Predicting Institutional Outbreak Characteristics in Toronto Using Bayesian Models
Findings TBA

Divya Gupta
November 26, 2024

Abstract

This paper investigates the factors influencing the type and duration of healthcare institutional outbreaks in Toronto. Using Bayesian logistic regression, we predict the likelihood of a respiratory versus enteric outbreak based on outbreak setting, time of year, and causative agents. Additionally, Bayesian survival analysis identifies factors associated with prolonged outbreaks. The study leverages simulated and cleaned datasets to develop models that provide actionable insights for public health policymakers.

1. Introduction

Healthcare-associated outbreaks in Toronto's institutions pose substantial risks to vulnerable populations. These outbreaks, which include respiratory and enteric infections, vary significantly in their causes, durations, and impacts. Understanding the factors that influence outbreak characteristics is critical for implementing effective prevention and control strategies. This study addresses the research question: *What factors influence the type and duration of healthcare-related outbreaks in Toronto institutions?*

Bayesian methods are particularly suited for this analysis due to their ability to incorporate prior knowledge and quantify uncertainty. Specifically:

1. **Bayesian Logistic Regression**: Predicts the likelihood of respiratory versus enteric outbreaks based on healthcare setting, causative agents, and the time of year.
2. **Bayesian Survival Analysis**: Models the duration of outbreaks, identifying factors associated with prolonged cases.

The significance of this work lies in its potential to inform outbreak management strategies in real time, ensuring resources are allocated efficiently and interventions are targeted effectively. Using both simulated and cleaned datasets, we demonstrate the applicability and robustness of Bayesian methods in this public health context.

The paper employs Bayesian logistic regression to predict outbreak types and Bayesian survival analysis to analyze outbreak durations. The remainder of the paper is structured as follows. Section 2 details the data. Section 3 outlines the modeling approach, while Section 4 presents the results. Section 5 discusses implications, limitations, and future directions.

2. Data

2.1 Dataset

A simulated dataset was constructed to test and refine analysis pipelines. The dataset included variables such as Institution Name, Outbreak Setting, Type of Outbreak, Causative Agents, and Duration. Random sampling methods ensured realistic variability, such as generating outbreak durations between 5 and 30 days and assigning probabilities for settings and outbreak types.

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**

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

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. These packages streamline Bayesian analysis by automating sampling and diagnostics.

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

Bayesian methods offer several advantages in this context:

- **Flexibility**: Priors allow the integration of domain knowledge, such as typical outbreak durations or pathogen characteristics.
- **Uncertainty Quantification**: Posterior distributions provide a natural way to capture and communicate uncertainty in parameter estimates.
- **Applicability to Public Health**: Real-time updates to the model are possible by incorporating new data, a crucial feature for dynamic outbreak management.

The logistic regression model is ideal for binary outcomes, enabling predictions about outbreak type. Survival analysis accounts for censored data, such as ongoing outbreaks, making it a robust choice for modeling durations.

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