Modeling Outbreak Frequency by Season and Pathogen: Insights and Predictions for 2025*

COVID-19 Dominates with Steady Frequency, Other Pathogens Peak in Summer

Tianrui Fu

December 2, 2024

We analyze the outbreak dynamics of respiratory pathogens in Toronto's healthcare settings, focusing on seasonal trends, pathogen-specific patterns, and predictive modeling. Using Poisson regression on historical data, we find COVID-19 dominates outbreak frequencies, with summer months showing peaks across most pathogens and lower activity in winter. Predictions for 2025 indicate similar seasonal patterns, reinforcing the importance of seasonally adjusted preparedness measures. These results highlight the need for strategic resource allocation to mitigate the burden of outbreaks while ensuring healthcare systems remain resilient to seasonal variability.

Table of contents

*Code and data are available at: https://github.com/FrankFU323/Toronto-Healthcare-outbreak-trends.git.

1 Introduction

Outbreaks of respiratory pathogens in healthcare institutions pose ongoing challenges to public health systems, especially in urban environments like Toronto where healthcare demands are high. Understanding the dynamics of these outbreaks across different seasons and pathogens is essential for developing timely and effective control strategies. Despite advancements in medical infrastructure and infection control measures, there is still a lack of systematic research on the seasonal variations of outbreaks caused by different pathogens. This variability presents significant challenges for resource allocation and the resilience of healthcare systems, particularly when responding to emerging infectious diseases.

This study utilizes data from 2024 related to Toronto healthcare institutions to explore the patterns and influencing factors of respiratory pathogen outbreaks. We provide a detailed description of the data sources, the measurement methods of key variables, and their distribution characteristics. Subsequently, the study employs a Poisson regression model, considering key variables such as pathogen type, seasonal variation, and outbreak setting, to assess their roles in outbreak dynamics and to offer a basis for future trend predictions. The predictive results showcase potential outbreak trends for 2025, serving as a reference for planning interventions. By analyzing the outbreak patterns of different pathogens, this study aims to offer actionable recommendations to enhance seasonal preparedness and response efficiency.

The results show that COVID-19 continues to dominate the frequency of outbreaks in healthcare institutions. Notably, most pathogens have outbreak peaks in the summer, with significantly reduced activity in winter. The predictive model indicates that these seasonal trends may persist, which emphasizes the importance of adjusting control measures to accommodate seasonal variations.

This study highlights the necessity of incorporating seasonal changes into outbreak management strategies. By using predictive models to identify high-risk periods in advance, healthcare institutions can allocate resources more effectively and strengthen system resilience. The structure of this paper is as follows: Section 2 provides a detailed introduction to data sources and the definition and exploration of variables; Section 3 describes the model construction process and theoretical basis; Section 4 presents the model summary and predictive results; finally, Section 5 discusses the practical significance of the findings in the context of policy and proposes specific recommendations for improving outbreak prevention capabilities in healthcare institutions.

2 Data

2.1 Overview

The data for this study was sourced from Open Data Toronto and downloaded using the Gelfand (2022) package. It consists of 846 observations from 2024 across 10 variables.

Our analysis was conducted using the statistical programming language R (R Core Team 2023), leveraging a suite of packages to facilitate efficient data manipulation, modeling, and visualization. Data cleaning, transformation, and visualization were handled using `tidyverse` Wickham et al. (2019) and `dplyr` Wickham et al. (2023). The core analysis used Poisson regression models to explore relationships between outbreak counts and key variables, including pathogen type, seasonal variation, and outbreak settings. These models were implemented using the flexible framework offered by `rstanarm` Goodrich et al. (2022) for model development and extensions. Efficient data integration was achieved through `arrow` Richardson et al. (2024), and modeling workflows were supported using `modelr` Wickham (2023) within the tidyverse framework. Model summaries were generated using `modelsummary` Arel-Bundock (2022), while `ggplot2` Wickham (2016) was employed for creating visualizations. Enhanced table outputs were produced with `kableExtra` Zhu (2024), and dynamic report generation was performed with `knitr` Xie (2023). File management, including reading parquet files, was streamlined with `here` Müller (2020).

The analysis framework followed the structured approach provided by Alexander (2023b), ensuring consistency and clarity in integrating data preparation and modeling workflows. Additionally, the critique and guidelines outlined in Alexander (2023a) informed our presentation of the data and results, emphasizing clarity and coherence.

2.2 Data cleaning

The cleaning process for the dataset focused on retaining key variables relevant to respiratory outbreaks in Toronto healthcare institutions. Initially, redundant or irrelevant columns were removed, and the dataset was narrowed to include only `institution_address`, `outbreak_setting`, `type_of_outbreak`, `causative_agent_1`, `date_outbreak_began`, and `active` variables. Outbreaks were filtered to focus on respiratory pathogens of interest, such as "COVID-19", "Coronavirus*",Rhinovirus", and "Parainfluenza". The season variable was derived from the `date_outbreak_began` column by mapping months to their respective seasons like December-February as Winter. Invalid or missing data were addressed by removing incomplete rows or invalid entries, ensuring data consistency. These steps ensured that the dataset was streamlined for meaningful analysis while retaining critical details.

The creation of the `count_data` dataset involved aggregating the cleaned data to prepare it for statistical modeling and visualization. Outbreak events were grouped by `causative_agent_1`, `season`, `month`, and `outbreak_setting`, and the `outbreak_count` variable was calculated as the total number of outbreaks for each combination of these factors. This aggregation provided a structured summary of outbreak frequency, enabling comparisons between pathogens, seasonal trends, and healthcare settings. By summarizing the data, `count_data` served as the foundation for constructing Poisson regression models and generating meaningful visualizations to explore outbreak dynamics.

2.3 Measurement

The data used in this study captures respiratory pathogen outbreaks in Toronto healthcare settings, collected through systematic reporting and published via Toronto (2024). Each outbreak is recorded as a discrete event, including the causative pathogen, the setting like hospital, long-term care facility, and key dates: when the outbreak began and when it was declared.

The dataset translates real-world outbreak events into structured entries, with the primary variable being outbreak counts for each pathogen. Dates provide a temporal framework for analyzing patterns, while outbreak settings add contextual detail. Data is derived from clinical diagnoses and laboratory confirmations, ensuring reliability, though differences in reporting practices may introduce variability.

This measurement framework ensures that the dataset provides a reliable overview of respiratory pathogen activity in Toronto healthcare settings. It transforms clinical and public health surveillance data into a structured format suitable for statistical analysis, offering a robust foundation for understanding outbreak dynamics.

2.4 Outcome variables

Figure 1 illustrates the outcome variable – total outbreak counts of four respiratory pathogens (Coronavirus, COVID-19, Parainfluenza, Rhinovirus) across four seasons (Fall, Spring, Summer, Winter). COVID-19 exhibits the highest outbreak frequency overall, peaking in Summer with 133 outbreaks and showing consistent activity across other seasons (Fall: 125, Winter: 114, Spring: 98). Rhinovirus demonstrates a strong seasonal trend, with 27 outbreaks in Fall but almost negligible activity in Winter (1 outbreak). Parainfluenza outbreaks are more evenly distributed, with notable peaks in Spring (23) and Summer (18), while Coronavirus is most active in Spring (21) and shows minimal activity in Fall (4). These seasonal patterns emphasize pathogen-specific dynamics, highlighting the importance of considering both pathogen type and season when planning outbreak prevention strategies.

The Figure 1 provides a more detailed and insightful view of the outcome variable `outbreak_count` by breaking it down into specific groups such as `causative_agent_1` and `season`. This grouping allows us to explore how outbreaks vary across different pathogens and seasons, providing a richer understanding of patterns within the data. The plot with just `outbreak_count` is included in Appendix A.2.

