# Analysing the Fatality of COVID-19 Cases Based on Aquisition

Chun Ki Yip - 1004532589

December 17, 2021

## Abstract

This report uses the "Confirmed Positive Cases of COVID-19 in Ontario" data set from Open Ontario Portal to answer to research question of "Does COVID-19 that are travel-related have a higher fatality rate?" The hypothesis drawn was that "Travel-related COVID-19 cases would have a higher fatality rate" because it travel-related COVID-19 cases have a higher chance to be a deadlier and new variant of the virus. Through propensity score matching and building a logistical regression, evidences show supports the complete opposite of the hypothesis and supports that travel-related to have a lower fatality rate, which could be explained by poor propensity score building which put a relatively healthier batch of observations into the travel-related "treatment" group, or that there exists a even deadlier acquisition of COVID-19, and a speculation is that it could be acquisition through outbreak.

**Keywords: COVID-19, Travel Restriction, Casual Inference, Propensity Score Matching, Observational Study, Ontario, Epidemiology.**

## Introduction

COVID-19 has been a global public health issue since the beginning of 2020. For the past 2 years Canada has implemented many measures to help reduce cases, one of the first measure being travel restrictions that began from March of 2020, restrictions include travel bans, proof of vaccination, and positive COVID-19 test results. This report is interested in examining whether cases acquired through travelling are more deadly by using the data set of all confirmed positive cases of COVID-19 in Ontario from Open Ontario Portal [5] and performing a propensity score matching to make quasi-experiment where acquiring COVID-19 through travel is the treatment of the experiment. This data set includes some preliminary data of all confirmed positive cases in Ontario since 2019 updated daily until this day. The data are collected by reports of Public Health Units across Ontario, and individual observations are collected by health workers by logging the data into the hospital system or they are existing information on the individual's health card. This report would analyze how the fatality of COVID-19 differs from cases acquired through travel and non-travel by using propensity score matching and logistic regression with the observational data. In other words, this report would be answering the research question of:

***Does COVID-19 that are travel-related have a higher fatality rate?***

This research question is especially important during this time as cases of the new variant of COVID-19 rises around the holidays, it is crucial for everyone, not only for government policy makers, to understand the effectiveness of travel restrictions and the danger of travelling so everyone could do their part in reducing cases.

The following are some terminologies that would be used in this report.

- PHU is the abbreviation for Public Health Unit which are hospitals and other government health institutions.
- Observational data is when researchers observe the effect of an independent variable instead of performing an experiment where they control the treatment.
- Logistic regression is a statistic model that predicts the probability of a binary result using various variables.
- Propensity Score Matching is a method that mimics the process of randomized placement of subjects in a treatment group by matching pairs of subjects with the same tendency to either have the treatment or not have the treatment by looking at all the other variables.

Hypothesis: ***Travel-related COVID-19 cases would have a higher fatality rate.***

Overall, travel-related COVID-19 cases would have a lower fatality rate due to strict travel restrictions. But using propensity score matching to match people with the same tendency to travel and not travel should reduce the effect of the hidden assumptions that only healthier and people with less risk of COVID-19 would travel. By only examining the danger of travel-related COVID-19 cases versus non-travel-related COVID-19 cases, travel-related COVID-19 cases should have a higher fatality rate due to increase chances of new variants from a different country. Although, the lack of more detailed data such as vaccination status, variant of the virus, and other history of disease might potentially decrease the accuracy of the propensity score matching where a relatively healthier person would still be matched with a relatively not as healthy person and still resulting in a lower fatality rate in travel-related COVID-19 cases.

# Data

## Data Collection Process

Data collection of individual observations in this data set is not specified but it probably differs case by case or differs from PHU to PHU. Some common methods of data collection in PHUs are data logging to the PHU's archive system by health workers as they observe the patients, surveying form filled in by the patient or health workers when the patient first enter the PHU, or existing information such as gender and age in the health card. After the data is collected within the PHU, Public Health Ontario collects these data and compile it to form this data set along with PHU specific data like the PHU's city, address, and website [6].

