Analyzing the Factors Influencing Outbreak Duration in Toronto Healthcare Institutions*

My subtitle if needed

Kevin Cai

November 28, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

Outbreaks of infectious diseases in healthcare institutions are a significant challenge, particularly in hospitals, long-term care homes, and retirement homes, where vulnerable populations are most at risk. Understanding the factors that influence the duration of these outbreaks is important for improving preparedness, response times, and resource allocation. Previous studies have looked at factors affecting outbreak durations, but more research is needed, especially in Toronto's healthcare system. This paper analyzes the key factors that influence the duration of outbreaks in Toronto healthcare institutions, focusing on outbreak setting, causative agents, and the timing of outbreaks.

This study uses a Bayesian negative binomial regression model to predict outbreak durations in Toronto's healthcare facilities. The model includes factors like the type of healthcare institution (outbreak setting), the causative agent, and the month the outbreak occurred. This research aims to identify which factors contribute most to variation in outbreak durations. It provides actionable information for public health officials and healthcare managers to improve their response strategies.

The results of this analysis help manage outbreaks, optimize patient care, and reduce the strain on healthcare facilities. Identifying high-risk settings or times of year when outbreaks last longer can improve resource allocation and response planning. This study also adds to the literature on healthcare outbreaks, with a focus on Toronto's healthcare system.

[*] Code and data are available at: https://github.com/kevicai/toronto-healthcare-outbreak-prediction.

The remainder of this paper is structured as follows: Section 2 covers the dataset used for analysis, **?@sec-model** describes the model setup and methodology, **?@sec-results** presents the results and their interpretation, and **?@sec-discussion** discusses the findings and future research directions.

2 Data

2.1 Overview

This report uses the Outbreaks in Toronto Healthcare Institutions dataset, contains data from January 2016 to November 2024. The dataset is provided by Toronto Public Health, through City of Toronto Open Data Portal (Toronto Public Health 2024). The dataset tracks reported outbreaks of gastroenteric and respiratory illnesses in Toronto healthcare institutions and contains detailed information on outbreak settings, causative agents, and outbreak durations. Following the principles from Telling Stories with Data (Alexander 2024), we examine how the characteristics of outbreaks, such as the type of healthcare institution, the causative agent, and the month the outbreak began, influence their durations. A sample of the cleaned dataset is shown in Table 1.

Table 1: Sample of Cleaned Outbreaks in Toronto Healthcare Institution Data

Outbreak Setting	Causative Agent	Month	Outbreak Duration
LTCH	Influenza	Dec	20
Hospital-Acute Care	Norovirus	Dec	5
LTCH	Respiratory syncytial virus	Dec	14
LTCH	Metapneumovirus	Dec	21
Retirement Home	Influenza	Dec	21

There is 5387 observations in the orginal dataset and 1119 observations were removed that contained missing, invalid, or irrelivant data of the variables we're interested in. The data was first downloaded using `Python` (Van Rossum and Drake 2009) and cleaned with the `pandas` package (team 2020). The cleaning process involved converting dates to a standardized date-time format, creating a "duration" variable representing the length of each outbreak, and extracting the month of the outbreak's start. Irrelevant columns were removed, and variables were renamed for clarity. Causative agents were grouped into broader categories, and rows with missing or invalid data were removed, including those with unidentifiable causative agents or certain outbreak settings. The final dataset was saved for further analysis.

`R` (R Core Team 2023) is used for the generation of figures, graphs, and tables throughout this paper. Specifically, the `rstanarm` package (Goodrich et al. 2024) was employed to fit the model. For data manipulation, the `dplyr` package (Wickham et al. 2023) was utilized to clean and transform the data efficiently. The caret package (Kuhn and Max 2008) was used for model training, while modelsummary (Arel-Bundock 2022) was used to produce concise tables summarizing the model output. The loo package (Vehtari et al. 2024) was used to perform leave-one-out cross-validation, which helped assess the model's predictive performance. Finally, the package ggplot2 is used to generate graphics and figures for this analysis. The starter code and the data analysis techniques used are from Telling Stories with Data (Alexander (2024)).

