Predicting Infection Outbreak Dynamics in Toronto Healthcare Institutions: A Bayesian Approach

Respiratory Outbreaks Are More Likely in Long-Term Care Homes and Retirement Homes, Driven by Seasonal Trends and Causative Agents Like COVID-19 and Influenza

Divya Gupta

December 3, 2024

Abstract

Identifying key trends in infection outbreaks in a city can help policymakers implement better reforms to support citizens and prevent outbreaks. This study uses cleaned datasets retrieved from Open Data Toronto to develop a model that can have insightful implications for the City of Toronto policymakers, especially regarding resource allocation for outbreak prevention. This study aims to help policymakers better understand the trends of infection outbreaks in Toronto, so that this data can be used to identify problem areas in the city (such as specific times of year). This study examines the factors influencing healthcare outbreaks in Toronto healthcare institutions, such as: type of infection (respiratory vs. gastroenteric), causative agent of infection, setting and month of infection. This study uses Bayesian logistic regression to identify the factors associated with infectious outbreaks and predict the likelihood of a specific outbreak type depending on the aforementioned variables. The findings show that COVID-19 is a significant causative agent for respiratory infections, while norovirus is a causative agent for enteric outbreaks. Moreover, there is a significant increase in the number of outbreaks in winter months and in long term care homes.

1. Introduction

Toronto has multitudes of healthcare institutions such as hospitals, long-term care homes and retirement homes that often house hundreds of people at a time and can serve as places of community for vulnerable population groups. Any healthcare-associated infectious outbreaks in such spaces pose substantial risk for all vulnerable populations in the space. Understanding the varying factors that influence and further an outbreak - such as its causes, duration, and type of outbreak - can help better understand the nature of the outbreak and plan interventions and outbreak prevention measures accordingly. Furthermore, analyzing data about infection outbreak trends in Toronto can help us predict the time, region and duration of the next outbreak occurrence which can be instrumental in helping vulnerable populations such as older people and children. This study aims to answer the research question: *How can Bayesian models help predict infection outbreak behavior, specifically based on the outbreak's healthcare setting, causative agent and month of outbreak?*

The estimand for this analysis is the probability of an infection outbreak being of a specific type (namely respiratory or enteric) as determined by the causative agent, month of outbreak, or healthcare setting.

Developing a robust Bayesian model that can predict outbreaks in the city and develop outbreak management strategies in real time, ensuring better resource allocation and healthcare management. I will be employing Bayesian Logistic Regression to predict the likelihood of respiratory versus enteric outbreaks and how they are influenced based on the healthcare institution, causative agent, duration of previous outbreaks, and time of the year. I will then be employing Bayesian Survival Analysis to model the duration of outbreaks and identify defining factors specifically in prolonged cases (duration of infection > 10 days).

The remainder of the paper is structured as follows. Section 2 details the data, measurement, along with helpful visualisations to understand the data. Section 3 outlines the modeling approach, while Section 4 presents the results. Section 5 discusses implications, limitations, and future directions. This is followed by the Appendix.

2. Data

This dataset, entitled "Outbreaks in Toronto Healthcare Institutions" was obtained from Open Data Toronto on November 25, 2024 (Open Data Toronto, 2024) and possesses data from January 2nd, 2024 until November 20, 2024.. It is published by Toronto Public Health and documents infection outbreaks reported by various healthcare institutions in the city (the dataset is updated weekly on Thursday to capture the latest information). This is done in compliance with Ontario's Health Protection and Promotion Act, which requires healthcare facilities to continuously report all infection outbreaks to their local authorities. This level of data collection allows timely identification and potential containment and reform in affected communities.

For the purpose of this study, an infection outbreak (here on referred to as "outbreak") is defined as a localized increase in the rate of infection above the baseline. In terms of settings, this dataset gathers information from hospitals (acute, psychiatric or chronic care), long term care homes (here on referred to as "LTCH"), retirement homes, and transitional care systems. Further, the dataset includes information about the type of outbreak (for eg., respiratory or enteric), location, the start and end date of the outbreak, and the causative agent(s). While Open Data Toronto possesses multiple other datasets regarding infections and outbreaks in Toronto, they were localised to information about homeless shelters or solely tracked COVID-19 outbreaks in the city. The chosen dataset encompasses a wider range of outbreak settings and includes various causative agents, including COVID-19, positioning it as a wider dataset - allowing us to extract greater insights and holistically understand the nature of infection outbreaks in Toronto.