Some foreseeable drawbacks of this data collection process is that the acquisition of the case might be hard to be completely accurate as it is impossible to trace back the exact time of acquisition. Data may also be inputted incorrectly when logging into the PHU system as health workers have to deal with an enormous amount of patients every day. The data set is also incomplete because not all positive cases are reported to a PHU. Some limitations is that it could include more information such as vaccination status, other history of disease and variant of the virus which are important variable for any analysis related to COVID-19, but there may be ethical issues in providing such information for a public data set.

## Important Variables

Some important variables in this data set are the following:

- Patient's age group: A discrete variable which categorizes the patient's age into everyone who is below 20, above 90, increments of 10 years between 20 and 90, and "UNKNOWN" for observations with unknown age.
- Gender: A discrete variable of the gender information of the patient which includes the values: "Female", "Male", "Gender Diverse", and "Unspecified."

- Case Acquisition: A discrete variable that gives the suspected method of exposure to COVID-19 if it's known. The values include: "Close Contact", "No Known Epidemiological Link", "Unspecified Epidemiological Link", "Outbreak", "Travel", "Missing Information".

- Outcome: A discrete variable that shows the outcome of the case which could have values: "Resolved", "Not Resolved", or "Fatal".

- Reporting PHU: A discrete variable that shows the PHU where this case is confirmed.

## Data Cleaning Process

1. The data was downloaded from the Open Ontario Portal as a .csv file and loaded into R studios [5].

2. The observations with age group "UNKNOWN" is removed using the filter function [8].

3. The observations with acquisition of "Outbreak", "Close Contact", and "Travel" are filtered using the filter function, i.e. the cases with acquisitions that are unknown, unspecified, and missing are removed [8].

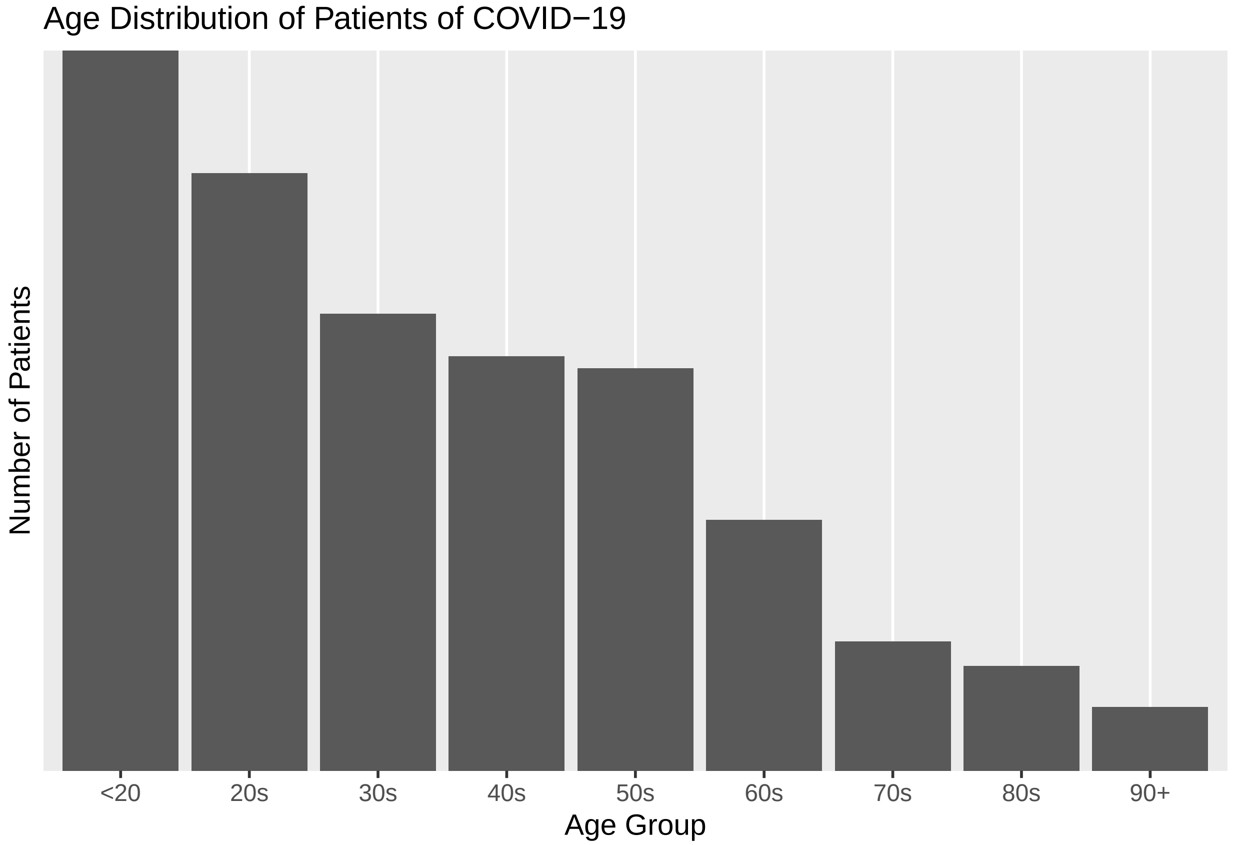
4. The observations with outcome of "Not Resolved" is removed using the filter function [8].

