Hospitalization Rates of Toronto COVID-19 Cases from January 2020 to February 2024*

Julian Marrero

March 16, 2024

This report accesses the hospitalization rates of Toronto COVID-19 cases from January 2020 to February 2024. A dataset, provided by Toronto Public Health, encompassing reported COVID-19 cases and associated hospitalizations over a period exceeding four years was analyzed. The study employed linear regression to evaluate the relationship between the progression of time (measured as days since the start of the dataset) and daily hospitalization rates. The linear regression analysis yielded an intercept indicative of the baseline hospitalization rate and a slight, non-significant slope suggesting a marginal increase in hospitalization rates over time. The low R-squared value pointed to the model's limited explanatory power, with time accounting for a small fraction of the variability in hospitalization rates

1 Introduction

The COVID-19 pandemic emerged in late 2019 in Wuhan, China, and quickly spread globally. Characterized by respiratory symptoms, it has led to significant health, economic, and social impacts worldwide (World Health Organization 2020). Governments and health organizations have responded with public health measures, with majority of nations going into a form of lockdown during the pandemic. Toronto was no different, having a lockdown that spanned for almost 2 years ("COVID-19 pandemic in Toronto" 2020). Despite the governments best efforts to reduce the spread of the virus, thousands of individuals were inflicted by COVID-19.

This paper aims to analyze the trends in hospitalization rate of Toronto COVID-19 cases from the start of the pandemic (January 2020) to the present (February 2024). This was accomplished by employing a linear regression model. This allows us to understand the trend and impact of the pandemic over time. This analysis can reveal the effectiveness of public

*Code for R scripts and data for this analysis are available at: https://github.com/julianmarrero/Hospitalization-Rates-of-COVID-19-Cases

health interventions, identify periods of increased strain on healthcare resources, and guide future policy decisions. It's crucial for planning, resource allocation, and evaluating the need for and timing of interventions like lockdowns and/or vaccination campaigns.

The linear regression analysis was conducted using the statistical programminig language R (R Core Team 2020). The following packages were used in the analysis: tidyverse (Wickham et al. 2019), lubridate (Grolemund and Wickham 2011), dplyr (Wickham et al. 2023), ggplot2 (Wickham 2016), and readr (Wickham, Hester, and Bryan 2024).

The paper begins with a discussion on the source of the data used, the methodologies employed, and a review of the variables used for the data reading, processing, analysis, and visualization. Following this, the linear model used in the analysis is explained. Next, the results from the analysis is discussed including summary statistics and a dotplot. The paper conlcudes with a discussion of the findings, the limitations of the paper, and future research that can be conducted.

2 Data

2.1 Source

The dataset used was accessed through Open Data Toronto and was provided by Toronto Public Health (TPH). The TPH shared anonymized information for all COVID-19 cases reported from the start of the COVID-19 pandemic in January 2020. The amount of information accessible for cases reduced as case and outbreak management recommendations changed and COVID-19 specialized resources were no longer supported (Toronto Public Health 2023). As a result, more current data are incomplete and not comparable to earlier years. As of February 14, 2024, TPH ceased producing this report, with the last refresh occurring on that date (Toronto Public Health 2023).

2.2 Methodology

The R script in 02-data_analysis.R focuses on analyzing the trend in hospitalization rates among COVID-19 cases over time, using a linear regression model to understand the relationship between the passage of time and the likelihood of hospitalization. Data cleaning steps include converting categorical and date data into formats suitable for analysis (binary and numeric variables, respectively). The visualization step helps in interpreting the model's findings, illustrating how hospitalization rates have evolved throughout the dataset's timeframe.

2.3 Data Reading and Processing

- 'covid_cases' is the dataframe created by reading a CSV file that contains data on COVID-19 cases.
  - 'Ever_Hospitalized_Binary' is a constructed binary variable within 'covid_cases' indicating whether a case was ever hospitalized (1 for Yes, 0 for No). This is derived from the 'Ever Hospitalized' column, converting the categorical Yes/No responses into a binary format for easier analysis.
  - 'Episode_Date' is another modified column within 'covid_cases' where the 'Episode Date' column is converted from a string or other format into a Date object. This allows for date-based operations and calculations.