2.2 Measurement

The data was primarily collected through mandatory reporting by healthcare institutions to Toronto Public Health under the Ontario Health Protection and Promotion Act (HPPA). Reports of suspected or confirmed outbreaks include both gastroenteric and respiratory illnesses. These reports are based on active monitoring by institutional staff, who observe and document signs and symptoms such as nausea, vomiting, fever, cough, or sore throat.

Some details, such as the causative agent group, may initially be unconfirmed and later identified through laboratory tests or clinical evaluations. However, these identifications are not always definitive. For instance, "Coronavirus*" in the dataset refers to seasonal coronaviruses, which are commonly implicated in respiratory outbreaks, and does not include COVID-19.

The unit of measurement for outbreak duration is in days. Other data fields, such as outbreak setting and causative agent group, are categorical features without numerical units. The dataset is updated weekly, ensuring it reflects the most recent outbreak data available.

2.3 Outcome variable

2.3.1 Duration

The Duration variable is numerical and indicates the total number of days each outbreak lasted. This reflects the severity and magnitude of the outbreak. It is constructed from the dataset by calculating the difference between the outbreak start and end dates.

Table 2: Summary of Outbreak Duration: Mean and Variance

Statistic	Value
Mean Duration	16.57873
Variance	110.89162

Longer outbreak durations may indicate challenges in containment, possibly influenced by the Outbreak Setting and Causative Agent.

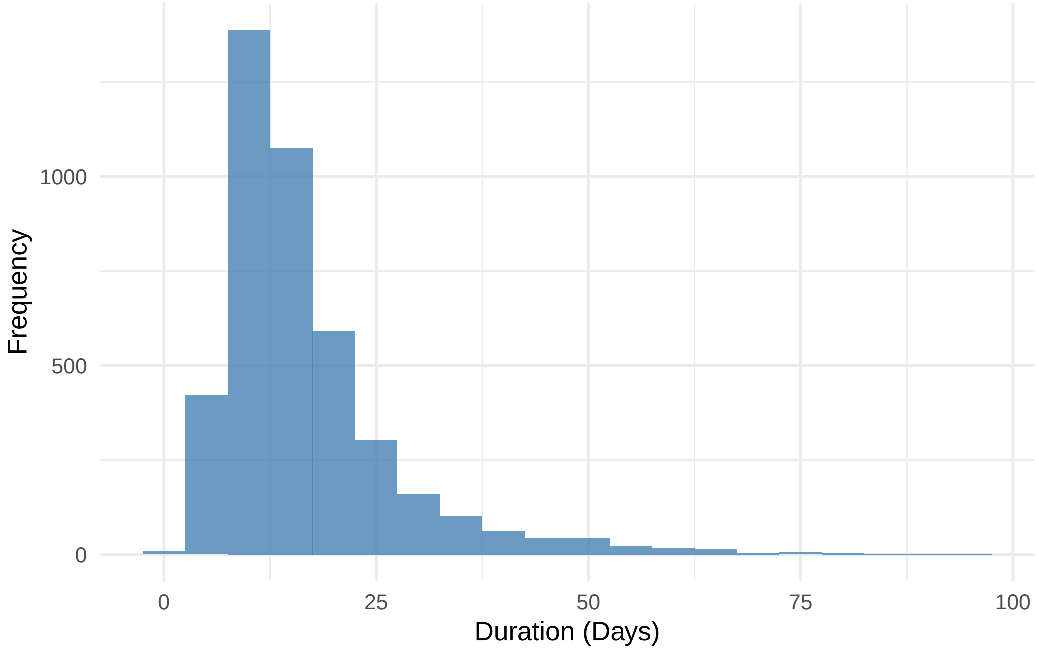


Figure 1: Distribution of Outbreak Duration

2.4 Predictor variables

2.4.1 Outbreak Setting

The Outbreak Setting variable is categorical and identifies the type of healthcare institution where the outbreak occurred, such as hospitals, long-term care homes (LTCH), or retirement homes. It provides insights into the environments most affected by outbreaks.

Figure 2 illustrates the count of outbreaks across different settings in the dataset.

LTCH (Long-Term Care Homes) accounts for a significant portion of outbreaks, likely due to the vulnerability of their populations. Comparing the frequency of outbreaks across settings can reveal risk patterns.

2.4.2 Causative Agent

The Causative Agent variable is categorical and reflects the infectious agents responsible for outbreaks. While the original dataset contains 55 agents, they are grouped into seven broader categories to simplify the analysis and enhance interpretability.