5. A new binary variable named "Travel-related" which indicates whether the case was acquired by travel or not was created using the mutate function [8].

6. The variables; age group, gender, reporting PHU, travel-related, and outcome were chosen to be the remaining variables that would be used and shown using the select function [8].

## Data Summary

Figure 1 [8]



This bar graph shows the age group distribution of the patients with positive case of COVID-19. This graph gives the readers a general idea of the demographics of this data set and it could also be used as a graphical comparison to the post propensity score matching data set to check if the general shape of distribution of the demographics has changed drastically.

Figure 2 [9]

Table 1: Table of Outcome of Patients of Different Acquisition of COVID-19

| Acquisition | Outcome | Count |
|---|---|---|
| Travel Related | Resolved | 14648 |
| Travel Related | Fatal | 108 |
| Non-Travel Related | Resolved | 367413 |
| Non-Travel Related | Fatal | 7326 |

This table aggregates 4 important groups of observations; patients with fatal outcomes and acquired COVID-19 through travel, patients with resolved outcomes and acquired COVID-19 through travel, patients with fatal outcomes and acquired COVID-19 through non-travel ways, and patients with resolved outcomes and acquired COVID-19 through non-travel ways. This table is important for readers to know the fatality rate of acquisitions travel versus non-travel COVID-19 before propensity score matching, and the difference in number in travel versus non-travel cases.

The fatality rate of travel-related COVID-19 cases is 0.0073191, and the fatality rate of the non-travel-related COVID-19 cases is 0.0195496. The fatality rate of travel-related COVID-19 cases are lower than non-travel-related COVID-19 cases which makes sense because of the strict travel restrictions which only allows lower risk people to travel, meaning even if the person got COVID-19 they should have a higher rate of survival as they are relatively healthier people with stronger immunity due to them being vaccinated for sure. This statistic is important because it is important for the readers to know the fatality rate before propensity score matching so readers could see the effect of the propensity score matching.

# Methods

The statistical methods that would be used to analyze the data are propensity score matching and logistic regression. Propensity score matching is a method that mimics an experiment with randomized treatment and non-treatment group by matching the each observation which is deciphered as “treated” to an observation which is deciphered as “not treated” with similar tendencies that to be either group. The use of propensity score matching entails that there is some variable in this observational study that could be viewed as the “treatment”. Logistic regression is a statistical model that predicts the probability of a binary outcome through various independent variables. The use of this statistical model entails that it is modeling something with a binary outcome.