2.4 Analysis and Visualization

- 'daily_hospitalization_rate' is a dataframe resulting from aggregating the 'covid_cases' data by 'Episode_Date'. It calculates the mean of 'Ever_Hospitalized_Binary' for each date, effectively giving the daily hospitalization rate. The .groups = 'drop' argument ensures the result is no longer grouped by any variable, simplifying further operations.
  - 'Hospitalized_Rate' is a variable within 'daily_hospitalization_rate' representing the calculated daily hospitalization rate.
  - 'Days_Since_Start' is a numeric variable added to 'daily_hospitalization_rate', representing the number of days since the first recorded episode date in the dataset. This is used as a predictor variable in the linear regression model.
- 'hospitalization_model' is a linear regression model that predicts the 'Hospitalized_Rate' based on 'Days_Since_Start'. This model aims to understand how the hospitalization rate changes over time.
- 'summary(hospitalization_model)' is the command that provides a summary of the linear regression model, including coefficients, R-squared value, and other statistical measures to assess the fit of the model.
- Plotting commands (using ggplot2): These commands create a scatter plot of the daily hospitalization rate over time, with points representing observed rates and a red line representing the linear regression model's prediction. The plot is titled "Daily COVID-19 Hospitalization Rate Over Time" and includes labels for both axes and a minimalistic theme for clarity.

3 Model

3.1 Model Description

The linear regression model developed aimed to explore the relationship between the progression of time since the onset of data recording ("Days Since Start") and the rate of hospitalization due to COVID-19 ("Hospitalized Rate") within a specific dataset. This dataset presumably contains daily counts of COVID-19 cases, along with whether each case resulted in hospitalization, among other pieces of information. The analysis sought to understand if there was a statistically significant trend in hospitalization rates over the time frame covered by the dataset.

The model formula can be expressed as:

Hospitalized Rate $= \beta_0 + \beta_1 \times$ Days Since Start $+\epsilon$

where:

- β_0 is the intercept, representing the estimated hospitalization rate at the start of the dataset (Day 0).
- β_1 is the coefficient for the"Days Since Start" variable, indicating the change in the hospitalization rate for each additional day since the start of the dataset.
- ϵ represents the error term, accounting for the variation in hospitalization rate not explained by the passage of time.

3.2 Model Justification

The choice to use linear regression stems from the initial hypothesis that the progression of the pandemic over time might influence the rate at which cases result in hospitalization, potentially due to factors like changes in virus virulence, healthcare system capacity, or public health interventions. Linear regression provides a straightforward method for quantifying and testing this relationship.

3.3 Model Interpretation

- Intercept($\beta_0 = 0.06267$): The model estimates that the hospitalization rate at the beginning of the dataset is approximately 6.27%. This value is statistically significant, suggesting confidence in the model's estimation of the baseline hospitalization rate.

- Slope($\beta_1 = 0.000007177$): The coefficient for "Days Since Start" suggests a marginal increase in the hospitalization rate by approximately 0.0007177% for each subsequent day. However, the p-value associated with this coefficient (0.0583) slightly exceeds the traditional threshold for statistical significance (0.05), suggesting caution in interpreting this as a definitive upward trend.

4 Results

Summary statistics of the linear model was created using the following R script:

```
# Aggregate data to calculate the daily hospitalization rate
daily_hospitalization_rate <- covid_cases %>%
  group_by(Episode_Date) %>%
  summarise(Hospitalized_Rate = mean(Ever_Hospitalized_Binary, na.rm = TRUE),
            .groups = 'drop')

# Convert 'Episode_Date' to a numeric variable (days since start)
daily_hospitalization_rate$Days_Since_Start <- as.numeric(daily_hospitalization_rate$Episode_

# Linear regression model
hospitalization_model <- lm(Hospitalized_Rate ~ Days_Since_Start, data=daily_hospitalization_

# Summary of the model
summary(hospitalization_model)
```

From this code, the following results were printed:

From Figure 1, the estimated hospitalization rate at the start of the dataset (Day 0) is approximately 6.27%. This value represents the model's prediction for the hospitalization rate before any days have passed since the start of the dataset. The p-value for the intercept is less than 0.05, indicating that the intercept is significantly different from 0 at the 5% significance level.

The coefficient for **Days_Since_Start** is approximately 0.000007177, suggesting a very slight increase in the hospitalization rate each day. However, the p-value for this coefficient is 0.0583, which is slightly above the conventional 0.05 cutoff for statistical significance. This means that while there is a trend of increasing hospitalization rate over time, we cannot confidently say it is statistically significant at the 5% level (though it might be considered significant at the 10% level, as indicated by the '.' sign).