Figure 3 illustrates the count and percentage distribution of causative agents in the dataset.

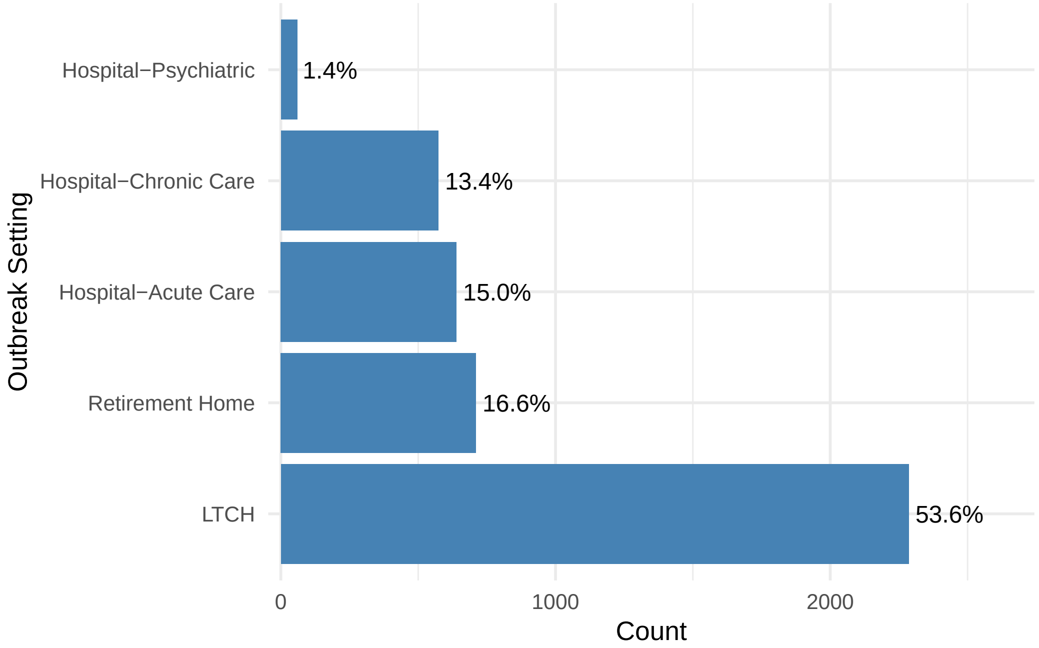


Figure 2: Outbreak occurrence in healthcare settings

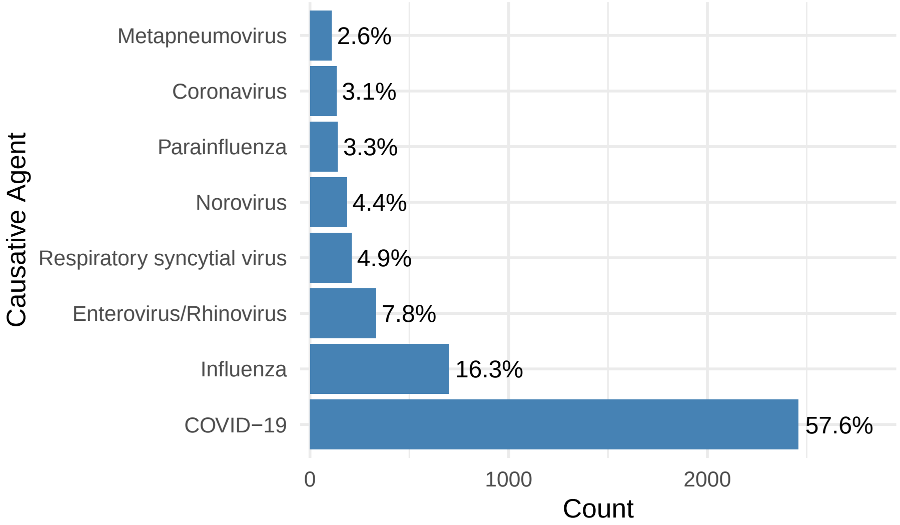


Figure 3: Outbreak causative agent count and percentage

2.4.3 Month Outbreak Began

The Month variable is categorical and records the calendar month when each outbreak started as a name (e.g., Jan, Feb). It reflects seasonal trends and potential patterns in infection rates. This variable is extracted from the date where each outbreak began from the original dataset and converted from a number to the corresponding month name.