Some assumptions of propensity score matching is there are characteristics which treated subjects could be compared to untreated subjects, i.e., a propensity score could be calculated [4]. In this case the ideal characteristics that the data would want to compare is how healthy the subject is, and that two subjects with similar propensity score should mean that they have the same tendency to acquire COVID-19 through non-travel and travel since they have the same health status. This assumption is somewhat fulfilled with variables such as age group and gender which could be used to compare subjects, but the propensity score matching would be more accurate if more health related variables such as vaccine status, disease history, and COVID-19 variant was provided. Another variable that would be useful for this comparison is the reporting PHU, as subjects that are treated in the same PHU are more likely to live in a similar area and have more comparable characteristics but this is also just more likely but not completely true.

The variables that was chosen to be the “treatment” was whether or not the case was travel-related because the research question was to answer whether COVID-19 cases acquired through travel had a higher fatality rate. The variables that are used to create a logistic regression model after the propensity matching are the patient’s age group, gender, whether it is travel-related or not, and the reporting PHU because these were the available variables that were meaningful and independent to each other when modeling fatality and were not repeated variables that represented the same thing like the longitude and latitude of the PHU, city of the PHU, website of the PHU, address of the PHU, and etc. Date related variables are also meaningful but were ultimately excluded due to the difficulty to implement in a model.

The equation of the logistic regression model that calculates the propensity score is the following:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{agegroup} + \beta_2 x_{PHU} + \beta_3 x_{gender} + \epsilon_i$$

Each $\beta$ represents a parameter variable and $\epsilon_i$ represents the error term.

The equation of the logistic regression model that calculates the fatality rate is the following:

$$log(\frac{p_i}{1-p_i}) = \beta_0 + \beta_1 x_{agegroup} + \beta_2 x_{PHU} + \beta_3 x_{gender} + \beta_4 x_{travel} + \epsilon_i$$

Similarily, each $\beta$ represents a parameter variable and $\epsilon_i$ represents the error term.

## Process of Propensity Score Matching and Logistic Regression

1. A model which is used to calculate the propensity score of the observation being travel-related or not is created using a logistic model with variables of age group, gender, and the reporting PHU.

2. Another variable called the propensity score is added to each observation by plugging in the observation's variable values into the propensity score model.

3. The binary variable of the observation being travel-related or not is duplicated and set as an integer value for matching purposes [8].