For data cleaning, analysis and visualisation, this study uses R (R Core Team 2023), along with packages such as tidyverse for visualisation (Wickham et al. 2019), dplyr for data wrangling, and lubridate for handling temporal data. Further, ggplot2 was used for creating visualizations (Wickham, Chang, et al. 2023) and forcats for working with categorical variables. For the Bayesian analysis, rstanarm was and brms were used. Readr was used to read the structured data and scales allowed further development upon the visualisations. We also used ggeffects for visualisations of the model predictions and testthat for unit testing.

Measurement

The measurement process aimed to transition real-world phenomena such as infection outbreaks in healthcare facilities into usable, numerical or categorical data that can be used for analysis. Each data entry in the data set refers to a case of infection outbreak in Toronto from January - November 2024. Variables of interest include:

Type of Outbreak: This refers to the nature of the outbreak, categorized as either respiratory or enteric - based on the symptoms of the infection. Respiratory infections are categorized by symptoms such as cough, fever, and sore throat. Enteric infections are categorized by symptoms such as diarrhea, nausea and vomiting.
Outbreak Setting: This includes variables such as hospitals (acute, psychiatric or chronic care), long-term care homes (LTCHs), retirement homes, or transitional care facilities.

Causative Agent: This refers to the variable, “Causative Agent - 1” in the dataset. This records the primary pathogen named to be responsible for the outbreak. Significant data entries included COVID-19, norovirus, and Influenza as the causative agents. This categorical variable can help one understand the nature of the infection and its instigating factors.

Date Outbreak Began: Records the first date, reported as YYYY-MM-DD, that the outbreak was reported.

Date Declared Over: Records, as a YYYY-MM-DD date, when all affected patients were declared infection-free, marking the end of the outbreak.

Constructed Variables

We also constructed some variables for the ease of data analysis from the provided variables from the raw dataset. These include:

Duration: This was calculated as the difference between Date Declared Over and Date Outbreak Began to give a numerical value for the number of days an outbreak lasted.

Month: This was constructed by monitoring the start date of an outbreak, from Date Outbreak Began, to track how seasonal changes may affect outbreak levels in the city.

Accuracy and Limitations

The Open Data Toronto portal provided the raw dataset that had some inconsistencies, namely - various fields had missing entries in the “Date Declared Over” tab, which rendered those data points useless as we could not track the duration of the outbreak. This value is essential for this analysis as our variable, Duration, was constructed as the difference between “Date Declared Over” and “Date Outbreak Began.” Further, some data points had no specific entry for “Type of Outbreak”, and had values like “other” which did not provide usable information for this analysis. For ease of analysis, these data entries were removed while constructing the analysis dataset.

Understanding the Data

Outcome Variable

Our model predicts the probability of an outbreak being labeled as respiratory or enteric based on its predictors. To that effect, our outcome variable is “Type_of_Outbreak”, with its results being in binary form of either respiratory or enteric.

Predictor Variables

Our analysis mainly focused on three predictor variables to help determine our outcome variable:

1. Outbreak Setting: This categorical variable utilises information about which type of healthcare institution housed an outbreak, such as hospitals of LTCH.
2. Causative Agent: This categorical variable houses information about the primary cause of the outbreak, such as COVID-19 or Influenza.
3. Month: This categorical variable represents the month of the year in which the outbreak began.

Visualising the Data

Figure 1 illustrates the difference between the levels of respiratory and enteric outbreaks in the city. As can be seen, the overwhelming amount of outbreaks reported are respiratory.