The residuals show how far off the model's predictions are from the actual data. The fact that the residuals range from -0.07332 to 0.93733 indicates a wide variance in the model's predictive

```
Residuals:
     Min       1Q   Median       3Q      Max
-0.07332 -0.03389 -0.01252  0.01782  0.93733

Coefficients:
                  Estimate Std. Error t value Pr(>|t|)
(Intercept)      6.267e-02  3.274e-03  19.141   <2e-16 ***
Days_Since_Start 7.177e-06  3.787e-06   1.895   0.0583 .
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06095 on 1457 degrees of freedom
Multiple R-squared:  0.002459,  Adjusted R-squared:  0.001774
F-statistic: 3.591 on 1 and 1457 DF,  p-value: 0.05829
```

Figure 1: Figure 1: Summary Statistics of Regression Model

accuracy. The large maximum value suggests there are outliers or instances where the model's prediction was significantly different from the actual hospitalization rate.

On average, the model's predictions deviate from the actual hospitalization rates by approximately 6.095%. This gives us an idea of the model's predictive accuracy.

As we have the multiple R-squared at 0.002459, this value indicates that only about 0.25% of the variance in the hospitalization rate is explained by the number of days since the start of the dataset. This is a very low value, suggesting that the model does not explain much of the variation in hospitalization rates.

The adjusted R-squared is 0.001774. This is a slight adjustment to the R-squared value that accounts for the number of predictors in the model (in this case, just one). It provides a more accurate measure of the goodness of fit for models with multiple predictors. The low value here further suggests that Days_Since_Start alone does not provide a good explanation for changes in the hospitalization rate.

The F-statistic is 3.591. This tests the overall significance of the model. With a p-value of 0.05829, it suggests that the model is not statistically significant at the 5% level but might be at the 10% level. This aligns with the p-value for the Days_Since_Start coefficient and indicates that the model might not be a good fit for the data.

The following R code was used to generate the dotplot displayed below:

```r
ggplot(daily_hospitalization_rate, aes(x = Days_Since_Start, y = Hospitalized_Rate)) +
  geom_point() +
  geom_smooth(method = "lm", color = "red") +
  labs(title = "Daily COVID-19 Hospitalization Rate Over Time",
       x = "Days Since Start of Dataset",
       y = "Hospitalization Rate") +
  theme_minimal()
```