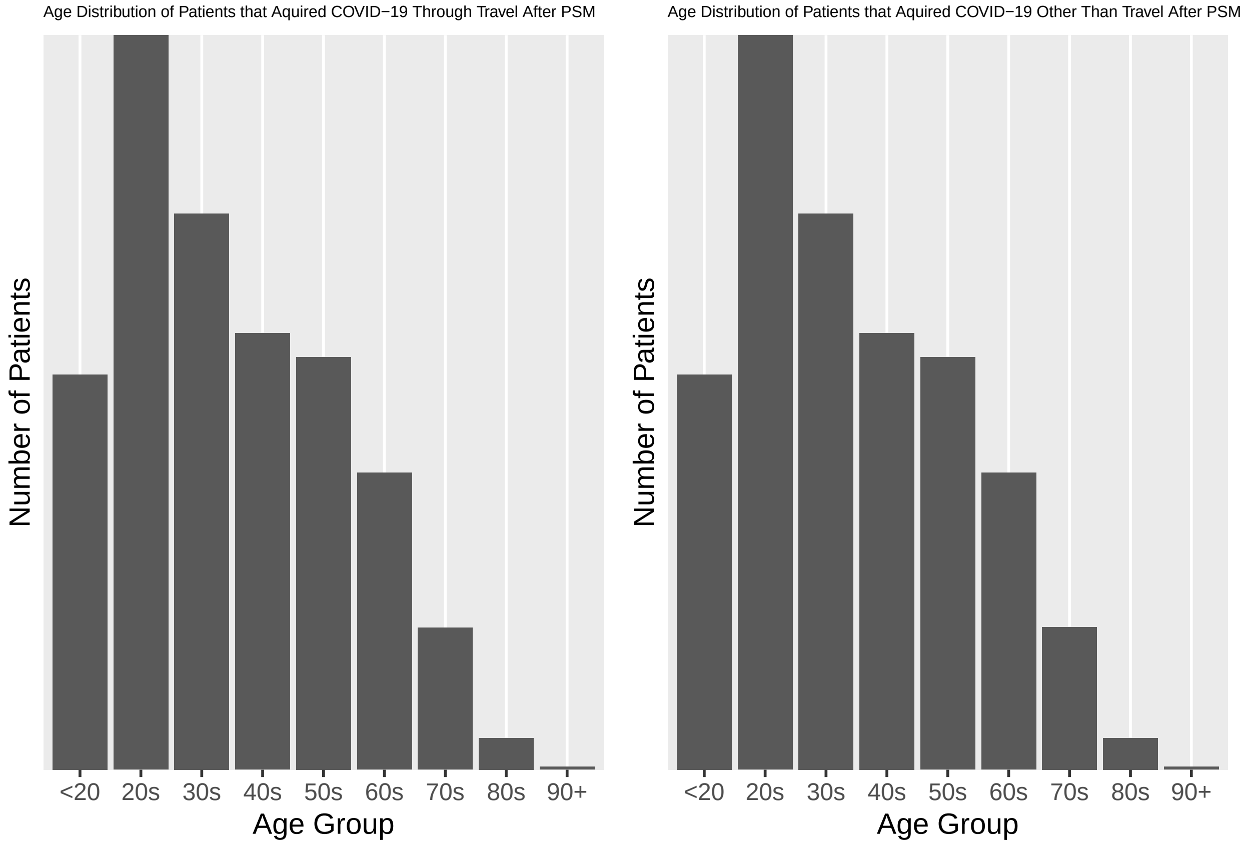
4. The arms function is used for matching each observation of the treated observation with a untreated observation with the closest propensity score [2].

5. The unmatched observations would be removed from the data set with the filter function [8].

6. The treated data set and the untreated data set would be separated to do a graphical [8] comparison [1] between the two demographics of age distribution of patients as a preliminary check of the legitimacy of the propensity matching. If the two graphs have similar distribution, it means that the propensity score matching process didn't disturb with other variables.

7. Finally, using the matched data to create a logistic model predicting the fatality rate of the patient by using the variables; whether it is travel-related, the age group, gender, and reporting PHU.

Figure 5 [1] [8]



This side by side bar graph shows the graphical comparison of the age distribution between the treated and non-treated groups. The similar shape in the graphs shows that the propensity score didn't disturb with the age variable, and from this it would be assumed that it didn't disturb with other variables too.

# Results

Figure 4 [9]

Table 2: Table of Outcome of Patients of Different Acquisition of COVID-19 Post PSM

| Acquisition | Outcome | Count |
|---|---|---:|
| Travel Related | Resolved | 14648 |
| Travel Related | Fatal | 108 |
| Non-Travel Related | Resolved | 14396 |
| Non-Travel Related | Fatal | 360 |

This table shows the same 4 groups of observations as figure 2 but this time after propensity score matching. The groups being; patients with fatal outcomes and acquired COVID-19 through travel, patients with resolved outcomes and acquired COVID-19 through travel, patients with fatal outcomes and acquired COVID-19 through non-travel ways, and patients with resolved outcomes and acquired COVID-19 through non-travel ways. This table is important because it shows the fatality rate of the patient when the variable travel-related is viewed as the "treatment" through propensity score matching, and this table could be used to calculate the fatality rate.