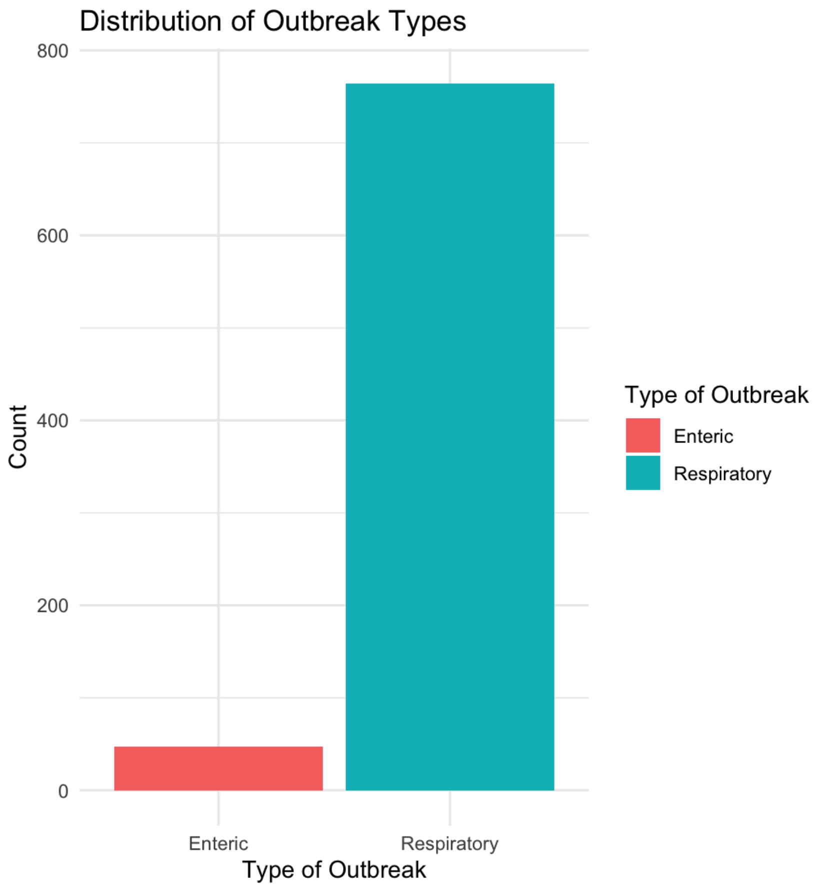


Figure 1. Distribution of Outbreak Types

Figure 2 takes this one step further and illustrates the number of outbreaks determined by healthcare settings. This shows that LTHC had an overwhelming majority in the number of outbreaks reported, followed by retirement homes.