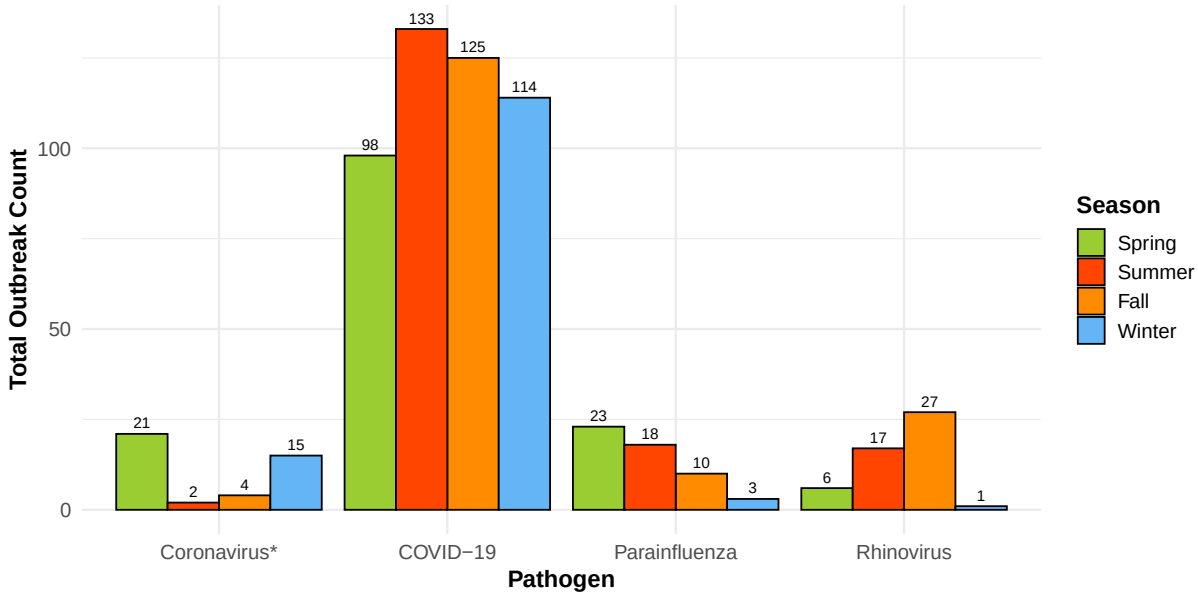

Figure 1: Outbreak Counts by Pathogen and Season

2.5 Predictor variables

2.5.1 Causative Agent

The `causative_agent_1` variable represents the primary pathogen responsible for outbreaks, such as COVID-19, Rhinovirus, and others. As shown in Figure 2, COVID-19 accounts for a significantly higher number of outbreaks (55 events) compared to other pathogens, such as Parainfluenza and Rhinovirus, which each account for 14–15 events. This stark disparity underscores the dominant role of COVID-19 in driving outbreaks and reflects its heightened transmissibility and prevalence. Including this predictor is critical to differentiate outbreak patterns across pathogens, as it enables tailored intervention strategies. For instance, targeting COVID-19 mitigation could significantly reduce the overall outbreak burden, given its disproportionate impact. This variable also highlights the need for pathogen-specific public health responses to minimize transmission risks effectively.

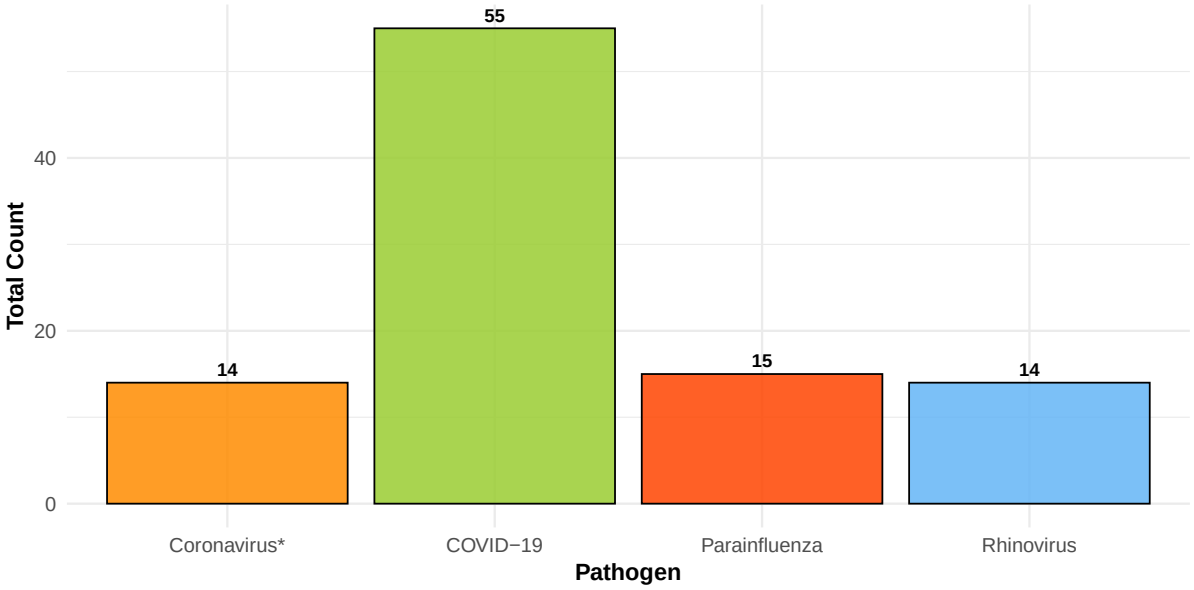

Figure 2: Outbreak Counts by Pathogen

2.5.2 Season

The `season` variable captures the seasonal variability in outbreaks, categorizing events into Spring, Summer, Fall, and Winter. As shown in Figure 3, Spring (30 events) and Fall (27 events) exhibit the highest number of outbreaks, while Winter (17 events) has the lowest. These trends align with the seasonal nature of many respiratory pathogens and the influence of climate conditions on transmission. By including this variable, the analysis accounts for temporal variability, enabling public health planners to allocate resources strategically during

high-risk seasons. For example, strengthening outbreak preparedness in Spring and Fall could reduce peak case loads. Additionally, this predictor emphasizes the importance of understanding seasonal drivers of infection to develop more effective prevention and control strategies.

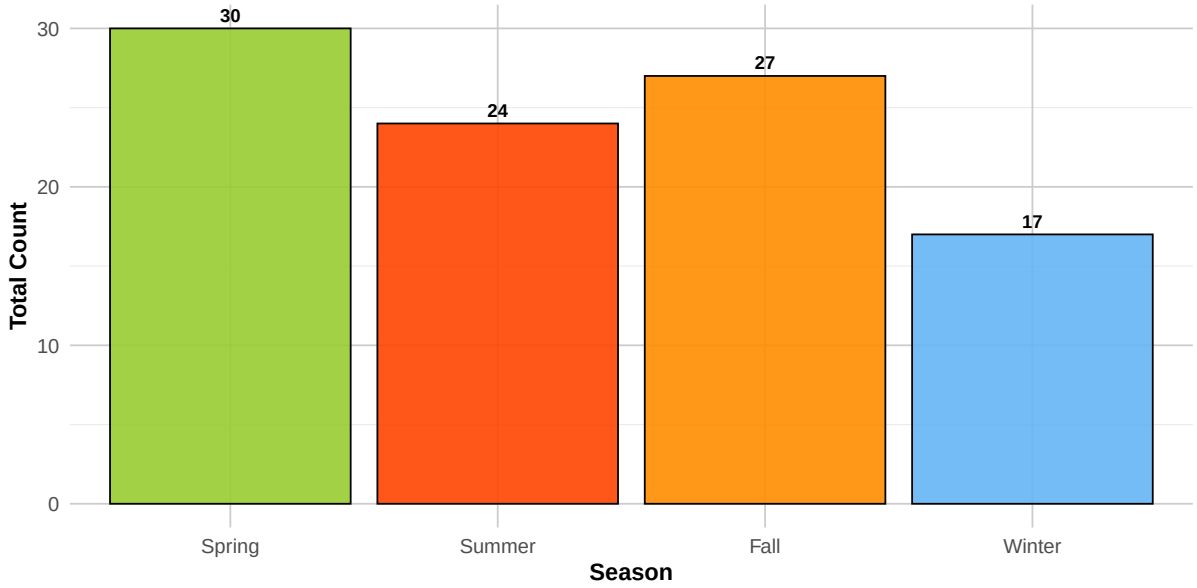

Figure 3: Outbreak Counts by Season

2.5.3 Month

The `month` variable provides granular temporal information about outbreaks throughout the year until now. As illustrated in Figure 4, May (12 events) and October (11 events) are the peak months for outbreaks, while most other months have relatively stable counts of 8–9 events. Monthly variation may stem from factors such as pathogen transmission cycles, seasonal transitions, or changes in human behaviors, such as increased social interactions during specific times of the year. This predictor allows for a finer temporal resolution in understanding outbreak dynamics and facilitates timely interventions, such as intensifying infection control measures in May and October. Month-level analysis also supports adaptive responses to temporal drivers that impact outbreak frequencies.