Figure 4 shows the the occurance of outbreaks in each month, with winter months having siginifiantly more outbreaks compared to other months. This suggest that seasons have effects on outbreak occurances.

Figure 5 the boxplot visualizes the distribution of outbreak durations for each month. The duration of months January to November outbreaks appears similar, while December has a noticeable increase in duration compared to other months.

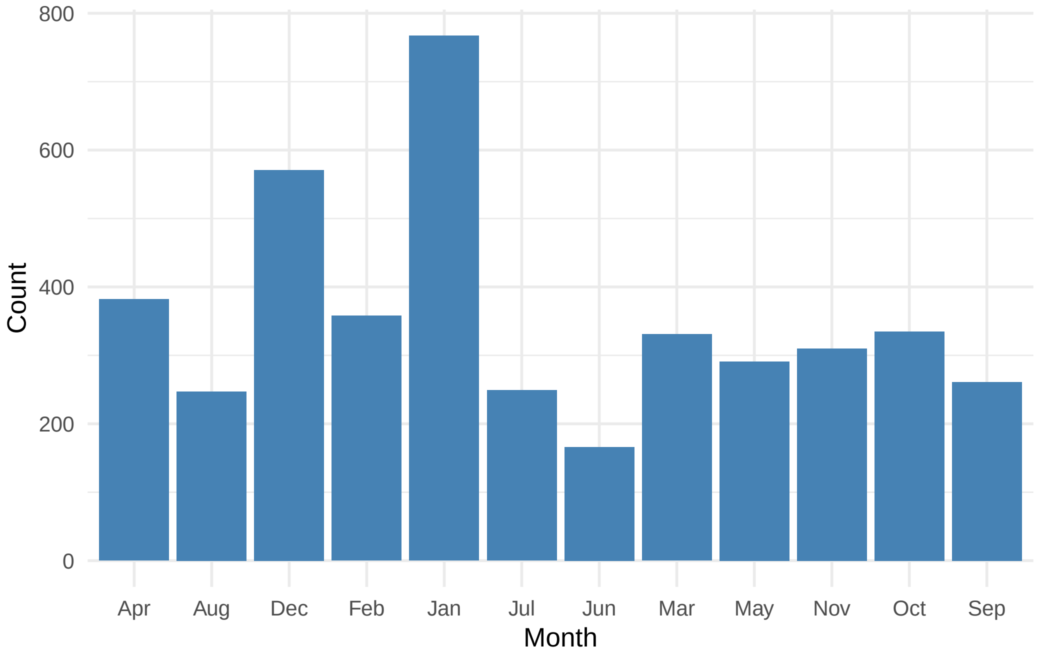


Figure 4: Seasonal trends in outbreak occurrence and percentage

3 Model

3.1 Model Overview

To better understand the factors influencing the duration of outbreaks in Toronto healthcare facilities, a statistical model was developed using the negative binomial regression framework.