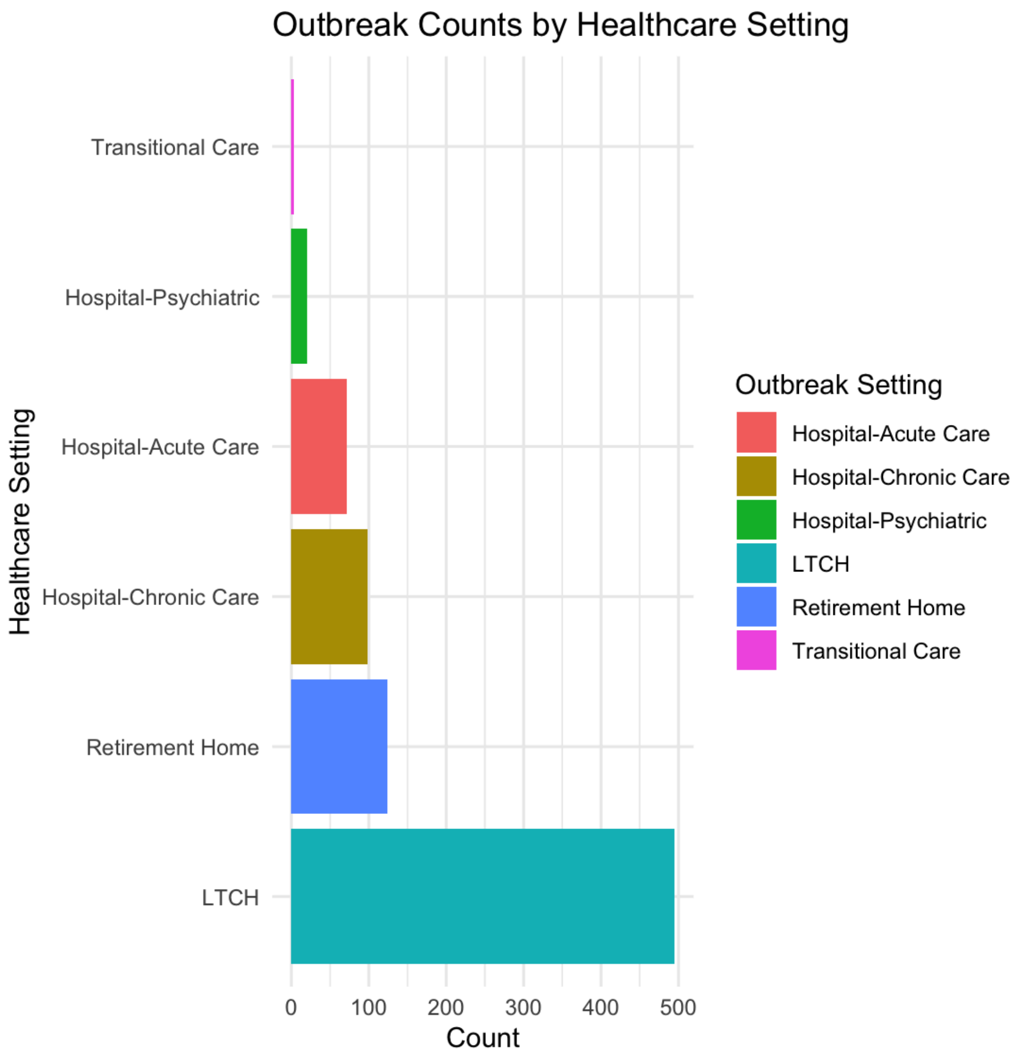


Figure 2. Outbreak Counts by Healthcare Setting

Next, we aimed to understand the interaction between different types of infection outbreaks and the setting in which they occurred. This is illustrated in Figure 3 which is a stacked bar chart. Key insights included the fact that respiratory infections dominated across all types of healthcare settings. While enteric outbreaks are less common, they are most prevalent in retirement homes and hospitals-acute care. Psychiatric care units in hospitals and transitional care systems reported no cases of enteric outbreaks in this dataset.