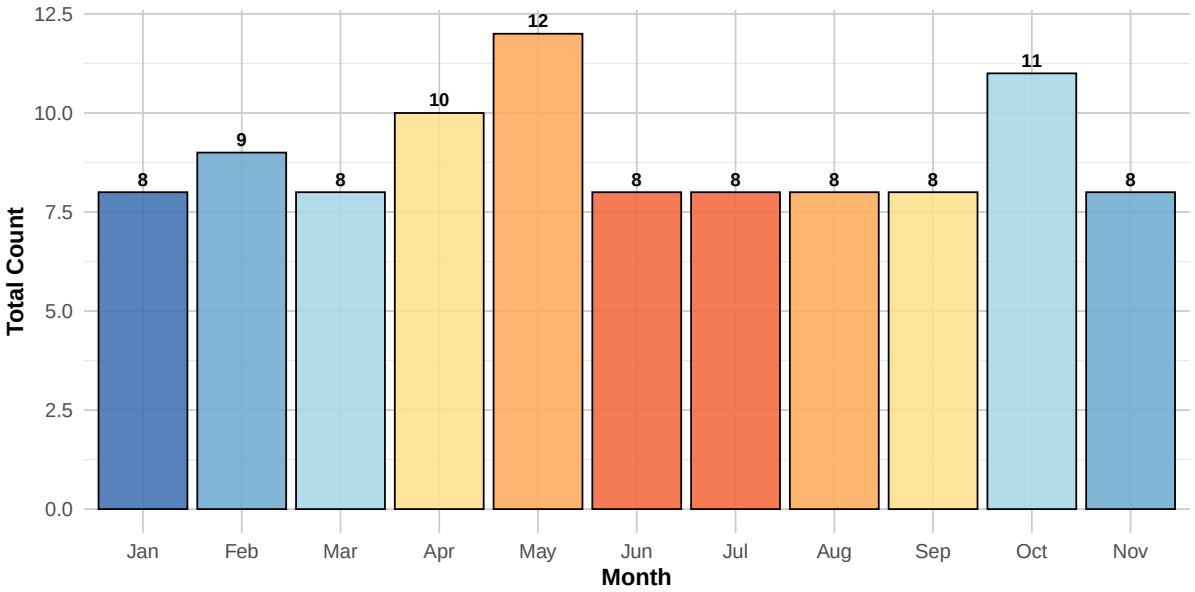

Figure 4: Monthly Outbreak Trends

2.5.4 Outbreak Setting

The `outbreak_setting` variable reflects the specific environments where outbreaks occur, such as long-term care homes (LTCH), hospitals (e.g., acute care or psychiatric wards), and retirement homes. As shown in Figure 5, LTCHs account for the highest number of outbreaks (37 events), while transitional care settings report the fewest (3 events). The elevated frequency of outbreaks in LTCHs highlights the vulnerability of their aging populations and the densely populated living conditions. Hospitals, particularly acute care wards, also exhibit notable outbreak counts (11 events). Including this variable captures the spatial distribution of outbreak risks, which is essential for resource prioritization and implementing targeted preventive measures. For instance, focusing on LTCHs with enhanced infection control and vaccination campaigns could mitigate their disproportionate outbreak burden. This predictor ensures that the analysis addresses key differences across institutional settings to design context-specific public health interventions effectively.

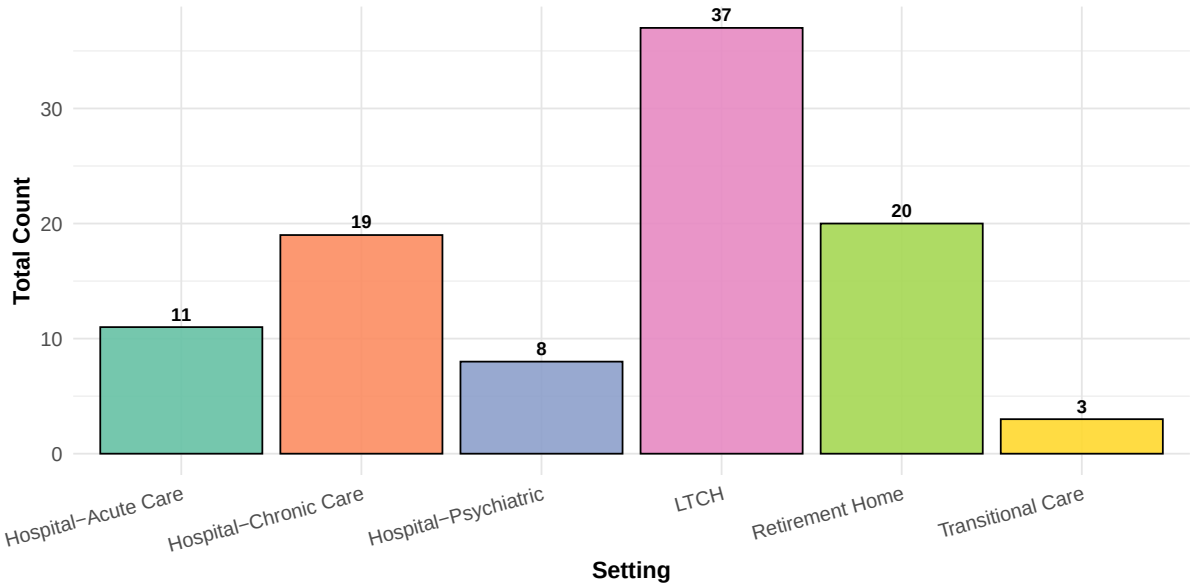

Figure 5: Outbreak Count by Setting

3 Model

The goal of our modelling strategy is to understand the factors influencing outbreak frequencies in Toronto healthcare institutions and to predict future outbreak trends. By employing Poisson regression models, we aim to examine how variables such as pathogen type, seasonality, month, and outbreak settings contribute to the observed counts of outbreaks. Interaction terms are included to explore whether temporal trends vary across pathogens and seasons, providing deeper insights into potential seasonal patterns and pathogen-specific behaviors. This approach enables us to generate predictions for future periods, such as 2025, which can inform public health planning and resource allocation while maintaining a model that balances interpretability and complexity.

In analyzing outbreak frequency, we developed multiple Poisson regression models to explore the effects of pathogens, time (`month` and `season`), and institutional settings (`outbreak_setting`) on outbreak rates. Starting with a baseline model focusing on main effects, we expanded the models by introducing interaction terms like pathogen with month, season with month to investigate dynamic relationships between variables. After comparing models, we identified that the Poisson interaction model, with significant pathogen-month interactions and the lowest AIC–507.71, best captured the monthly variation in outbreak frequency for specific pathogens such as Rhinovirus. Therefore, this model was selected as the optimal tool for studying the relationship between pathogens and months.

Diagnostics are included in Appendix B.

3.1 Model set-up

The model predicts the number of outbreaks in healthcare settings using the following predictor variables:

1. **Causative Agent**: Represents the primary pathogen identified in the outbreak (e.g., COVID-19, influenza). This variable captures pathogen-specific effects on outbreak counts.
2. **Season**: Denotes the season during which the outbreak occurred (Winter, Spring, Summer, Fall). It accounts for seasonal variations in respiratory disease transmission.
3. **Month**: Represents the calendar month when the outbreak began. It captures finer temporal patterns within seasons.
4. **Outbreak Setting**: Refers to the type of healthcare facility where the outbreak occurred (e.g., hospitals, long-term care facilities). This variable accounts for variations across healthcare contexts.

Define $ y_i$ as the number of outbreaks observed for i_{th} record. Let `causative_agent_1` represent the primary pathogen type, `month` denote the month the outbreak began, `season` denote the season, and `outbreak_setting` denote the type of facility.

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (1)\\
\log(\lambda_i) &= \beta_0 + \beta_1 \cdot \texttt{causative\_agent\_1}_i + \beta_2 \cdot \texttt{month}_i + \beta_3 \cdot \texttt{season}_i && (2)\\
&\quad + \beta_4 \cdot (\texttt{month}_i \times \texttt{season}_i) + \beta_5 \cdot \texttt{outbreak\_setting}_i && (3)\\
\beta_0 &\sim \text{Normal}(0, 10) && (4)\\
\beta_1, \beta_2, \beta_3, \beta_4, \beta_5 &\sim \text{Normal}(0, 10) && (5)
\end{aligned}
$$

Where:

- y_i represents the outbreak count for record i_{th}.
- λ_i is the expected count of outbreaks for record i_{th}.
- β_0 is the intercept, while β_1 through β_5 are coefficients describing the effects of the predictors.