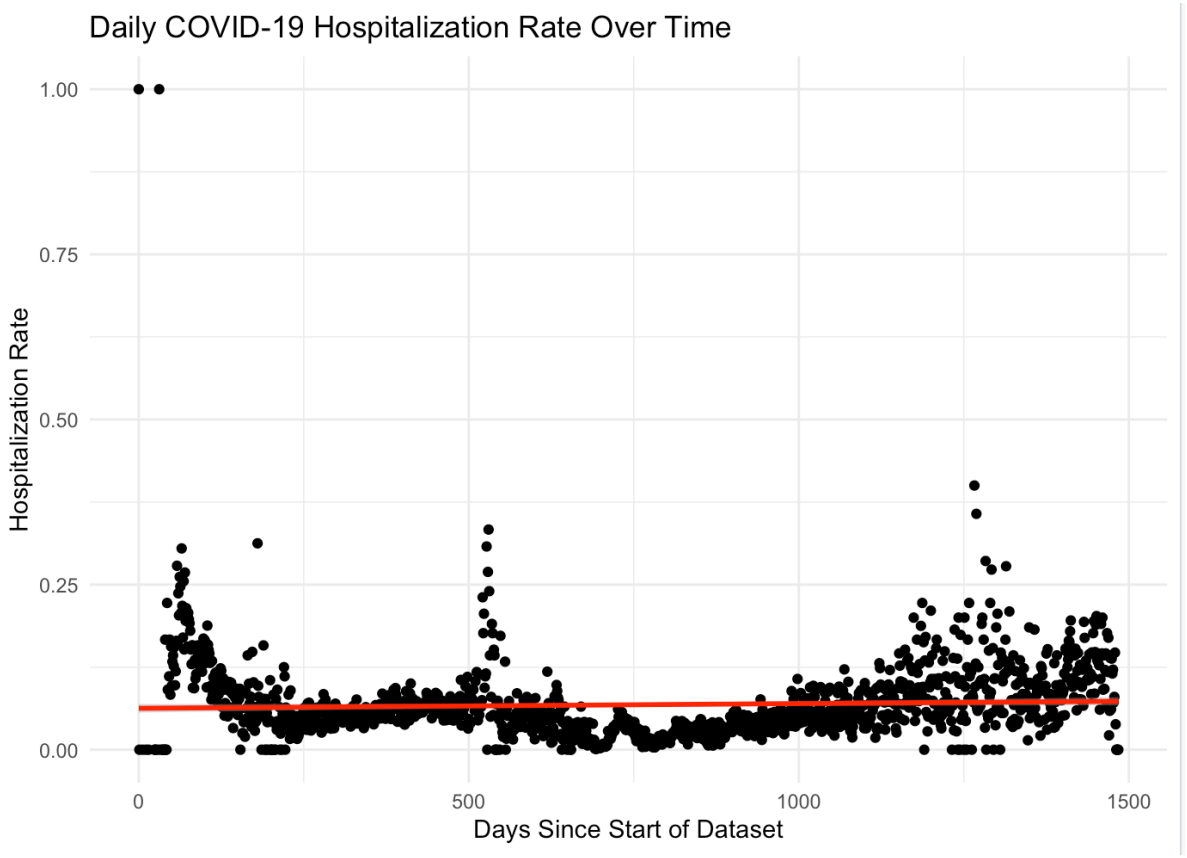

Figure 2: Figure 2: Hospitalization Rate of Toronto COVID-19 Cases From 2020-2024

Figure 2 displays the daily COVID-19 hospitalization rate over time, where each dot represents the hospitalization rate for a particular day as a function of the number of days since the start of the dataset. The x-axis, labeled "Days Since Start of Dataset," shows the timeline of the dataset from day 0 to beyond day 1500, suggesting the data spans over four years if each increment represents a day. The y-axis, labeled "Hospitalization Rate," represents the proportion of hospitalized cases and appears to range from 0 to 1, which is typical for rates expressed as a fraction or percentage.

The bulk of the data points are clustered near the lower end of the hospitalization rate axis, suggesting a generally low rate of hospitalization relative to the number of COVID-19 cases recorded. A linear regression line (in red) is superimposed on the dot plot, showing a relatively flat trajectory. This indicates that the linear model found only a slight, if any, upward trend in the hospitalization rates over the duration of the dataset.

There are noticeable vertical clusters of dots at different points along the x-axis, showing days with higher variability in hospitalization rates. This could suggest occasional spikes in hospitalization rates, perhaps due to outbreaks, changes in public health policy, or other factors not directly captured by the time variable. Several dots appear well above the main cluster, indicating days with unusually high hospitalization rates. These outliers may represent specific events or data reporting anomalies and could have a significant influence on the regression model's fit and interpretation.

5 Discussion

5.1 Findings

While there is a slight trend indicating an increase in hospitalization rate over time, Figure 1 suggests that the number of days since the start of the dataset is not a strong predictor of hospitalization rate. The low R-squared values indicate that other variables not included in the model may be influencing the hospitalization rate more significantly. Additionally, considering the potential for outliers or influential points suggested by the residuals, further investigation into the data's distribution and potential data quality issues might be warranted.

The regression analysis suggests that the severity of cases requiring hospitalization has remained relatively stable over time. This stability may reflect the success of medical interventions and the healthcare system's ability to adapt to the evolving pandemic. Improved treatment protocols and early interventions could be keeping the need for hospitalization in check, even as the virus continues to spread. This insight highlights the importance of medical research and healthcare preparedness in managing the severity of infectious diseases.

The pandemic is influenced by a myriad of factors, including biological (virus mutations), epidemiological (patterns of transmission), social (public health measures), and healthcare system capacities. The analysis emphasizes that no single variable, such as time, can fully capture the dynamics at play. It underscores the need for comprehensive data collection and analysis that includes various determinants of health and disease spread, allowing for nuanced understanding and response strategies.

The slight trend in hospitalization rates may reflect the cumulative impact of public health interventions, from lockdowns and social distancing to mask mandates and vaccination campaigns. These measures, aimed at controlling the spread of the virus and protecting the most vulnerable, likely contributed to preventing a significant increase in hospitalization rates. This

reinforces the value of evidence-based public health policies and community compliance in mitigating the pandemic's impact.