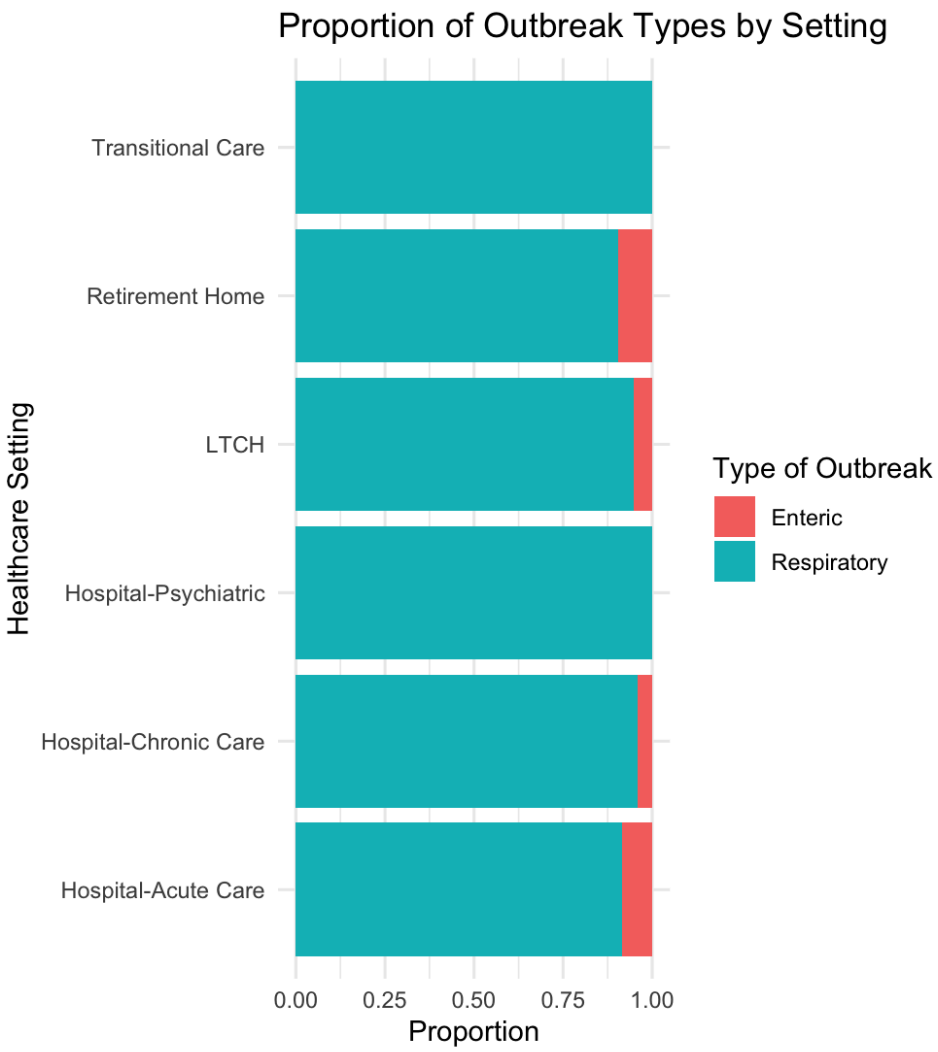


Figure 3. Proportion of Outbreak Types by Setting

Lastly, we aimed to understand the overall interaction trends between the outbreak duration, the causative agent, and the setting of the outbreak. This is illustrated in Figure 4.