This Poisson regression model uses a log-link function to ensure that the predicted outbreak counts λ_i remain positive. It incorporates main effects for pathogen type, month, season, and outbreak setting, as well as an interaction between month and season.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022).

3.2 Model justification

The Poisson regression model was chosen for its suitability in handling count data, such as outbreak frequencies in healthcare settings, where the dependent variable `outbreak_count` reflects non-negative integers. This aligns naturally with the properties of the Poisson distribution. The model captures the relationship between the frequency of outbreaks and key predictors, including `causative_agent_1`, `month`, `season`, and `outbreak_setting`, which were identified as relevant through a thorough examination of the dataset. These variables represent temporal, environmental, and pathogen-specific factors influencing outbreak patterns.

To account for interactions between variables, terms such as `month * season` and `month * causative_agent_1` were incorporated. These interaction terms allow the model to capture variations in outbreak frequencies across seasons and pathogen types over time. For example, some pathogens may show pronounced seasonality, while others may follow a steadier pattern. This choice was informed by observed trends in the data, where both temporal and pathogen-specific factors appeared to interact meaningfully.

The model was designed to balance interpretability and complexity. Variables such as `month` were included as numeric predictors to preserve temporal trends, while categorical variables like `season` and `outbreak_setting` were retained to reflect their discrete nature. This treatment ensures consistency with the definitions provided in the data section while facilitating a clear interpretation of the results. The model remains neither overly simplistic nor unnecessarily complex, making it appropriate for the given research question.

To further evaluate the model's adequacy, a Reduced Model was created through variable selection by systematically excluding non-significant predictors to simplify the framework. However, comparison results indicated in Table 1 that the Poisson Season Time Model outperformed the Reduced Model, evidenced by a lower AIC (449.29 vs. 508.79) and RMSE (2.60 vs. 4.43). These findings underscore the importance of retaining interaction terms like `month * season` and pathogen-specific variables to effectively capture temporal and outbreak-specific dynamics. The Poisson Season Time Model was thus deemed the most robust and informative framework for understanding outbreak patterns.

Model diagnostics, including residual analysis and dispersion tests, were performed to validate the assumptions of the Poisson model. Overdispersion was addressed by testing alternative models such as the negative binomial regression, but the Poisson model was ultimately retained due to its interpretability, stability, and alignment with the dataset. The model was implemented in R (R Core Team 2023) using the `glm()` function, and predictions for outbreak frequencies in 2025 further demonstrate its potential utility for public health planning. While assumptions such as independent events and the exclusion of potential confounders like population density are acknowledged as limitations, the model provides a robust framework for analyzing outbreak patterns.

Table 1: Comparison of full and reduced model

Metric	Full.Model	Reduced.Model
AIC	449.29	508.79
BIC	490.65	539.81
RMSE	2.60	4.43
Likelihood Ratio Test (p-value)	0.00	N/A

4 Results

The analysis employed a Poisson regression model to investigate the relationship between the outcome variable (`outbreak_count`) and key predictor variables, including pathogen type, month, season, and outbreak setting. The fourth section presents detailed results on the associations between these variables and outbreak frequency, as described in Section 4.1. The model highlights significant relationships, such as the notable influence of specific pathogens like COVID-19 and outbreak settings like long-term care homes on the frequency of outbreaks.

Additionally, in Section 4.2, the model was used to predict potential outbreak trends for 2025. These predictions assume consistent patterns in the predictor variables and similar environmental conditions to those observed in 2024. However, the predictions are limited by the exclusion of other potential influencing factors, such as changes in public health interventions, population dynamics, or pathogen evolution. The more particular information see in Section 5.

4.1 Model results

Our model summary result are summarized in Table 2. The Poisson regression model demonstrates significant relationships between outbreak counts and several predictors. COVID-19 is strongly associated with higher outbreak counts ($\beta = 1.57$, $p < 0.001$), highlighting its dominant role compared to other pathogens, such as Parainfluenza and Rhinovirus, which show no significant associations. Seasonal effects are notable, with Spring and Summer significantly reducing outbreak counts compared to Fall ($\beta =$ -5.21) and ($\beta =$ -4.95, $p < 0.001$). The interaction between month and season further refines these effects, showing that seasonal impacts vary across months.

Outbreak settings also contribute significantly; long-term care homes (LTCHs) show the highest association with outbreak counts ($\beta = 1.64$, $p < 0.001$), while transitional care settings exhibit significantly lower outbreak frequencies ($\beta =$ -1.64, $p = 0.006$). The model achieves a good fit with an AIC of 449.29, a BIC of 490.6, and a log-likelihood of -208.64. While the RMSE of 2.6 indicates some variability in the data, the overall model performance ($F = 35.11$) suggests strong predictive capacity.

Table 2: Summary of the chosen model

[!h]

	Model Summary
	(1)
(Intercept)	3.135
	(1.010)
causative_agent_1COVID-19	1.568
	(0.164)
causative_agent_1Parainfluenza	0.025
	(0.208)
causative_agent_1Rhinovirus	0.038
	(0.213)
month	−0.326
	(0.100)
seasonSpring	−5.211
	(1.092)
seasonSummer	−4.948
	(1.194)
seasonWinter	−1.561
	(1.016)
outbreak_settingHospital-Chronic Care	0.320
	(0.183)
outbreak_settingHospital-Psychiatric	−0.806
	(0.296)
outbreak_settingLTCH	1.644
	(0.155)
outbreak_settingRetirement Home	0.514
	(0.176)
outbreak_settingTransitional Care	−1.636
	(0.597)
month × seasonSpring	0.761
	(0.147)
month × seasonSummer	0.575
	(0.138)
month × seasonWinter	−0.706
	(0.216)
Num.Obs.	98
AIC	449.3
BIC	490.6
Log.Lik.	−208.644
RMSE	2.60

Figure 6 illustrates the coefficient estimates and their 95% confidence intervals for the Poisson regression model predictors, highlighting the direction, magnitude, and statistical significance of their effects on outbreak counts.

Causative Agent Effects

Among the pathogens, COVID-19 has a strongly positive and statistically significant association with outbreak counts, as its confidence interval does not overlap zero. This finding underscores the dominant role of COVID-19 in driving outbreaks compared to the reference pathogen. In contrast, Rhinovirus and Parainfluenza show smaller coefficients with confidence intervals that include zero, indicating no statistically significant effects relative to the reference pathogen. These results emphasize the disproportionately high burden of COVID-19 on outbreak frequencies.

Seasonal Effects

The seasonal predictors reveal distinct trends. Both Spring and Summer exhibit negative coefficients with confidence intervals that exclude zero, suggesting significant decreases in outbreak counts compared to the reference season (Fall). Winter, while also having a negative coefficient, shows a confidence interval overlapping zero, implying that its effect is not statistically significant. These findings highlight the seasonal variability in outbreaks, with Fall serving as a high-risk period for certain pathogens.

Temporal Trends

The coefficient for month is negative and statistically significant, indicating a general decline in outbreak counts as the year progresses. This temporal trend likely reflects the waning of seasonal pathogen transmission cycles and shifts in public health interventions or behaviors later in the year. Interactions between month and seasons reveal more nuanced patterns: Month:Spring and Month:Summer have positive and statistically significant coefficients, suggesting increases in outbreaks over the months during these seasons. Conversely, Month:Winter shows a significant negative effect, reflecting a decline in outbreaks over the course of the winter months.

Setting Effects

Outbreak settings significantly influence the frequency of outbreaks. Long-Term Care Homes (LTCHs) have a positive and statistically significant coefficient, reflecting their heightened vulnerability due to factors like densely populated environments and vulnerable residents. Hospital-Chronic Care and Retirement Homes also show positive coefficients, but their confidence intervals marginally overlap zero, indicating borderline significance. On the other hand, Hospital-Psychiatric and Transitional Care settings have negative coefficients with confidence intervals overlapping zero, suggesting no statistically significant effects relative to the reference setting.