It could be seen that the travel-related fatal and resolved case remained the same as figure 2 meaning the fatality rate of travel-related cases also remained the same as 0.0073191. This wasn't expected when drawing the hypothesis but makes sense because there are less travel-related cases compared to non-travel-related cases, so during the process of propensity matching, which the closest propensity score are matched, all the travel-related cases would be matched up. Another surprising result is that the fatality rate of non-travel-related cases actually increased from 0.0195496 to 0.0243969 after propensity score matching, this may be explained by the fact the characteristics used to determine the propensity score wasn't good enough to pair two subjects with the same exact same health conditions so people who acquire COVID-19 through travel was still relatively more healthy than people who didn't. Another possible reason was that there were other methods of acquisitions that were potentially more dangerous than travel, such as an outbreak, as an outbreak of the same area would put a lot of capacity pressure on local PHUs and increase the fatality rate due to lack of resources.

Figure 5 [7]

| | (1) |
|---|---|
| (Intercept) | -0.019 |
| | (0.016) |
| Travel_Related | -0.017 *** |
| | (0.001) |
| Age_Group20s | 0.000 |

This is a preview of the logistic regression parameter table that shows the parameter of each variable of the logistic regression and the P-value of each parameter. This logistic regression parameter table is important because it further confirms the result from the fatality rate and also shows the P-value of the parameters to show that this treatment was statistically significant or not.

The preview logistic regression parameter table shows that being travel-related has a -0.017 effect on the fatality rate, meaning there would be a lower fatality rate by 0.017 whenever the case was travel-related. This matches with the fatality rate result above and could be explained with the same reasoning. The logistic parameter table also shows that the parameter of the case being travel-related has a P-value of less than 0.001 meaning that this parameter was very statistically significant, showing that this "treatment" was significant, but to the opposite direction of which was hypothesized.

# Conclusions

The hypothesis of the research question is that ***Travel-related COVID-19 cases would have a higher fatality rate.*** Which was supported by the reasoning that travel-related COVID-19 cases would have a higher chance of being a new and more deadly variant of the virus. The methods that were used to verify this hypothesis was propensity score matching and logistic regression modeling. Propensity score matching was used to treat the variable of travel-related as the "treatment" of an experiment to mimic a randomized experiment where the treatment and non-treatment groups are 50/50 split and are randomized. A logistic regression model was created to see the effect of the travel-related parameter in the model and to see the P-value to see whether or not this "treatment" was significant.

The result shows that after the propensity score matching, the difference in fatality rate between travel-related and non-travel-related cases were even bigger with the travel-related cases being lower, as the fatality rate of travel-related cases were 0.0073191 and non-travel-related cases were 0.0243969 which doesn't support the hypothesis. The parameter of the logistic regression further confirms this, as being travel-related lowers the fatality rate by 0.017, and the parameter also has a P-value lower than 0.001 meaning this parameter is statistically significant. This result could be reasoned by not having good enough characteristics to build the propensity score which matches pairs of observations with the same health conditions [3], resulting in the treatment group to have a relatively healthier batch of subjects compared to the "non-treatment" group as travel restrictions by the government would make this treatment bias. Another reason would be that there is an acquisition that is even deadlier than travel within the non-travel group, for example, an outbreak, because an outbreak would put stress on local PHU capacity and potentially increase fatality rate due to lack of resources.

## Weaknesses

Some limitations of this analysis are again the lack of more detailed data points of individual observations which allows a better propensity score building, if the data set includes more health related variable such as previous health conditions, vaccination status, and COVID-19 variant, it would greatly help the propensity score building. Another weakness of this analysis is that propensity score matching excludes a lot of data points which may create an imbalance in the natural data set, this was mentioned in the King and Nielsen paper [3]. Another limitation of this analysis was the exclusion of any variables related to dates due to the difficulty in modeling time related variables. Dates are really important when it comes to epidemiology as there are new developments of treatment of diseases as time passes and that would impact the fatality of a patient greatly, even those with the exact same conditions would have a drastically different fatality rate if the context of time was different.

## Next Steps and Discussion

The immediate next step of this analysis would be to use outbreak as the "treatment" variable and do the exact same analysis to see whether or not "outbreak" was the cause of the significantly higher fatality rate of non-travel cases. Other acquisitions could also be used as the "treatment" to see its effect on the fatality rate. Future reports should also consider time as variables by implementing more complicated statistical models.