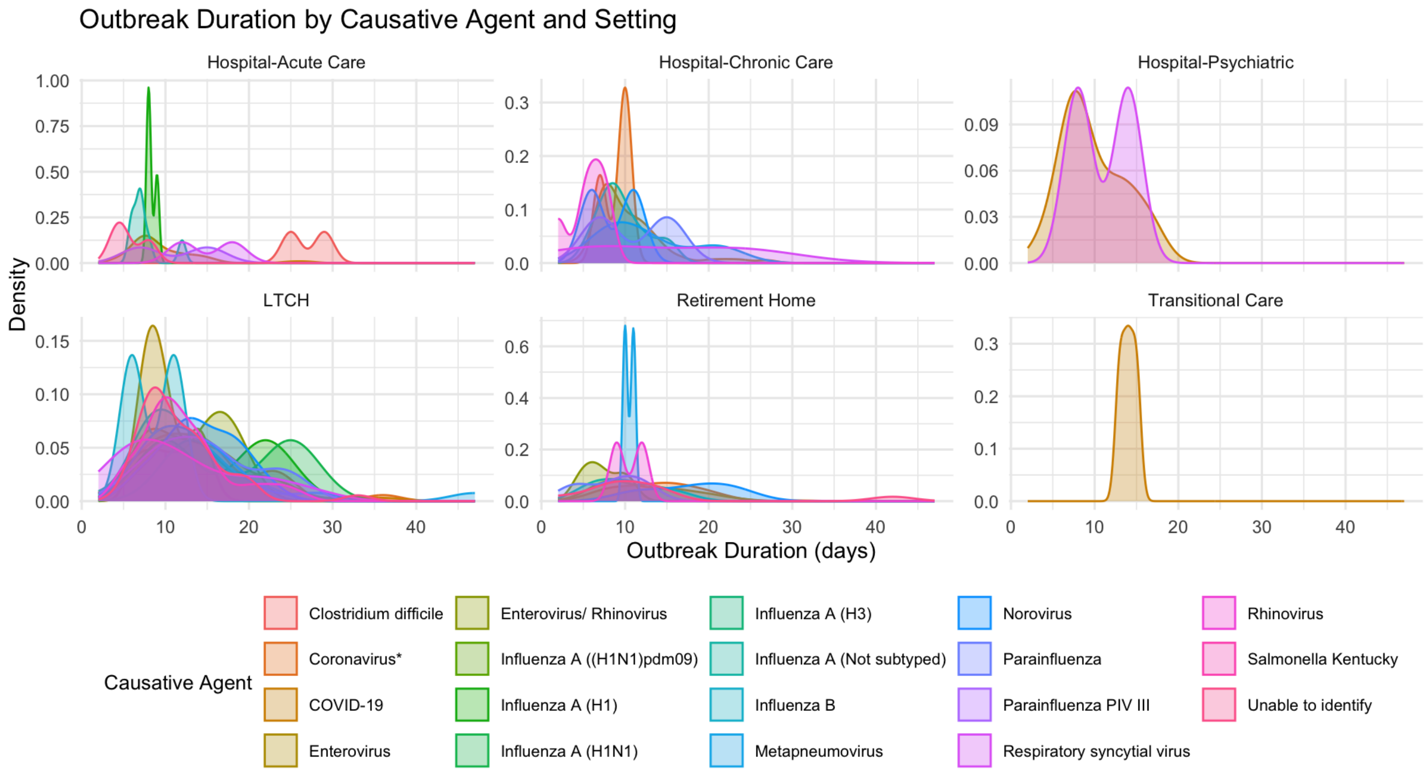


Figure 4. Outbreak Duration by Causative Agent and Setting

This series of density plots illustrates the duration of outbreaks as stratified by the causative agents and various healthcare settings included in the dataset. This shows that LTHC and Retirement homes report significantly longer outbreak durations, specifically for causative agents such as norovirus and Influenza. Further, transitional care settings have a short outbreak duration. This may be due to strict infection control practices - this raises a question about outbreak control policies in less-monitored settings such as retirement homes which have reported longer outbreak durations. These differences in durations according to healthcare settings also raises a need for setting-tailored outbreak interventions.

Overall, these visualisations help understand key parts of the data that will inform our model. Namely, they illustrate that an overwhelming majority of infections are respiratory in nature. Next, there is a difference in the duration of an outbreak according to the healthcare setting, and that there are potential impacts of the causative agent and the month in which the outbreak occurred upon the duration of the outbreak.

3. Model

3.1 Model Set-Up

I employed two Bayesian models:

1. **Logistic Regression**

$yi \sim \mathrm{Bernoulli}(\pi i), \mathrm{logit}(\pi i)=\beta 0 + \beta 1 \times Xsetting + \beta 2 \times Xagent + \beta 3 \times Xmonth$

Priors:

($\beta \sim N(0, 2.5)$) -> weakly informative normal priors that we can ensure the regularization of the model without imposing strong assumptions.