Implications

The results reveal important drivers of outbreak frequencies, including the dominant role of COVID-19, seasonal variability, and the heightened risk associated with specific settings like LTCHs. These findings underscore the need for tailored public health strategies, such as increased surveillance during high-risk seasons like Fall, year-round measures for COVID-19, and targeted interventions in vulnerable settings. The significance of temporal and seasonal interactions highlights the importance of adapting outbreak responses to both seasonal and monthly trends.

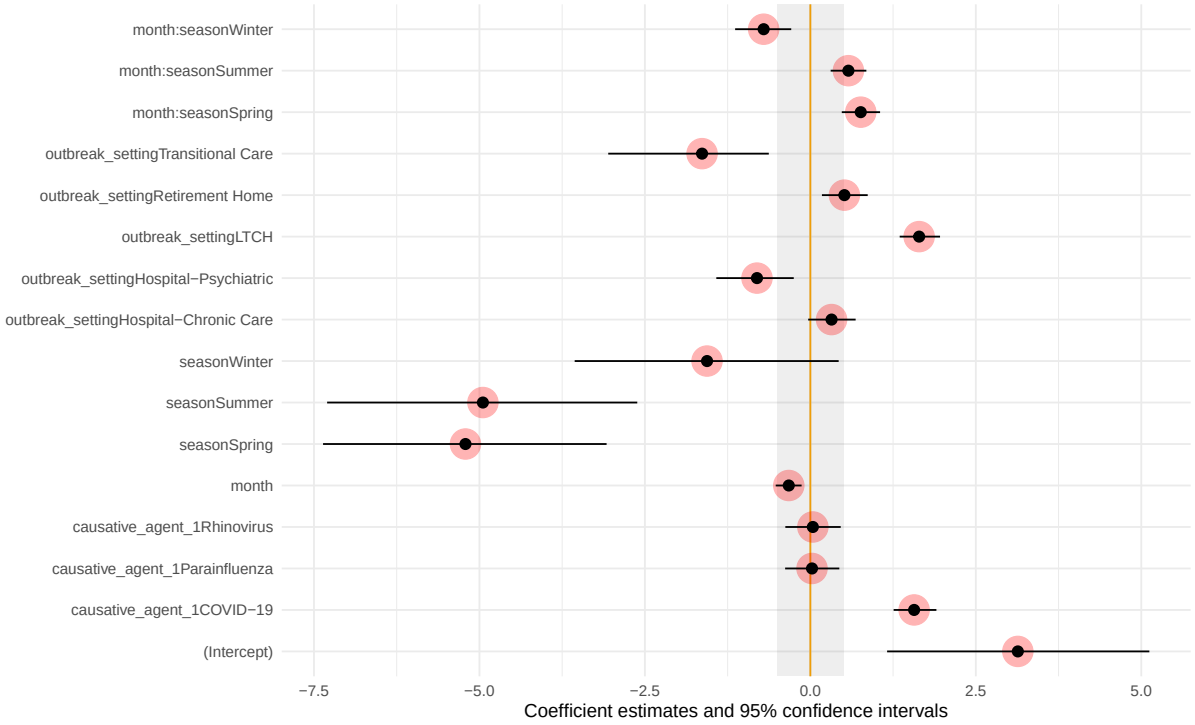

Figure 6: Summary of the confidence interval for predict variable

Figure 7 shows the predicted outbreak counts for different pathogens `causative_agent_1` across months and seasons, highlighting distinct temporal patterns.

Coronavirus: Outbreak counts are higher during the winter months (January and February) and decline sharply in spring and summer, consistent with the seasonality of respiratory viruses that thrive in colder conditions.

COVID-19: Unlike other pathogens, COVID-19 maintains relatively stable outbreak counts throughout the year, with minor fluctuations. This suggests that its transmission is less affected by seasonal variations, reflecting its unique epidemiological dynamics.

Parainfluenza: Parainfluenza outbreaks peak in late summer (August and September) and drop significantly during the winter, showing a strong seasonal pattern likely influenced by

environmental factors.

Rhinovirus: Rhinovirus shows bimodal peaks in early spring (March) and late summer (August), which align with known seasonal transitions favoring its transmission.

In summary, the model reveals that while some pathogens exhibit clear seasonality, others, such as COVID-19, are less dependent on seasonal factors. These insights can inform targeted, pathogen-specific intervention strategies for outbreak control.

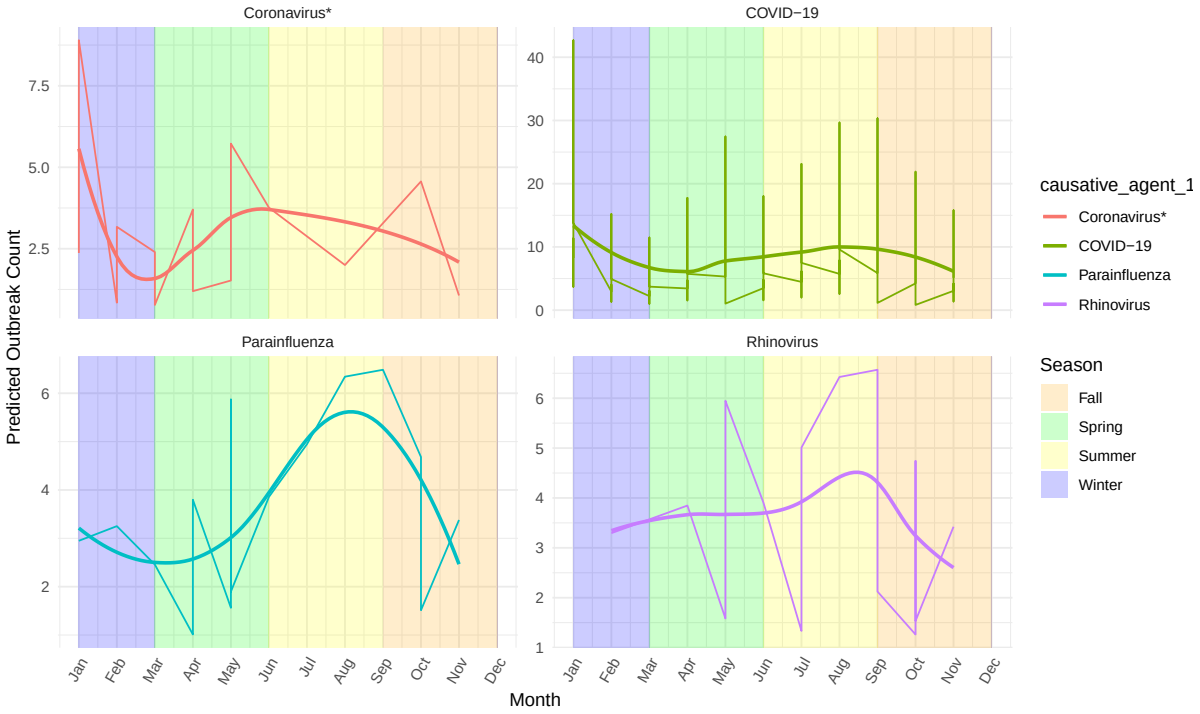

Figure 7: Outbreak Frequency by Pathogen and Month with Seasonal Effects

4.2 Model prediction results

The 2025 predicted outbreak frequencies reveal distinct temporal patterns across different pathogens. From Figure 8, Coronavirus* exhibits a gradual decline in outbreak counts from winter through summer, consistent with its seasonally driven nature. In contrast, COVID-19 maintains a steady outbreak frequency throughout the year, with only minor fluctuations, suggesting limited sensitivity to seasonal variation. Parainfluenza is projected to peak in late summer and early fall, reflecting established seasonal transmission patterns. Similarly, Rhinovirus demonstrates a bimodal distribution, with notable peaks in spring and late summer, aligning with its known epidemiological behavior. These pathogen-specific predictions underline the necessity of targeted strategies to address seasonal outbreaks effectively.

