Outbreaks in Toronto: Examining Disease Outbreaks Trends Between 2020 to 2024*

An exploration of the location and frequency of disease outbreaks in Toronto between the years 2020 and 2024

Sindhu Priya Mallavarapu

April 18, 2024

This paper analyzes the disease outbreak patterns across the city of Toronto in the years between 2020 - 2024 based on location and models future outbreak trends. It is found that the highest number of disease outbreaks within these five years happened Long-Term Care Facilities. The results of this paper can help government officials see where more healthcare funding and stricter safety practices are needed.

1 Introduction

The pandemic has put a clear spotlight on all the cracks in the medical system. As Toronto is the most populous city in Canada, there is particular drain on this city's medical system. Thus, identifying where there are a higher number of outbreaks can show where the funding needs to be directed to in order to improve the medical system. This paper uses disease outbreak data form Open Data Toronto (Gelfand 2022) to show where the highest number of outbreaks took place in Toronto and model outbreak trends given the type of healthcare facility (setting) and year. The estimand in this paper is the average disease outbreaks in each setting and year.

It is found that no matter what year between 2020 to 2024 was examined, the most outbreaks occurred in Long-Term Care Facilities (LTCF) and the least amount of outbreaks happened in Transitional Care. Moreover, between the examined years, 2022 was the year with the highest amount of total disease outbreaks. These findings are important as they can help government officials see where there is a need for more funding and health care professionals.

The paper is structured in the following way. The Data section, Section 2, goes into detail about the data that was used in this paper, including details about its measurement. The Model

[*] Code and data are available at: https://github.com/MSindhuPriya/toronto_outbreaks.

section, Section 3, talks about the Poisson distribution model used in the paper. The Results section, Section 4, shows what was found from analyzing the data, that is, the highest number of outbreaks occurred in Long-Term Care Facilities and in the year 2022. Further discussion of the results, limitations, and future research pathways are detailed in the Discussion section, Section 5. The Appendix provides additional details about the model that were not included in the main body of the paper.

2 Data

2.1 Overview

The disease outbreak data that was used in this paper is sourced from Open Data Toronto (Gelfand 2022) and maintained by the Communicable Disease Surveillance Unit. The analysis in this paper is done through R (R Core Team 2022) along with other packages that are compatible with R, such as, `tidyverse` (Wickham et al. 2019), `marginaleffects` (Arel-Bundock 2024), `arrow` (Richardson et al. 2024), `rstanarm` (Goodrich et al. 2022),`ggplot2` (Wickham 2016), `knitr` (Y 2023), `lintr` (Hester et al. 2024), and `dplyr` (Wickham et al. 2023).

Other than the data that was used here, similar datasets were available though the province of Ontario. However, this paper is only concerned with outbreak patterns in the city of Toronto and the datasets available through the province of Ontario did not have any data solely specific to this city. Therefore, we chose to only use the disease outbreak dataset available through Open Data Toronto so that we had data that was relevant to the region of study.

2.2 Measurement

Several decisions were made with regards to the measurement of outbreaks in the city of Toronto. First, only outbreaks reported in healthcare facilities were measured. This was done to ensure reliability of data as outbreaks reported by individuals in their residential areas are not reliable. Moreover, only data between the years 2020 and 2024 (inclusive) were considered. During data cleaning we added in variables that counted the number of outbreaks to happen in that setting based on the year. As some outbreaks spanned over multiple years, the year of the outbreak was determined by the year the outbreak began rather than the year it ended.

2.3 Description of Variables

A brief look at the relevant variables in the dataset is seen in Figure 1. In this paper, we focus on three main variables - `outbreak_setting`, `setting_count`, and `outbreak_year`. The type of healthcare facility that the outbreak happened in is represented as a string in

Outbreak Setting	# of Outbreaks	Outbreak Year
LTCH	477	2022
LTCH	601	2023
LTCH	187	2021
LTCH	169	2024
Hospital-Chronic Care	49	2020
Hospital-Psychiatric	8	2024
Retirement Home	205	2022
Retirement Home	55	2021
Hospital-Acute Care	82	2020
LTCH	316	2020
Hospital-Chronic Care	47	2021
Hospital-Acute Care	107	2023
Retirement Home	195	2023
Retirement Home	97	2020
Retirement Home	35	2024
Hospital-Acute Care	102	2021
Hospital-Chronic Care	177	2022
Hospital-Chronic Care	33	2024
Hospital-Acute Care	248	2022
Hospital-Acute Care	27	2024
Hospital-Chronic Care	131	2023
Hospital-Psychiatric	19	2023
Hospital-Psychiatric	17	2022
Transitional Care	9	2022
Hospital-Psychiatric	3	2021
Hospital-Psychiatric	9	2020
Transitional Care	13	2023