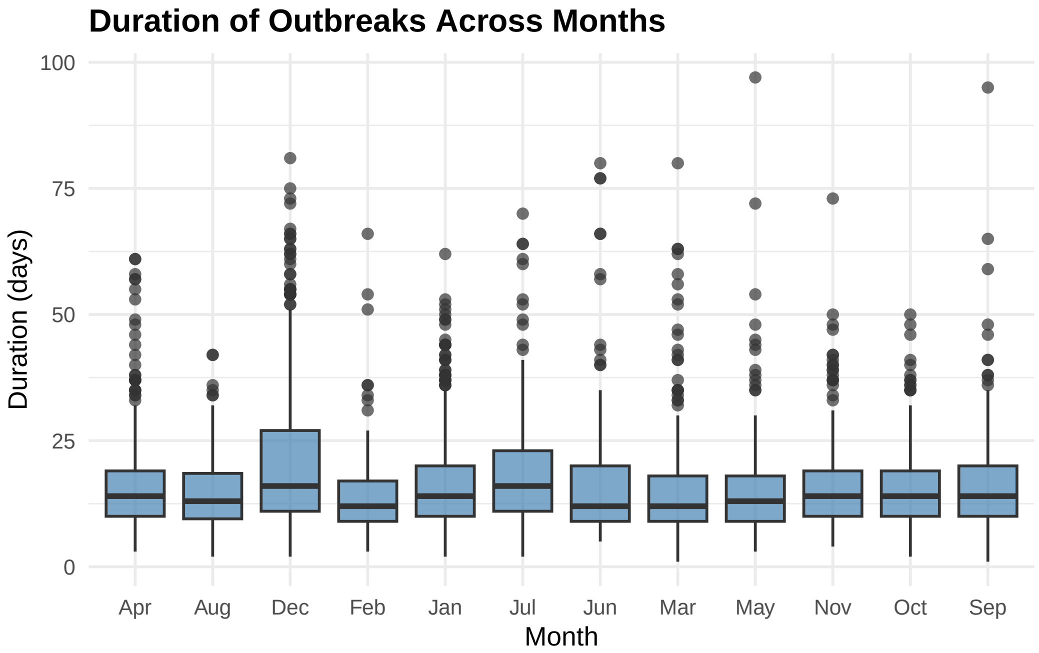


Figure 5: Duration of outbreaks across different months

This model was chosen because the outcome variable of interest, outbreak duration, is a count variable with evidence of overdispersion—where the variance exceeds the mean (Alexander 2024). Additionally, this model was Bayesian, meaning the parameters were treated as random variables with prior probability distributions reflecting initial beliefs about their values before considering the data.

3.2 Model Setup

The setup for the Bayesian negative binomial regression model used in this analysis is as follows:

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Negative Binomial}(\lambda_i, \phi) && (1)\\
\log(\lambda_i) &= \beta_0 + \beta_1 \times \text{outbreak\_setting}_i + \beta_2 \times \text{causative\_agent}_i + \beta_3 \times \text{month}_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
\beta_2 &\sim \text{Normal}(0, 2.5) && (5)\\
\beta_3 &\sim \text{Normal}(0, 2.5) && (6)\\
\phi &\sim \text{Exponential}(1) && (7)
\end{aligned}
$$

In the above model:

- λ_i is the expected duration of outbreak i, modeled through a log link.
- β_0 is the intercept term.
- β_1 is the coefficient for the **outbreak setting**.
- β_2 is the coefficient for the **causative agent**.
- β_3 is the coefficient for the **month** when the outbreak started.
- ϕ is the **dispersion parameter** that controls the degree of overdispersion in the negative binomial distribution.

All coefficients $(\beta_0, \beta_1, \beta_2, \beta_3)$ are assigned the prior of Normal(0, 2.5), which is the `rstanarm` package's default priors. The choice of the prior is sufficient for the model in this analysis and is a common non-informative prior that allows for reasonable variability (Alexander 2024). The dispersion parameter ϕ is assigned an **Exponential(1)** prior, which reflects the observation that the variance is greater than the mean. Since we have categorical predictor variables, these priors allow the coefficients to adjust based on the data.

3.3 Model Selection

Both negative binomial model and Poisson model for the dataset was constructed using the `rstanarm` package (Goodrich et al. 2024) and R (R Core Team 2023). But the negative binomial model was chosen over the Poisson model for several reasons. First, as shown in **?@tbl-modelresults**, the variance of the outcome variable, duration, is significantly higher than the mean, indicating overdispersion. The Poisson model assumes equal mean and variance, which is not suitable in this case. The negative binomial model relaxes this assumption, allowing for overdispersion and providing a better fit for the data (Alexander 2024).Additionally, the Leave-One-Out Cross Validation (LOO-CV) results in Table 4 show that the negative binomial model has a higher ELPD (Expected Log Pointwise Predictive Density) compared to the Poisson model. The ELPD is a metric that measures the model's predictive performance, with higher values indicating a better fit to the data (Alexander 2024). The fact that the negative binomial model outperforms the Poisson model in this regard suggests that it is more effective at capturing the underlying patterns of the outbreak duration data.

Other regression models like logistic regression were not chosen because logistic regression is designed for modeling binary outcomes. Since our outcome variable, duration, is a continuous count variable representing the number of days an outbreak lasts, logistic regression is not appropriate because it cannot model continuous or count data. Linear regression was also not chosen because Poission and negative binomial distributions are more suitable for modeling count data like outbreak duration in days, where as linear regression is more suitible for continuous data.

