Predicting Institutional Outbreak Characteristics in Toronto Using Bayesian Models

Findings TBA

Divya Gupta

November 26, 2024

Abstract

This study aims to examine the factors influencing healthcare outbreaks in Toronto healthcare institutions. These factors include: type of infection (respiratory vs. gastroenteric), causative agent of infection, and duration of infection. These variables are potentially influenced by the setting (type of healthcare institution), time of year, and causative agent. This study uses logistic regression and Bayesian survival analysis to identify the factors associated with infectious outbreaks and predict the likelihood of a specific outbreak type depending on the aforementioned variables. This study uses cleaned datasets retrieved from Open Data Toronto to develop models that can have insightful implications for the City of Toronto policymakers, especially regarding resource allocation for outbreak prevention.

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: *What factors influence the type and duration of healthcare-related outbreaks in Toronto institutions, and how can Bayesian models help predict infection outbreak behavior?*

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

2.1 Dataset

2.2 Data Cleaning

The Toronto Open Data portal provided the raw dataset, which was cleaned and preprocessed as follows:

- Missing values in Date Declared Over were removed.
- Entries with Type of Outbreak as "Other" were excluded.
- Duration was calculated as the difference between Date Declared Over and Date Outbreak Began.

2.3 Measurement

2.4 Exploratory Data Analysis

Initial analyses revealed:

- Respiratory outbreaks were more common than enteric outbreaks. (Figure 1)
- Long-Term Care Homes experienced the highest number of outbreaks. (Figure 2)
- Duration varied significantly by setting and causative agent. (Figure 3)

3. Model

3.1 Model Set-Up

I employed two Bayesian models:

1. **Logistic Regression**

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

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.
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.

4.2 Survival Analysis

5. Discussion

5.1 Insights

5.2 Limitations

5.3 Future Directions

Appendix

A. Data Details

B. Model Details

References