Figure 1: Frequncy of outbreaks grouped by setting and year

`outbreak_setting`. The year that the outbreak began in (a year between 2020 to 2024, inclusive) is represented as an integer in `outbreak_year`. The number of outbreaks at that setting in that particular year is given in `setting_count`. Figure 1 shows the 27 occurrences of `setting_count`, that is, the 27 different counts of how man outbreaks happened in each setting in each of the observed years.

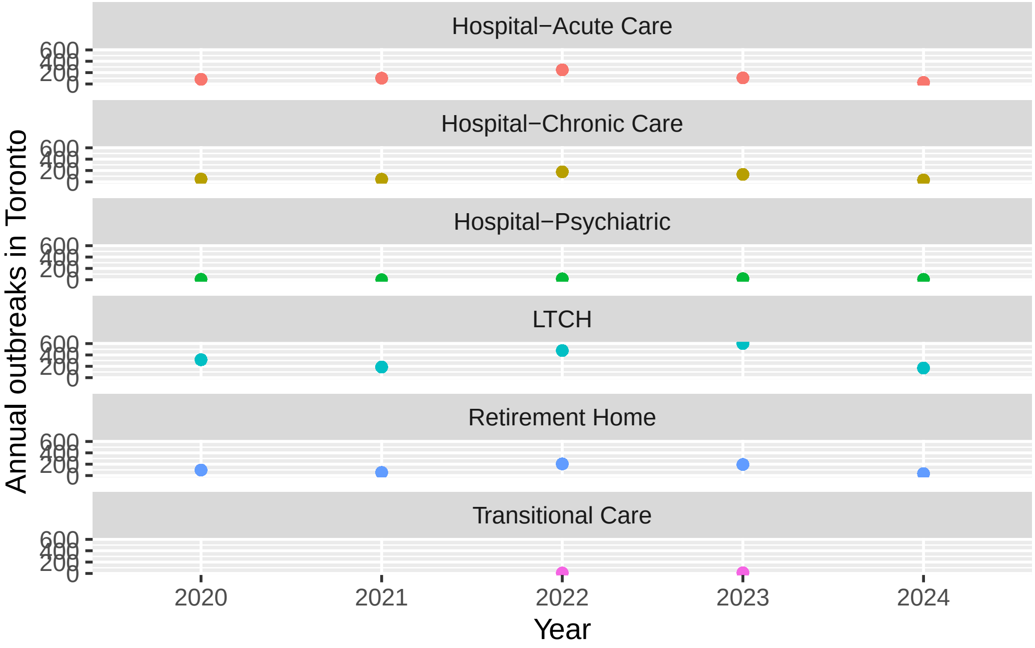


Figure 2: Outbreaks per year grouped by setting

3 Model

The goal of our modelling strategy is to use Poisson distribution to model the number of disease outbreaks in Toronto based on type of healthcare facility and the outbreak year. We use Poisson distribution as this is commonly used when given the average rate of occurrence of the event to model the number of those events occurring in a fixed interval of time or space.

In this study, we treat each setting and year as a separate interval of time and space. The number of occurrences in each of those intervals follows a Poisson distribution. This distribution is characterized by a single parameter, , which in this case represents the average number of outbreaks in each setting and year interval.

3.1 Assumptions

Poisson distributions assume that events occur independently and at a constant average rate. That is, in our study, this assumption implies that occurrence of an outbreak of a particular disease in one year or setting does not imply that an outbreak of a different disease in another setting or year will occur.