# Bibliography

1. Auguie, Baptiste. “Grid Extra.” Arranging multiple grobs on a page, September 9, 2017. https://cran.r-project.org/web/packages/gridExtra/vignettes/arrangeGrob.html.

2. Gelman, Andrew, et al. “Data Analysis Using Regression and Multilevel/Hierarchical Models [R Package Arm Version 1.12-2].” The Comprehensive R Archive Network. Comprehensive R Archive Network (CRAN), October 15, 2021. https://cran.r-project.org/web/packages/arm/index.html.

3. Noah. “Propensity Score Matching - What Is the Problem?” Stacks Exchange. Stack Exchange Inc., August 5, 2020. https://stats.stackexchange.com/questions/481110/propensity-score-matching-what-is-the-problem.

4. “Propensity Score Matching.” DIME Wiki. The World Bank Group. Accessed December 20, 2021. https://dimewiki.worldbank.org/Propensity_Score_Matching.

5. Public Health Ontario. “Confirmed Positive Cases of COVID-19 in Ontario.” Confirmed positive cases of COVID-19 in Ontario . Public Health Ontario, December 17, 2021. https://data.ontario.ca/en/dataset/confirmed-positive-cases-of-covid-19-in-ontario.

6. Public Health Ontario. “DAILY EPIDEMIOLOGICAL SUMMARY of Covid-19 in Ontario: January 15, 2020 to December 18, 2021.” Ontraio.ca. Public Health Ontario, December 18, 2021. https://files.ontario.ca/moh-covid-19-report-en-2021-12-19.pdf.

7. Robinson, David, et al. “Convert Statistical Objects into Tidy Tibbles [R Package Broom Version 0.7.10].” The Comprehensive R Archive Network. Comprehensive R Archive Network (CRAN), October 31, 2021. https://cran.r-project.org/web/packages/broom/index.html.

8. Wickham, H., et al. Tidyverse. Accessed December 20, 2021. https://tidyverse.org/.

9. Xie, Yihui. “A General-Purpose Package for Dynamic Report Generation in R [R Package Knitr Version 1.37].” The Comprehensive R Archive Network. Comprehensive R Archive Network (CRAN), December 16, 2021. https://cran.r-project.org/web/packages/knitr/index.html.

# Appendix

## A1: Ethics Statement

One ethical consideration that was made within this report was to keep the unspecified and diverse gender in the analysis even though they are only an extremely small portion of the data set (roughly 100 out of 29000 after propensity score matching). Keeping these two non-binary genders means that this analysis would be using gender instead of the binary sex which is a more modern and ethical way to do statistics. Another ethical consideration was using open data and properly citing it, health related data could be especially sensitive due to privacy reasons, this data set may be limited in its variables which caused some issues in generating a reliable propensity score but it was the best that was openly available from the government.

## A2: Materials

```
##   Age_Group  Client_Gender                    Reporting_PHU Travel_Related
## 1       <20 GENDER DIVERSE Windsor-Essex County Health Unit              0
## 2       90+         FEMALE         Timiskaming Health Unit              0
## 3       90+         FEMALE           Porcupine Health Unit              0
## 4       90+         FEMALE           Porcupine Health Unit              0
## 5       90+         FEMALE           Porcupine Health Unit              0
## 6       90+         FEMALE           Porcupine Health Unit              0
##   Outcome1      .fitted treated match.ind cnts pairs
## 1 Resolved 7.006916e-05       0         0    0    NA
## 2 Resolved 3.460940e-04       0         0    0    NA
## 3    Fatal 5.287361e-04       0         0    0    NA
## 4 Resolved 5.287361e-04       0         0    0    NA
## 5 Resolved 5.287361e-04       0         0    0    NA
## 6 Resolved 5.287361e-04       0         0    0    NA
```

This is a glimpse of the cleaned data used for analysis

All analysis for this report was programmed using `R version 4.0.4`.