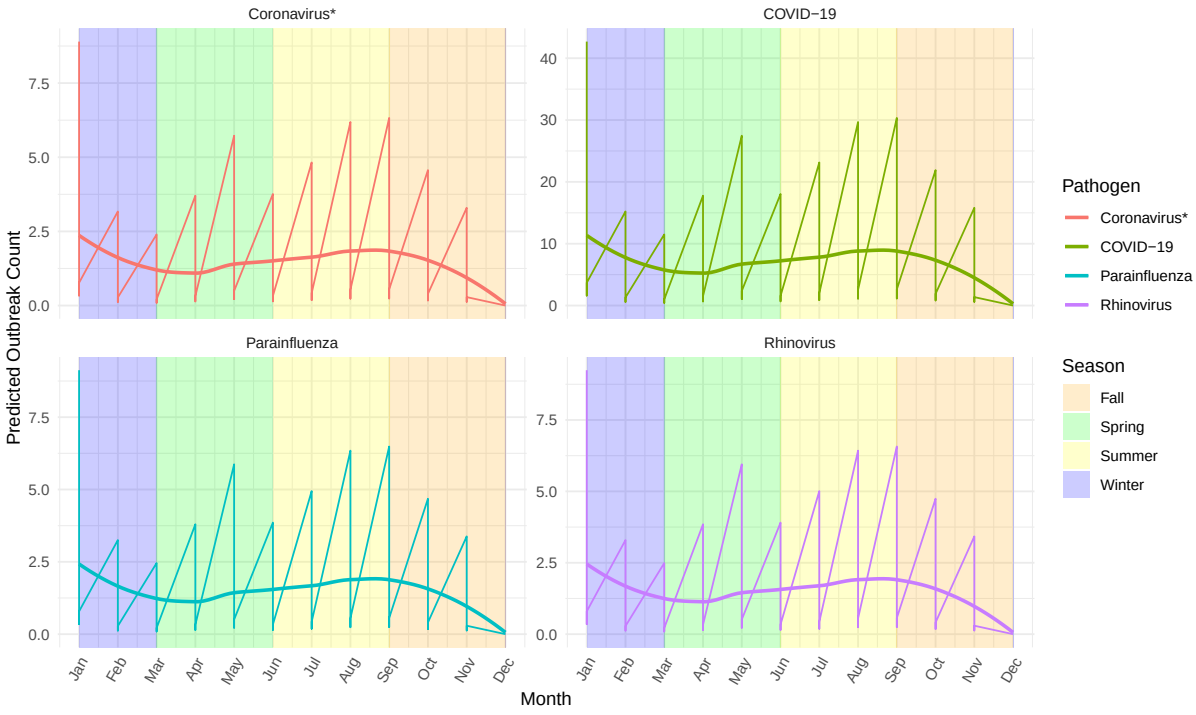

Figure 8: Predicted Outbreak Frequency by Pathogen in 2025

The Figure 9 shows comparison between actual 2024 outbreak data and model predictions highlights strong alignment in capturing overall patterns, though some deviations are observed. COVID-19 predictions align closely with actual frequencies, showcasing consistency and robustness in modeling. For Parainfluenza, the model effectively captures the general seasonal peak but slightly overestimates outbreaks during summer and underestimates them in fall. Rhinovirus predictions accurately reflect its bimodal distribution, though discrepancies occur during spring and summer months. For Coronavirus*, the model captures the downward trend but underestimates outbreak counts in early months. Overall, the model demonstrates its capacity to replicate key seasonal and temporal dynamics, providing a reliable basis for future projections while highlighting areas for potential refinement in capturing finer nuances.

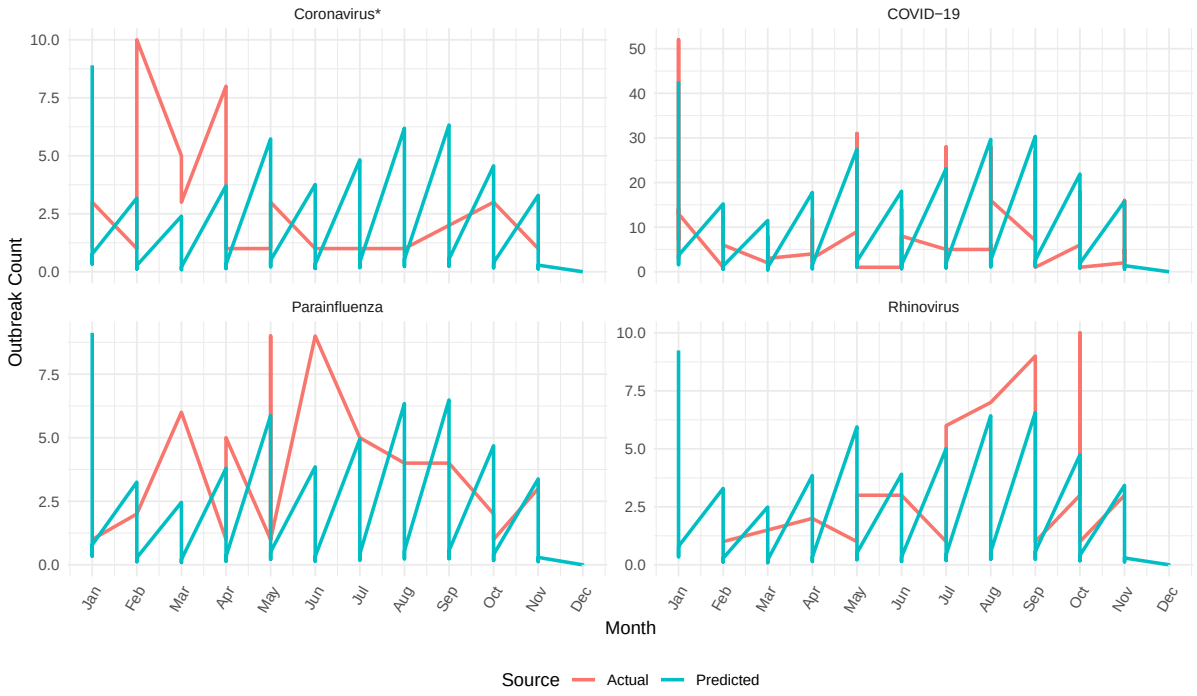

Figure 9: Comparison of Actual and Predicted Outbreak Frequencies

5 Discussion

This study employed a Poisson regression model to analyze seasonal and pathogen-specific outbreak patterns and predict outbreak frequencies for four pathogens in 2025. The analysis revealed distinct seasonal trends, particularly for Rhinovirus and Parainfluenza, while COVID-19 exhibited relatively consistent transmission rates throughout the year. Comparing the 2025 predictions with the actual outbreak data from 2024 demonstrated the model's ability to capture key temporal and seasonal patterns. However, deviations for smaller-scale pathogens highlighted potential areas for refinement, underscoring the need for enhanced modeling approaches to address variability in outbreak dynamics.

5.1 Seasonal Sensitivity of Pathogens

The findings revealed marked seasonal sensitivity for pathogens like Rhinovirus and Parainfluenza, which showed distinct peaks in spring and late summer to early fall, respectively. In contrast, COVID-19 displayed a near-steady transmission rate across the year, reflecting a lack of climatic influence. This consistency aligns with guidance from the World Health Organization, which notes that COVID-19 transmission is driven by human-to-human contact rather

than environmental factors Organization (n.d.). Rhinovirus and Parainfluenza, however, exhibited seasonality driven by factors such as weather transitions, increased indoor activity, and environmental stability favorable to these pathogens.

The observed seasonal variations emphasize the need for targeted interventions during high-risk periods. For example, scaling vaccination campaigns before Rhinovirus peaks in spring or bolstering healthcare resources for Parainfluenza in late summer could effectively reduce outbreaks. Conversely, COVID-19's steady transmission underscores the necessity of maintaining year-round public health measures such as vaccination drives, consistent mask use, and robust surveillance systems.

5.2 Influence of Outbreak Settings

The analysis showed substantial variation in outbreak frequencies across settings, with long-term care homes (LTCHs) reporting the highest outbreak counts. This finding underscores the vulnerability of residents in these facilities, attributed to factors such as shared living spaces, advanced age, and pre-existing health conditions. COVID-19, in particular, demonstrated disproportionately high frequencies in LTCHs, highlighting the need for rigorous infection control measures, including regular testing, enhanced hygiene protocols, and vaccination prioritization in these facilities.

In contrast, lower outbreak frequencies were observed in settings such as transitional care centers, suggesting that operational protocols, population density, or resident health status might play a mitigating role. These findings call for tailored public health strategies. For LTCHs, enhanced access to healthcare resources and strict infection control practices are critical. In lower-risk settings, routine monitoring and flexible outbreak response protocols could be sufficient to manage risks without overburdening resources.