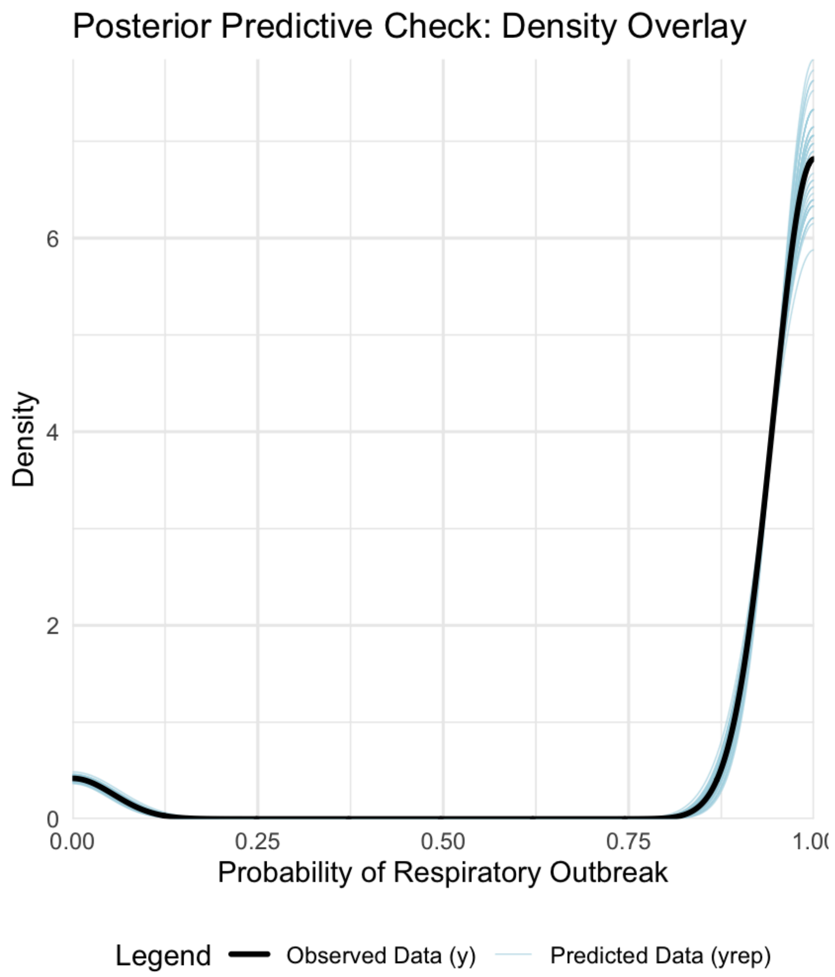
2. **Survival Analysis**

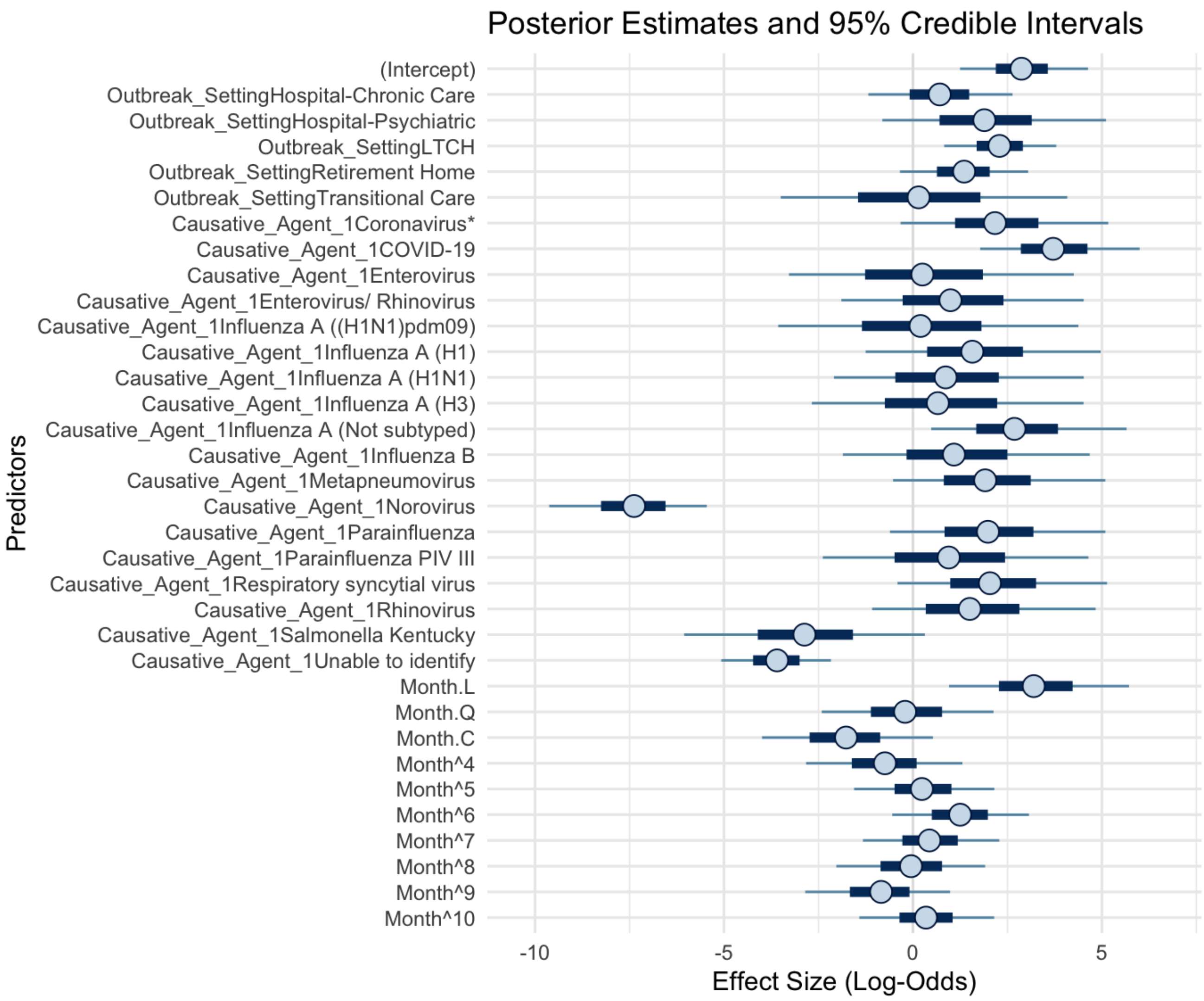
The models were implemented using the **rstanarm** and **brms** packages in R.