3.2 Model set-up

Define y_i as the number of disease outbreaks given a setting or year. To implement our Poisson model, we first calculated which is the average disease outbreaks given a year and a type of healthcare facility. Then we used that as our parameter in our Poisson distribution to model the number of outbreaks in Toronto given the year and setting.

$$\begin{aligned}
y_i|\lambda &\sim \text{Poisson}(\lambda_i) && (1)\\
log(\lambda_i) &= \beta_0 + \beta_1 \times \text{Outbreak Setting}_i + \beta_2 \times \text{ Outbreak Year}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)
\end{aligned}$$

We run the model in R (R Core Team 2022) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`.

3.2.1 Model justification

We expect an upside down parabola as the relationship between the outbreaks given the setting and year. This is because we expect an increase in outbreaks due to the COVID-19 pandemic and then a decrease afterwards. We find that our model accurately predicted the number of disease outbreaks based on setting and year. Therefore, we can say that the Poisson distribution accurately captures the frequency of outbreaks in Toronto if given the healthcare facility and year. This aids us in being able to see which facilities need additional funding and healthcare workers.

4 Results

As seen in Figure 2 we found that the highest number of occurred in Long-Term Care Facilities in the year 2022. Overall, the year 2022 had the highest number of outbreaks and Long-Term

Care Facilities were the healthcare facilities with the highest number of outbreaks. Moreover, it is found that Transitional Care facilities had the least amount of outbreaks.

4.1 Poisson Distribution Analysis

We found that our Poisson distribution can closely model the number of outbreaks given year and type of healthcare facility. This model acts according to our expectation that that there is a bell shaped distribution between frequency of outbreaks given setting and year. Our results are summarized in Table 2 where we found that the intercept for our model is 4.73. Further, the accuracy checks for our models are given in Section .1 and additional diagnostic details are in Section .2.

5 Discussion

In this paper, we found the number of outbreaks that occurred in a given setting and year. Our goal was to investigate the relationship between the frequency of the outbreaks and the year and type of healthcare facility that the outbreaks occurred in. This can provide us with valuable insights on where to increase medical funding and where there is an additional need for healthcare workers. This can be a step towards having a more stable medical system.

Our analysis revealed that the highest number of outbreaks occurred during the year 2022 and Long-Term Care Facilities (LTCF). We also found that overall Long-Term Care Facilities consistently had the highest outbreak rates regardless of year. Additionally, we found that the year 2022 was the year with the highest number of outbreaks regardless of setting which can be explained by the COVID-19 pandemic. Further, we were also able to model the number of outbreaks given the year and the type of healthcare facility using a Poisson distribution model.

5.1 Weaknesses

Our study only focused on the city of Toronto and did not look at the outbreak trends of Canada as a whole. Further, we also only looked at outbreaks that were reported in health care facilities. We did not map the location of those facilities, nor did we look at outbreaks in residential areas. Moreover, our study did not look at any outbreak trends before the year 2020.

Table 2: Explanatory model of the total disease outbreaks based on setting and year

	First model
(Intercept)	4.73
	(0.00)
outbreak_settingHospital-Chronic Care	−0.32
	(0.01)
outbreak_settingHospital-Psychiatric	−2.44
	(0.04)
outbreak_settingLTCH	0.97
	(0.00)
outbreak_settingRetirement Home	−0.07
	(0.00)
outbreak_settingTransitional Care	−2.91
	(0.06)
outbreak_year2021	−0.42
	(0.01)
outbreak_year2022	0.56
	(0.00)
outbreak_year2023	0.65
	(0.00)
outbreak_year2024	−0.64
	(0.01)
Num.Obs.	3418
Log.Lik.	−22 207.036
ELPD	−22 236.7
ELPD s.e.	346.4
LOOIC	44 473.4
LOOIC s.e.	692.8
WAIC	44 473.3
RMSE	35.02