5.3 Weaknesses and next steps

Despite its strengths, this study has several limitations. First, the focus on four pathogens restricts the generalizability of the findings. Outbreak data relied on administrative reporting, which may be incomplete or inconsistent, potentially skewing the model's predictions. Moreover, the dominance of COVID-19 in the dataset may have overshadowed smaller-scale pathogens, reducing their statistical significance and limiting the model's ability to account for their unique dynamics. This imbalance highlights the need to construct datasets that better represent diverse pathogen behaviors.

The predictions for 2025 also assume static conditions based on 2024 data, making them less adaptable to emerging pathogens or shifts in environmental or public health contexts. External factors such as population density, climatic changes, or new interventions (e.g., novel vaccines) were not incorporated, potentially affecting predictive accuracy.

Future research should address these limitations by expanding the scope of the dataset to include additional pathogens, longer time periods, and diverse geographic regions. Incorporating environmental variables such as temperature and humidity, along with metrics for public health interventions, could improve model precision. Additionally, examining the impact of targeted strategies, such as vaccination campaigns or changes in healthcare policies, would provide actionable insights for outbreak management. This expanded approach could better inform public health planning, ensuring preparedness for both established and emerging pathogens.

A Appendix

A.1 Optimized Methodology of Data Collection and Reporting

Outbreak Data Framework and Governance

The dataset utilized in this study centers on outbreak data collected from healthcare institutions in Toronto, following stringent regulatory guidelines. The data collection is governed by the Ontario Health Protection and Promotion Act (HPPA) Government (1990), which mandates healthcare institutions—including hospitals, long-term care homes, and retirement homes—to monitor and report signs of respiratory and gastroenteric infections among staff and residents. An outbreak is defined as an unusual increase in infection rates beyond the expected baseline within a specific institution or ward. Reporting is carried out by the institutions to Toronto Public Health (TPH), which investigates the reported outbreaks and collaborates with institutions to implement control measures Toronto (2024).

Population, Framework, and Sampling Coverage

The dataset includes all reported outbreaks from Toronto's healthcare institutions, representing a near-complete population of healthcare-associated outbreaks. The framework, defined by HPPA requirements, ensures comprehensive coverage of outbreaks within institutions mandated to report. The data essentially serves as a full census of reported outbreaks rather than a traditional sample, as all institutions are legally obligated to report under HPPA. However, certain cases are excluded, such as mild infections managed outside healthcare settings or individuals unable to seek timely medical care, resulting in potential underrepresentation.

Reporting Process and Data Collection

Outbreaks are identified and reported by healthcare institutions based on clinical symptoms or laboratory-confirmed cases. These reports are then submitted to Toronto Public Health, with updates provided weekly to maintain real-time relevance Ministry of Long-Term Care (2024). However, smaller outbreaks or those diagnosed late may be underreported, leading to data gaps. Variability in reporting practices across institution types further complicates data consistency; for instance, long-term care homes may report more frequently due to heightened surveillance requirements, whereas transitional care centers might report less rigorously.

Observational Nature of the Data

This dataset is observational, relying on mandatory administrative reporting rather than experimental sampling or randomized control. Its strength lies in the extensive coverage across diverse healthcare settings. However, inconsistencies in reporting practices between institutions can introduce biases. For example, retirement homes may demonstrate heightened vigilance in outbreak reporting compared to hospitals, potentially skewing the dataset's representation of outbreak frequencies.

Additionally, the observational nature of the data does not account for unmeasured confounders such as population density, healthcare access, or staff-to-patient ratios, which may influence outbreak rates. Temporal gaps may also arise from unreported or delayed diagnoses, particularly during holidays or periods of resource strain. Unlike randomized datasets, observational data lacks randomization, leading to potential selection biases as institutions with better reporting practices are more prominently represented. Despite these limitations, this dataset provides invaluable insights into naturally occurring outbreak trends, enabling analyses that reflect the complexities of real-world healthcare settings.

Strengths and Limitations of the Dataset

The dataset offers notable strengths that support its use for analyzing outbreak patterns. Comprehensive temporal coverage, with detailed outbreak timelines, facilitates longitudinal analyses to identify trends such as seasonal or monthly peaks. Granular data, including variables like outbreak duration, pathogen type, and affected settings, allows for nuanced exploration of interactions between key factors. Additionally, the dataset's foundation in legally mandated reporting ensures broad coverage of significant outbreaks, reducing the likelihood of selective underreporting. Its utility for public health planning is enhanced by these attributes, supporting targeted interventions such as vaccination campaigns during high-risk periods.

However, the dataset is not without limitations. Cases involving mild infections managed outside healthcare settings or resolved without public health involvement are often excluded, leading to underreported outbreak frequencies. Inconsistent reporting practices across institutions further complicate analyses, as long-term care homes may report more consistently than hospitals or transitional care facilities. Delays in confirming or reporting outbreaks also create potential data gaps. Furthermore, errors in diagnosis or administrative reporting can introduce inaccuracies, while the dataset's geographic scope is limited to Toronto, restricting the generalizability of findings to other regions.

Reporting Standards and Data Quality

The dataset adheres to rigorous reporting standards mandated by the HPPA Government (1990), ensuring uniformity in data collection across institutions. Outbreaks are defined using standardized criteria, such as an increase in infection rates above expected baselines within a specific ward or institution. This ensures consistency in reported data. Institutions work closely with Toronto Public Health (TPH) to verify and classify outbreaks, further enhancing the dataset's reliability.

Real-time updates are a key strength, allowing for near-immediate tracking of outbreak trends. However, administrative lags or delays in outbreak confirmation may temporarily affect data completeness. The granularity of the data, including variables like outbreak start and end dates, pathogen type, and institution settings, enables detailed analyses but increases the risk of errors during data entry, especially in high-pressure environments like hospitals. Differences in diagnostic resources across institutions can also lead to variability, with smaller or resource-limited facilities potentially underreporting certain pathogens. Furthermore, institutions with

robust surveillance systems may report more frequently, creating an overrepresentation of outbreaks in these settings.

Despite these challenges, the dataset provides a reliable foundation for analyzing healthcare-associated outbreaks. By balancing its strengths—such as high granularity and comprehensive coverage—against its limitations, it remains a valuable tool for informing public health policy and outbreak preparedness strategies.

Connection to Literature and Data Applications

The dataset supports broader public health efforts to manage and mitigate healthcare-associated infections, aligning with the objectives outlined in the 2023-2024 Canadian Public Health Plan Canada (2024). The outbreak monitoring framework is consistent with international practices, such as the World Health Organization's guidelines for disease surveillance and transmission Organization (n.d.). By linking local data collection to global health standards, this dataset highlights the critical role of systematic outbreak reporting in improving public health outcomes and guiding evidence-based policy decisions.

A.2 More details about plot

The Figure 10 showcases the distribution of outbreak counts, revealing a heavily skewed pattern where most instances report lower outbreak counts. The majority of the data lies in the range of 0–2 outbreaks, indicating that many facilities or contexts experienced minimal disruptions. As the outbreak count increases, the frequency declines sharply, highlighting the rarity of high-count outbreaks. Notably, there are a few outlier cases, such as counts above 30 or even 50, which likely correspond to specific high-risk pathogens (e.g., COVID-19) or settings like long-term care homes. These extreme cases underline the need for targeted interventions in high-risk contexts.

In contrast, the Figure 10 aggregates all data without considering the influence of key predictors like pathogens or seasons. While it gives a general distribution of outbreak counts, it lacks the depth needed to examine contextual factors that affect outbreak patterns. This makes it less suitable for understanding the drivers of variation in outbreak frequencies across multiple factors. By visualizing outbreaks grouped by pathogen and season in Figure 1, we can identify trends or anomalies, such as seasonal peaks for specific pathogens, which are crucial for drawing actionable conclusions and targeting interventions.