3.2 Model Validation

Validation is critical to ensure the models are reliable and provide accurate predictions. The following steps were undertaken to validate the models:

1. **Posterior Predictive Checks**:
   - Compared observed data distributions with data simulated from the posterior predictive distribution.
   - Ensured the models captured key patterns without overfitting.


Posterior Predictive Check: Density Overlay
Density
6
4
2
0
0.00
0.25
0.50
0.75
1.0
Probability of Respiratory Outbreak
Legend
Observed Data (y)
Predicted Data (yrep)



2. **Cross-Validation**:
   - Employed leave-one-out cross-validation (LOO-CV) to compare models and assess predictive performance.
   - Metrics such as **WAIC** (Widely Applicable Information Criterion) and **LOOIC** (Leave-One-Out Information Criterion) provided a robust basis for model selection.
3. **Sensitivity Analysis**:
   - Tested the robustness of the models by varying priors and including interaction terms.
   - Assessed the impact of removing outliers and reanalyzing the data.
4. **Out-of-Sample Testing**:
   - Reserved a portion of the data for testing, comparing predicted probabilities and durations against actual outcomes.

3.3 Model Justification

The priors used in the model allow the integration of domain knowledge, such as typical outbreak durations or pathogen characteristics. Moreover, real-time updates to the model by acquiring new data will allow for it to continue being relevant to Toronto public health, making it a dynamic model that can help manage outbreaks.

4. Results

4.1 Logistic Regression

Key predictors of outbreak type included:

- Outbreak Setting: Hospitals were more likely to experience respiratory outbreaks.
- Causative Agent-1: Norovirus was strongly associated with enteric outbreaks.

Causative agent COVID-19 (β=3.8\beta = 3.8β=3.8) strongly increases the likelihood of respiratory outbreaks, while Norovirus (β=−7.4\beta = -7.4β=−7.4) is strongly associated with enteric outbreaks.

Long-Term Care Homes (LTCH) (β=2.3\beta = 2.3β=2.3) have a higher likelihood of respiratory outbreaks than hospitals.

Some month-related predictors (e.g., Month.L\text{Month.L}Month.L) show positive associations, suggesting seasonal trends.

Causative Agents Are Key Predictors:

- COVID-19 and other respiratory viruses (e.g., Influenza A) significantly increase the likelihood of respiratory outbreaks.
- Norovirus and Salmonella are strong predictors of enteric outbreaks.

Healthcare Settings Matter:

- Long-Term Care Homes and Retirement Homes are more prone to respiratory outbreaks compared to hospitals.
- This suggests differences in patient demographics, environmental factors, or infection control practices.

Seasonality Is Relevant:

- Month-related predictors indicate seasonal variation in respiratory outbreaks, with winter months likely contributing to higher probabilities.

Model Fit and Validity:

- The model fits well, with R^=1.0\hat{R} = 1.0R^=1.0 and high effective sample sizes. This suggests reliable and converged posterior distributions.

5. Discussion

5.1 Insights

COVID-19 remains a significant driver of respiratory outbreaks, highlighting the importance of ongoing prevention measures.
Norovirus's strong association with enteric outbreaks suggests a need for targeted hygiene protocols.

- Long-Term Care Homes and Retirement Homes could benefit from additional respiratory infection control measures, such as air quality monitoring and staff training.

5.2 Limitations

5.3 Future Directions

Appendix

A. Data Details

B. Model Details

References