3.4 Model Diagnostics and Validation

We conducted several key validation checks to assess its predictive performance and overall adequacy. Aside from using LOO Cross Validation technique, we also calculated the Mean Absolute Error (MAE) for both models as a metric to assess the predictive performance of the Negative Binomial model over the Poisson model. To ensure the model doesn't over fit the training data, we first split the data into training and test sets. The data was randomly divided using the `caret` package (Kuhn and Max 2008), with 80% used for model training and the remaining 20% reserved for testing. We used both models to predict the outcome variable (outbreak duration) on the test set and compared the predicted values to the actual values from the test set to compute the MAE for each model.

Table 3: Comparison of Mean Absolute Error (MAE) for Poisson and Negative Binomial Models

Model	MAE
Poisson Model	6.53
Negative Binomial Model	6.52

From Table 3, the MAE for the Poisson model is 6.53, while the MAE for the Negative Binomial model is 6.52. The difference between the MAEs is minimal, suggesting that both models perform similarly in terms of prediction accuracy. However, the Negative Binomial model may still be preferred as it accounts for overdispersion, which is more appropriate for count data. The MAE of 6.52 means that, on average, the predicted outbreak duration from the Negative Binomial model deviates from the actual duration by approximately 6.52 days. In other words, for any given outbreak in the test data, the model's prediction of the outbreak's duration is off by around 6.5 days, either overestimating or underestimating the actual duration.

In Figure 8, the MCMC algorithm is also used to check for potential issues with the model. The trace plot is constructed using a subset of categorical values from each predictor variable, as there are many values for each category. As seen in the trace plot in Figure 8 (a), the lines bounce around but remain horizontal, with a nice overlap between the chains. This indicates that the chains have effectively explored the posterior distribution (Alexander 2024). The Rhat plot evaluates whether the chains have converged to a common distribution. As seen in the Rhat plot in Figure 8 (b), the values are close to 1 and fall below 1.1, suggesting no problems with convergence (Alexander 2024). Therefore, the model appears to be properly converged, and we do not need to remove or modify predictors, adjust the priors, or re-run the model.

4 Results

To determine the reference levels for the categorical data outbreak settings, causative agent, and month, the model automatically selects the first category of each factor based on alphabetical orders of the category names. The intercept represents that the baseline outbreak duration is 2.257 days when all predictors are at their reference levels. Figure 6 shows that the

4.1 Outbreak Setting

Outbreak setting is referenced based on "Hospital-Acute Care", we compare other categories with it to evaluate how each category influence the model: - Hospital-Chronic Care (0.047): Outbreaks in this setting last 0.047 days longer than Hospital-Acute Care. - Hospital-Psychiatric (-0.080): Outbreaks in psychiatric hospitals last 0.080 days shorter than Hospital-Acute Care. - LTCH (Long-Term Care Homes) (0.465): Outbreaks in LTCHs last 0.465 days longer than in Hospital-Acute Care. - Retirement Home (0.343): Outbreaks in retirement homes last 0.343 days longer than in Hospital-Acute Care.

4.2 Causative Agent

Outbreak setting is referenced based on "Hospital-Acute Care", we compare other categories with it to evaluate how each category influence the model:

Num. Obs. (3416): Based on 3416 observations. Algorithm (sampling): The model used Bayesian sampling for estimation.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