The analysis indirectly shines a light on the importance of equitable healthcare access. Stable hospitalization rates may indicate that individuals in the studied population have relatively equal access to care, preventing severe outcomes among marginalized groups. This points to the broader issue of health equity, highlighting the need for policies and systems that ensure all individuals, regardless of socio-economic status, have access to the care and resources needed to combat COVID-19 effectively.

The ability of healthcare systems to maintain stable hospitalization rates amidst the challenges of a pandemic is a testament to their resilience and adaptability. This includes scaling up hospital capacities, optimizing patient triage and care processes, and leveraging telehealth. It reflects the critical role of healthcare planning, investment in healthcare infrastructure, and the dedication of healthcare workers. Ensuring the continuous adaptability of healthcare systems is essential for addressing current and future public health crises.

5.2 Limitations

The model's simplicity, focusing solely on the relationship between time (as days since the start of the dataset) and hospitalization rates, overlooks numerous other factors that influence hospitalization. These include virus variants, population immunity levels, changes in public health policies, and individual behaviors. A more complex model that incorporates these additional variables could provide a fuller understanding of the factors driving hospitalization rates.

The linear regression model assumes a linear relationship between time and hospitalization rates. However, the progression of a pandemic and its impact on hospitalizations can be highly non-linear, influenced by complex dynamics such as policy changes, public compliance with guidelines, and the emergence of new virus variants. The model does not capture these potential non-linearities and their impact on hospitalization trends.

The analysis uses aggregated data (daily or weekly rates) to assess the trend in hospitalizations. This aggregation can mask important variations and patterns within the data, such as differences across age groups, geographical areas, or socio-economic statuses. More granular analyses could reveal disparities in hospitalization rates and the effectiveness of interventions across different segments of the population.

The model implicitly assumes that time is a proxy for various underlying factors affecting hospitalization rates. However, this approach does not establish causal relationships between specific interventions or events and changes in hospitalization rates. Identifying causality would require more detailed data and analytical techniques, such as time-series analysis or causal inference methods.

The analysis is based on data from Toronto and may not be generalizable to other contexts or populations. Factors influencing hospitalization rates can vary significantly by location due to differences in healthcare systems, population demographics, and public health responses. Findings from this analysis should be applied to other settings with caution.

The analysis depends on the quality and completeness of the data available. Issues such as underreporting of cases, changes in testing rates, or delays in data reporting can affect the accuracy of the hospitalization rates calculated. Additionally, the availability of detailed data on other potentially relevant factors (e.g., vaccination status, compliance with public health measures) is crucial for more comprehensive analyses.

The slight exceedance of the p-value above the conventional threshold for statistical significance (0.05) and the low R-squared value indicate that the model does not strongly predict hospitalization rates based on time alone. This suggests that other unmodeled factors play a significant role in determining hospitalization rates, and the model's explanatory power is limited.

5.3 Future Research

Future studies could include more variables that are likely to influence hospitalization rates, such as:

- Vaccination status: Including data on vaccination rates over time to assess their impact on hospitalizations.
- Public health interventions: Analyzing the effect of specific interventions (lockdowns, mask mandates) on hospitalization rates.
- Virus variants: Considering the emergence of more contagious or virulent variants and their impact on hospitalization trends.
- Socio-economic factors: Exploring how factors like income, housing, and occupation affect the risk of hospitalization.

Additionally, exploring non-linear models and time series analysis techniques could capture the complex dynamics of the pandemic more accurately. Techniques such as ARIMA (AutoRegressive Integrated Moving Average) models, seasonal decomposition, or exponential smoothing models might reveal patterns and trends not captured by simple linear regression.

References

"COVID-19 pandemic in Toronto." 2020. https://en.wikipedia.org/wiki/COVID-19_pandemic_in_Toronto.

Grolemund, Garrett, and Hadley Wickham. 2011. "Dates and Times Made Easy with lubridate." *Journal of Statistical Software* 40 (3): 1–25. https://www.jstatsoft.org/v40/i03/.

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

Toronto Public Health. 2023. "COVID-19 Cases in Toronto." https://open.toronto.ca/dataset/covid-19-cases-in-toronto/.

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2019. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

World Health Organization. 2020. "Coronavirus Disease (COVID-19) Pandemic." https://www.who.int/emergencies/diseases/novel-coronavirus-2019.