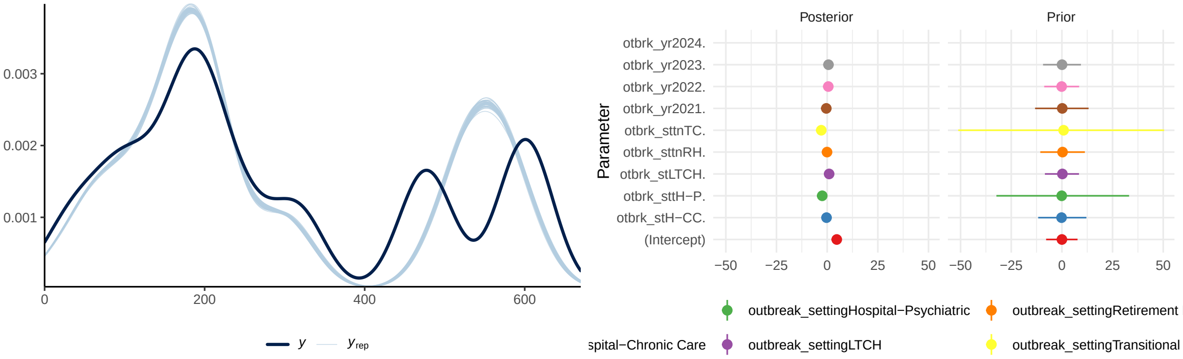
5.2 Future Research

Future research can include looking at outbreaks across Canada and not just Toronto. Along with that, we can include outbreak data from additional settings, not just healthcare facilities. Additionally, future studies can look at outbreak data from before 2020 and compare them to current data. Our study lays out a foundation of where we can direct funding to help aid the medical system, however, further research can be more specific as to what exactly can be done to ensure that the aid is effective.

Appendix

.1 Posterior predictive check

In Figure 3a we implement a posterior predictive check. This shows our model's ability to accurately predict the frequency of outbreaks given the setting and year. From Figure 3a we see that our model is close to the actual data. This means that our Poisson distribution model has the ability to accurately predict the number of outbreaks given the type of healthcare facility and the year.



(a) Posterior prediction check

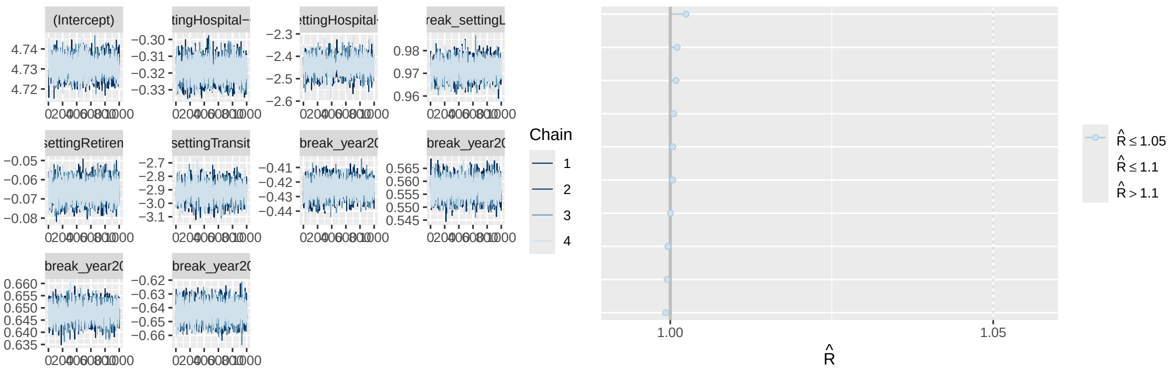
(b) Comparing the posterior with the prior

Figure 3: Examining how the model fits, and is affected by, the data

.2 Diagnostics

Figure 4a is a trace plot. Here, each line shows the sampled values of a parameter at each iteration of the MCMC algorithm which shows the convergence of the chain.

Figure 4b is a Rhat plot. This compares the variance within chains to the variance between chains. Our Rhat plot shows that they fall by 1 which indicates convergence.



(a) Trace plot

(b) Rhat plot

Figure 4: Checking the convergence of the MCMC algorithm

References

Arel-Bundock, Vincent. 2024. *Marginaleffects: Predictions, Comparisons, Slopes, Marginal Means, and Hypothesis Tests*. https://CRAN.R-project.org/package=marginaleffects.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Hester, Jim, Florent Angly, Russ Hyde, Michael Chirico, Kun Ren, Alexander Rosenstock, and Indrajeet Patil. 2024. *Lintr: A 'Linter' for r Code*. https://CRAN.R-project.org/package=lintr.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Richardson, Neal, Ian Cook, Nic Crane, Dewey Dunnington, Romain François, Jonathan Keane, Dragoș Moldovan-Grünfeld, Jeroen Ooms, Jacob Wujciak-Jens, and Apache Arrow. 2024. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

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://CRAN.R-project.org/package=dplyr.

Y, Xie. 2023. *Knitr a General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.