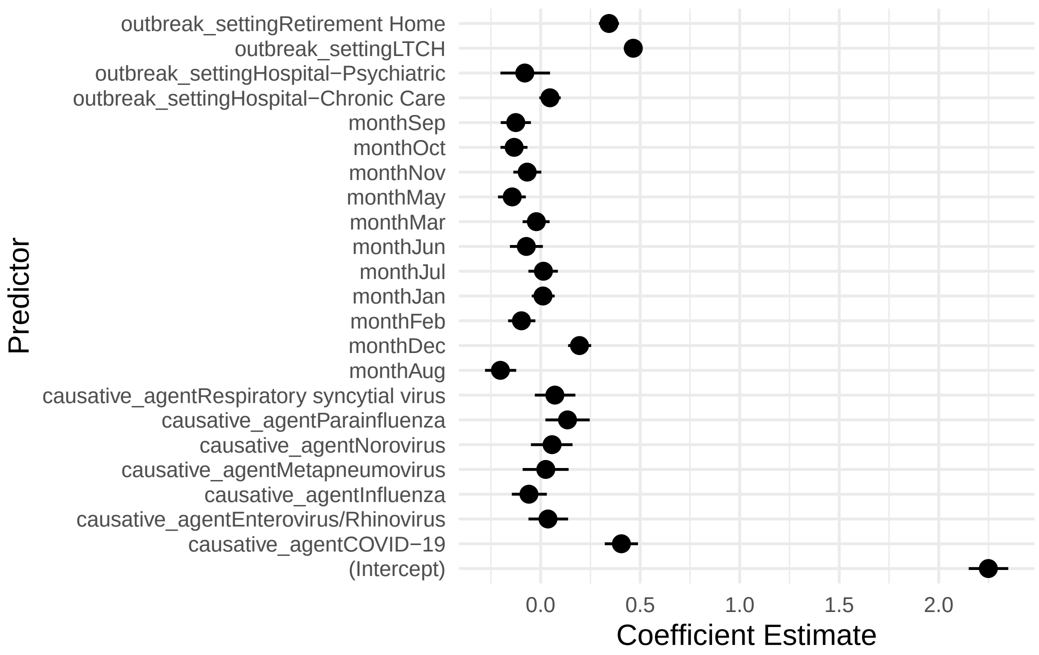


Figure 6: 90 percent credibility interval for coefficients

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

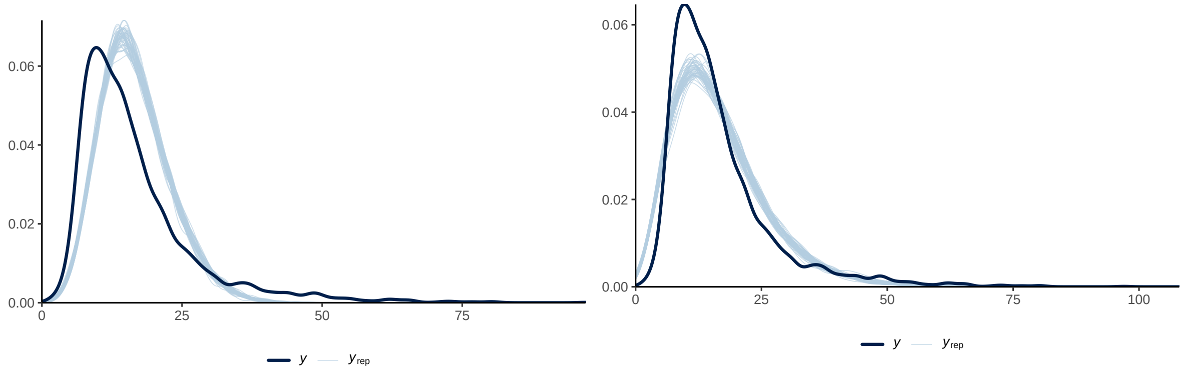
A Additional data details

B Model details

B.1 Outcome Variable Variance and Mean

B.2 Posterior predictive check

In Figure 7, using code adapted from Alexander (2024), posterior prediction checks were performed for both the Poisson model and the negative binomial model. The figure show how well the model is able to predict the observed outcomes.



(a) Poisson model posterior prediction check

(b) Negative binomial model posterior prediction check

Figure 7: Comparing posterior prediction checks for the Poisson model and the negative binomial model

B.3 Leave-One-Out (LOO) Cross Validation (CV) Comparison

In Table 4, we compare LOO performance of the Poission model against the negative binomial model based on the expected log pointwise predictive density (ELPD) and find that the negative binomial model has a higher ELPD value.

