly influence the probability that a COVID-19 infected individual gets a fatal outcome. The factor of whether the patient is asymptomatic may have a minor influence. The male and other gender people older than 60, who are symptomatic, get hospitalized in ICU, resides in Prairies regions have the highest risk of getting fatal outcome. Thus, people in this group

should take extra precautions to reduce the risk of getting sick from COVID-19. Second, although not very strong, there is evidence that medical resources inequality exists in Canada. The developed regions, such as Ontario and Quebec, tend to have more medical resurces than other regions in Canada. The Canadian government should consider allocating more medical resources to the less developed regions to achieve health equity and, more importantly, reduce the fatal rate in the less developed areas due to COVID-19.

Weaknesses The weaknesses for this analysis are:

1. To ensure the confidentiality, Statistics Canada grouped some detailed information into large groups, such as the provinces or territories have been grouped into 5 broad regions (Statistics Canada, 2020). Since the regions are grouped according to the geographical location rather than economic factors, inaccuracy is introduced to the causal inference, as some developed areas in Canada are grouped with less developed areas.
2. When doing propensity score matching, the cases in the treatment groups and the comparison groups are not identical before treatment. This weakens the conclusion drawn from the causal inference. There might be more appropriate and sophisticated methods to perform data matching, and a more accurate conclusion can be drawn.

Next Steps

1. Find another COVID-19 dataset with more fine-grained provinces or territories information for each case, and use them to perform the same causual inference again. A more appropriate way to divide the Canadian provinces and territories into developed and less developed regions is by the mean GDP. The provinces or territories with GDP higher than the average are grouped into the developed regions, and the provinces or territories with GDP lower than the average are grouped into the less developed regions. This division method is impractical for the dataset used in this paper, as many provinces and territories are grouped based on geographical location.
2. According the user guide (Statistics Canada, 2020), this dataset "Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada" is updated regularly given that the COVID-19 pandemic is still progressing. So, I will perform the analysis periodically to see if there are any changes to the conclusion. The results can be compared throughout time to see the trend.
3. As mentioned in the weaknesses of analysis, propensity score matching conducted in this paper does not have perfect matching quality. Therefore, more sophisticated methods can be tried to perform data matching, and more accurate conclusions can be drawn.

References

1. WHO Coronavirus Disease (COVID-19) Dashboard. (2020). Retrieved December 20, 2020, from https://covid19.who.int/
2. Public Health Agency of Canada. (2020). Government of Canada People who are at risk of more severe disease or outcomes from COVID-19. Retrieved December 21, 2020, from https://www.canada.ca/en/public-health/services/publications/diseases-conditions/people-high-risk-for-severe-illness-covid-19.html
3. Chai, K., Zhang, Y., & Chang, K. (2020). Regional Disparity of Medical Resources and Its Effect on Mortality Rates in China. Frontiers in Public Health, 8. doi:10.3389/fpubh.2020.00008
4. Wu, C., & Thompson, M. E. (2020). Sampling Theory and Practice. Springer International Publishing.
5. Rosenbaum, P. R., & Rubin, D. B. (1982). The central role of the propensity score in observational studies for causal effects. Madison, WI: Wisconsin Clinical Cancer Center, Biostatistics.
6. Statistics Canada. (2020). Preliminary dataset on confirmed cases of COVID-19, Public Health Agency of Canada. Retrieved December 22, 2020, from https://www150.statcan.gc.ca/n1/pub/13-26-0003/132600032020001-eng.htm

7. Robinson, David, Alex Hayes, and Simon Couch. (2020). Broom: Convert Statistical Objects into Tidy Tibbles, from https://CRAN.R-project.org/package=broom.

8. 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, from https://doi.org/10.21105/joss.01686.

9. Robinson, David, Alex Hayes, and Simon Couch. (2020). Broom: Convert Statistical Objects into Tidy Tibbles, from https://CRAN.R-project.org/package=broom.

10. Tierney, N., & Hughes, S. (2019). Preliminary Visualisation of Data [R package visdat version 0.5.3]. Retrieved December 22, 2020, from https://cran.r-project.org/web/packages/visdat/index.html

11. Canadian Institutes of Health Research. (2020). Why sex and gender need to be considered in COVID-19 research. Retrieved December 22, 2020, from https://cihr-irsc.gc.ca/e/51939.html

12. Gelman, A., Su, Y., Yajima, M., Hill, J., Pittau, M. G., Kerman, J., . . . Dorie, V. (2020). Data Analysis Using Regression and Multilevel/Hierarchical Models [R package arm version 1.11-2]. Retrieved December 22, 2020, from https://cran.r-project.org/web/packages/arm/index.html

13. Duffin, E. (2020). Canada: Real Gross Domestic Product (GDP), by province 2019. Retrieved December 22, 2020, from https://www.statista.com/statistics/463905/canada-real-gross-domestic-product-by-province/

14. Hugh-Jones, D. (2020). Easily Create and Style Tables for LaTeX, HTML and Other Formats [R package huxtable version 5.1.1]. Retrieved December 22, 2020, from https://cran.r-project.org/web/packages/huxtable/index.html

15. Alexander, R. (2020). Difference in differences. Retrieved December 23, 2020, from https://www.tellingstorieswithdata.com/06-03-matching_and_differences.html

16. Robinson, David, Alex Hayes, and Simon Couch. (2020). Broom: Convert Statistical Objects into Tidy Tibbles, from https://CRAN.R-project.org/package=broom.

17. Auguie, B. (2019, July 13). Extensions for 'ggplot2': Custom Geom, Custom Themes, Plot Alignment, Labelled Panels, Symmetric Scales, and Fixed Panel Size [R package egg version 0.4.5]. Retrieved December 23, 2020, from https://cran.r-project.org/web/packages/egg/index.html