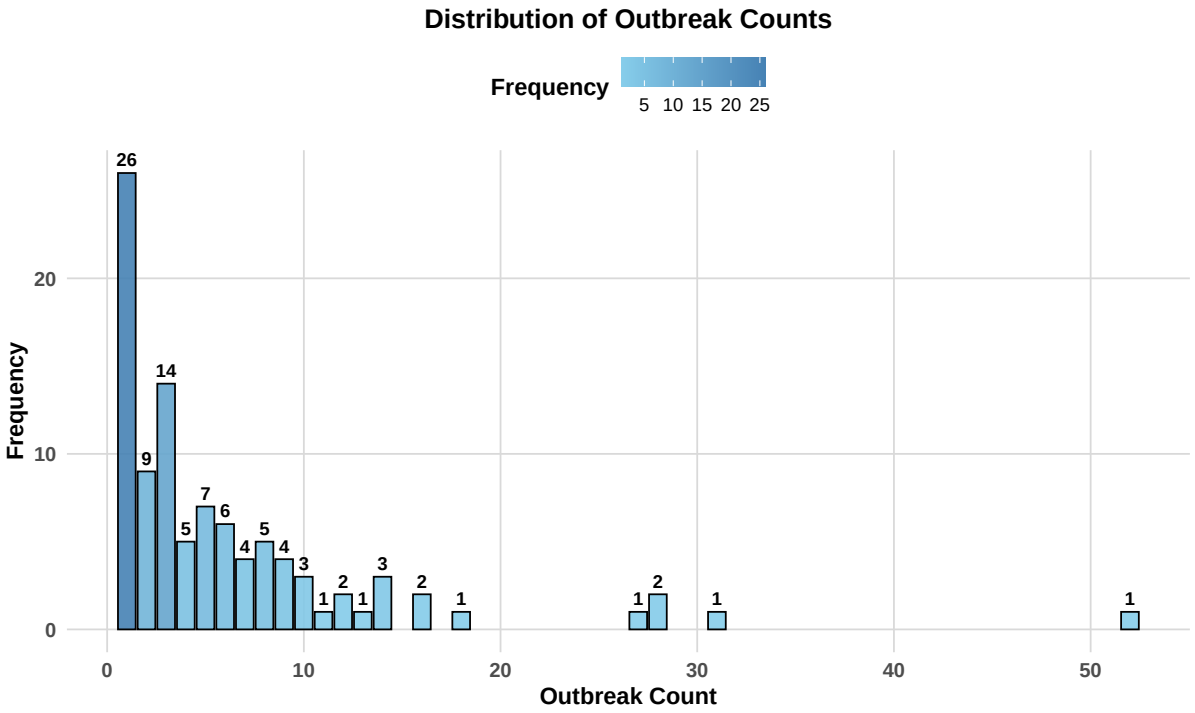

Figure 10: Distribution of outbreak Count

B Model details

The diagnostic plots Figure 11 provide insights into the goodness-of-fit and assumptions of the Poisson regression model:

Residuals vs. Fitted: This plot reveals a moderate deviation from randomness, suggesting potential non-linearity or issues with model fit. The points generally cluster around zero, but certain observations (e.g., #14) exhibit higher residuals, indicating possible outliers or influential points.

Q-Q Plot of Residuals: The Q-Q plot demonstrates that the residuals mostly follow a theoretical normal distribution, though slight deviations are observed in the tails. Observation #14 again emerges as an outlier, potentially impacting the model fit.

Scale-Location Plot: The Scale-Location plot shows variability in the spread of residuals across predicted values. The slight upward trend suggests heteroscedasticity, where variance increases with higher fitted values, which may indicate a limitation of the Poisson model in capturing variance accurately.

Cook's Distance: The Cook's distance plot identifies influential observations, particularly #14 and #63, which have a disproportionate impact on model coefficients. These points

warrant further examination to determine if they represent true outliers or errors in data collection.

Residuals vs. Leverage: This plot highlights observation #63 as a high-leverage point with substantial influence on the regression line, as indicated by its position above the Cook's distance threshold. Such points can disproportionately affect the model and may require additional attention or sensitivity analysis.

Cook's Distance vs. Leverage: This combined plot reiterates the influence of high-leverage points like #14 and #63, which appear beyond acceptable boundaries for leverage and Cook's distance, reinforcing the need to assess these observations carefully.

In summary, the diagnostics suggest that while the model captures the overall trend reasonably well, there are notable concerns with influential observations (#14 and #63) and mild heteroscedasticity. These findings indicate the need for potential model refinement, such as considering alternative distributions (e.g., negative binomial) or investigating data anomalies.

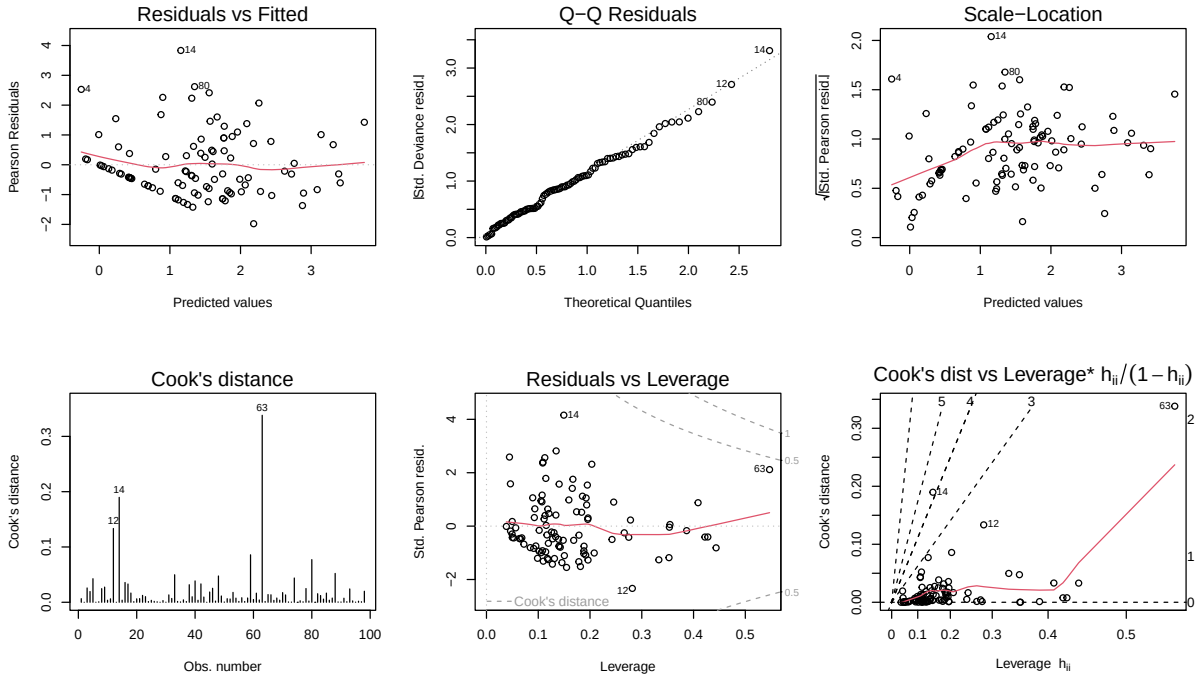

Figure 11: Diagnostic Plots of model

References

Alexander, Rohan. 2023a. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

———. 2023b. *Telling Stories with Data: With Applications in r*. Chapman; Hall/CRC.

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.

Canada, Government of. 2024. "Public Health Agency of Canada 2023-2024 Departmental Plan." https://www.canada.ca/en/public-health/corporate/transparency/corporate-management-reporting/reports-plans-priorities/2023-2024-departmental-plan.html.

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/.

Government, Ontario. 1990. "Health Protection and Promotion Act (HPPA), r.s.o. 1990." https://www.ontario.ca/laws/statute/90h07.

Ministry of Long-Term Care, Ontario. 2024. "Ministry of Long-Term Care." https://www.ontario.ca/page/ministry-long-term-care.

Müller, Kirill. 2020. *here: A Simpler Way to Find Your Files*. https://CRAN.R-project.org/package=here.

Organization, World Health. n.d. "Coronavirus Disease (COVID-19): How Is It Transmitted?" https://www.who.int/news-room/questions-and-answers/item/coronavirus-disease-covid-19-how-is-it-transmitted.

R Core Team. 2023. *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.

Toronto, City of. 2024. "Active Outbreaks in Toronto Healthcare Institutions." https://www.toronto.ca/community-people/health-wellness-care/health-inspections-monitoring/active-outbreaks-in-toronto-healthcare-institutions/.

Wickham, Hadley. 2016. *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

———. 2023. *modelr: Modelling Functions that Work with the Pipe*. https://CRAN.R-project.org/package=modelr.

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.

Xie, Yihui. 2023. *knitr: A General-Purpose Package for Dynamic Report Generation in R*.

ttps://yihui.org/knitr/.
Zhu, Hao. 2024. *kableExtra: Construct Complex Table with 'kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.