Table 4: Comparing LOO for Poisson and negative binomial models

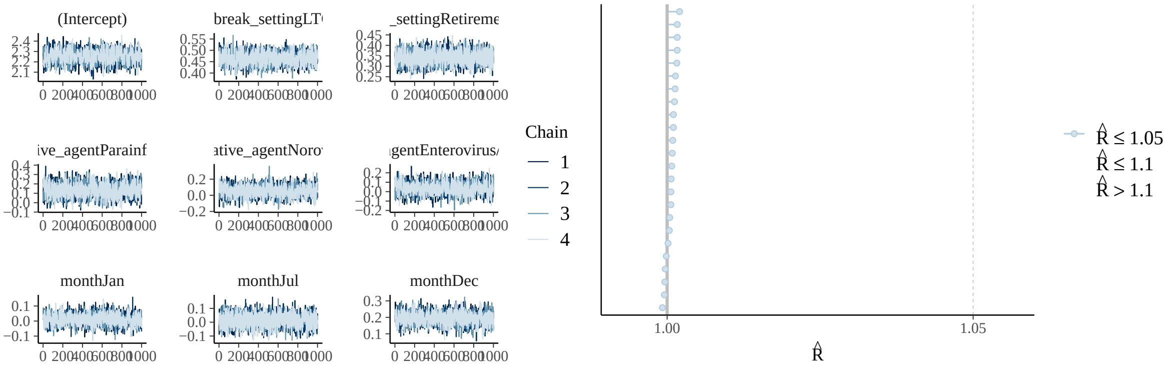
```
                   elpd_diff se_diff
neg_binomial_model     0.0       0.0
poisson_model      -3234.5     188.8
```

B.4 Model summary

Table 5 presents a summary of the model used in the analysis, which includes the intercept and the coefficients for predictor variables, and the model fitting process.

B.5 Diagnostics

Figure 8 presents the diagnostic plots for the MCMC algorithm used to estimate the parameters of our model. These plots are essential for assessing the convergence of the sampling process and ensuring the reliability of the Bayesian estimates.



(a) Trace plot

(b) Rhat plot

Figure 8: Checking the convergence of the MCMC algorithm

Table 5: Explanatory model of outbreak duration for Model (1), the negative binomial model

	(1)
(Intercept)	2.249
	(0.060)
outbreak_settingHospital-Chronic Care	0.047
	(0.033)
outbreak_settingHospital-Psychiatric	−0.080
	(0.075)
outbreak_settingLTCH	0.465
	(0.026)
outbreak_settingRetirement Home	0.343
	(0.031)
causative_agentCOVID-19	0.406
	(0.051)
causative_agentEnterovirus/Rhinovirus	0.037
	(0.058)
causative_agentInfluenza	−0.059
	(0.054)
causative_agentMetapneumovirus	0.026
	(0.070)
causative_agentNorovirus	0.057
	(0.063)
causative_agentParainfluenza	0.135
	(0.068)
causative_agentRespiratory syncytial virus	0.071
	(0.062)
monthAug	−0.202
	(0.047)
monthDec	0.195
	(0.035)
monthFeb	−0.097
	(0.042)
monthJan	0.012
	(0.035)
monthJul	0.014
	(0.046)
monthJun	−0.072
	(0.050)
monthMar	−0.022
	(0.040)
monthMay	−0.143
	(0.040)
monthNov	−0.068
	(0.041)
monthOct	−0.133
	(0.041)
monthSep	−0.125
	(0.047)
Num.Obs.	3416
algorithm	sampling
pss	4000

References

Alexander, Rohan. 2024. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Arel-Bundock, Vincent. 2022. "modelsummary: Data and Model Summaries in R." *Journal of Statistical Software* 103 (1): 1–23. https://doi.org/10.18637/jss.v103.i01.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2024. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Kuhn, and Max. 2008. "Building Predictive Models in r Using the Caret Package." *Journal of Statistical Software* 28 (5): 1–26. https://doi.org/10.18637/jss.v028.i05.

R Core Team. 2023. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

team, The pandas development. 2020. "Pandas-Dev/Pandas: Pandas." Zenodo. https://doi.org/10.5281/zenodo.3509134.

Toronto Public Health. 2024. *Outbreaks in Toronto Healthcare Institutions*. https://open.toronto.ca/dataset/outbreaks-in-toronto-healthcare-institutions/.

Van Rossum, Guido, and Fred L. Drake. 2009. *Python 3 Reference Manual*. Scotts Valley, CA: CreateSpace.

Vehtari, Aki, Jonah Gabry, Måns Magnusson, Yuling Yao, Paul-Christian Bürkner, Topi Paananen, and Andrew Gelman. 2024. "Loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models." https://mc-stan.org/loo/.

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.