My title*

My subtitle if needed

First author Another author

January 22, 2024

This paper presents a comprehensive analysis of outbreak data from long-term care homes in 2023. The key finding reveals a significant correlation between outbreak durations and the type of causative agents involved. This insight is crucial for healthcare policy planning and management of future outbreaks in similar settings.

1 Introduction

Long-term care homes (LTCHs) serve a pivotal role in the healthcare system, especially in catering to the needs of the elderly and those with chronic health conditions. However, these facilities are often hotspots for infectious disease outbreaks, presenting a significant challenge to public health and patient safety. In this paper, we delve into a comprehensive analysis of data pertaining to outbreaks in LTCHs during the year 2023. Our focus is primarily on identifying the causative agents behind these outbreaks, analyzing the duration over which they persisted, and evaluating their overall impact on the health and wellbeing of residents and staff.

This investigation is critical, as it sheds light on the dynamic nature of infectious diseases within LTCH environments. By examining the specifics of these outbreaks, including the types of pathogens involved and their behavior in a long-term care setting, we gain valuable insights into how these diseases spread and persist. This understanding is fundamental in enhancing the quality of care provided to vulnerable populations and in devising robust preventive strategies aimed at mitigating similar incidents in the future.

To facilitate a comprehensive understanding, the remainder of this paper is organized into several key sections. Firstly, we present a thorough analysis of the outbreak data, which includes a breakdown of the types of causative agents identified, the duration of each outbreak, and an assessment of the subsequent impacts. Following this, we engage in a detailed discussion

*Code and data are available at: LINK.

of our findings, exploring their significance and the potential implications for healthcare policy and practice in LTCHs. In the final sections, we draw conclusions from our study, highlighting the key takeaways and their relevance to the broader context of public health. Additionally, we outline avenues for future research, emphasizing areas where further investigation could contribute to a deeper understanding and better management of infectious disease outbreaks in long-term care settings.

2 Data

2.1 Data Collection

The dataset sourced from Open Data Toronto (Gelfand 2022), is a collection of data on reported outbreaks, presumably compiled by a public health department or a similar entity. The primary aim of this dataset is to monitor public health concerns, assist in policy making, and inform the public, fitting within a broader initiative by governmental or health organizations to track health-related incidents. As Gebru et al. discuss in their work on datasheets for datasets (Gebru et al. 2021), the structure and transparency of dataset documentation are crucial for effective use in research and policy development. The data in this case is structured in a way that each row represents a reported outbreak, with columns detailing identifiers, institution names and addresses, outbreak settings, types, causative agents, dates of occurrence, and the active status of each outbreak. It includes mostly categorical data, such as the types of outbreak and causative agents, and date fields for the outbreak timelines. This kind of data categorization can be pivotal in statistical analysis within R (R Core Team 2022).

The dataset sourced from Open Data Toronto Gelfand (2022), is a collection of data on reported outbreaks, presumably compiled by a public health department or a similar entity. The primary aim of this dataset is to monitor public health concerns, assist in policymaking, and inform the public, fitting within a broader initiative by governmental or health organizations to track health-related incidents. The data is structured in a way that each row represents a reported outbreak, with columns detailing identifiers, institution names and addresses, outbreak settings, types, causative agents, dates of occurrence, and the active status of each outbreak. It includes mostly categorical data, such as the types of outbreak and causative agents, and date fields for the outbreak timelines.

2.2 Variables

The dataset, comprising of key variables, offers a comprehensive picture of outbreak occurrences in Toronto. The Institution Name variable provides critical insights into the geographical spread and institutional vulnerability to outbreaks. Specifically, analyzing these variables can reveal patterns in outbreak occurrences across different regions and institution types, highlighting areas or institutions that may be more susceptible to health crises. Crucial to understanding the spread and control of diseases is the Causative Agent-1 variable. It sheds light on the pathogens or factors responsible for the outbreaks, essential for tracking specific disease spread, identifying emerging health threats, and formulating responsive strategies. The temporal variable, Date Outbreak Began, allow for an analysis of outbreak duration and the identification of seasonal patterns or trends, which is vital for future preparedness and preventive measures. Lastly, the Active status of each outbreak provides immediate information on current public health challenges, enabling swift responses and resource allocation to active health threats. There are other varibles that could be produced from the data set, but utilmetley the ones that were chosen are the most important.

2.3 Data Processing

The process of handling and analyzing the dataset in R (R Core Team 2022) involves a comprehensive series of steps, beginning with the loading of essential libraries and culminating in the extraction of actionable insights. Initially, libraries such as tidyverse (Wickham 2023) for data manipulation and visualization, lubridate (Spinu, Grolemund, and Wickham 2023) for handling date-time data, and stringi for string operations are loaded. These libraries provide a robust toolkit for various data processing tasks.

The dataset is then read into R (R Core Team 2022) using the read_csv function from the readr package, a part of tidyverse (Wickham 2023). This function efficiently converts the CSV file into a dataframe, R's (R Core Team 2022) fundamental data structure for handling tabular data. Once loaded, the data undergoes an initial inspection using functions like head, summary, and str to understand its structure, identify any inconsistencies, missing values, or incorrect data types.

Data cleaning is a crucial next step where issues identified during inspection are addressed. This may involve handling missing values, filtering out irrelevant data, or converting data into correct formats, such as using lubridate (Spinu, Grolemund, and Wickham 2023) to parse and format dates properly. Additionally, new variables that could provide further insights might be created from the existing data.

Following the cleaning process, data transformation is performed, primarily using the dplyr (Wickham et al. 2023) package. This could include grouping the data based on certain variables like Outbreak Setting and summarizing or calculating statistics for each group. Textual data, such as institution names and addresses, are manipulated as needed using stringi

Finally, the processed data or the results of the analysis are exported for reporting or further use. This can be in the form of CSV files using write_csv.

This entire process transforms the raw data into meaningful and actionable insights, essential for informed decision-making and policy formulation, especially in fields like public health as evident in this dataset. Each step, from initial reading to final exporting, adds layers of

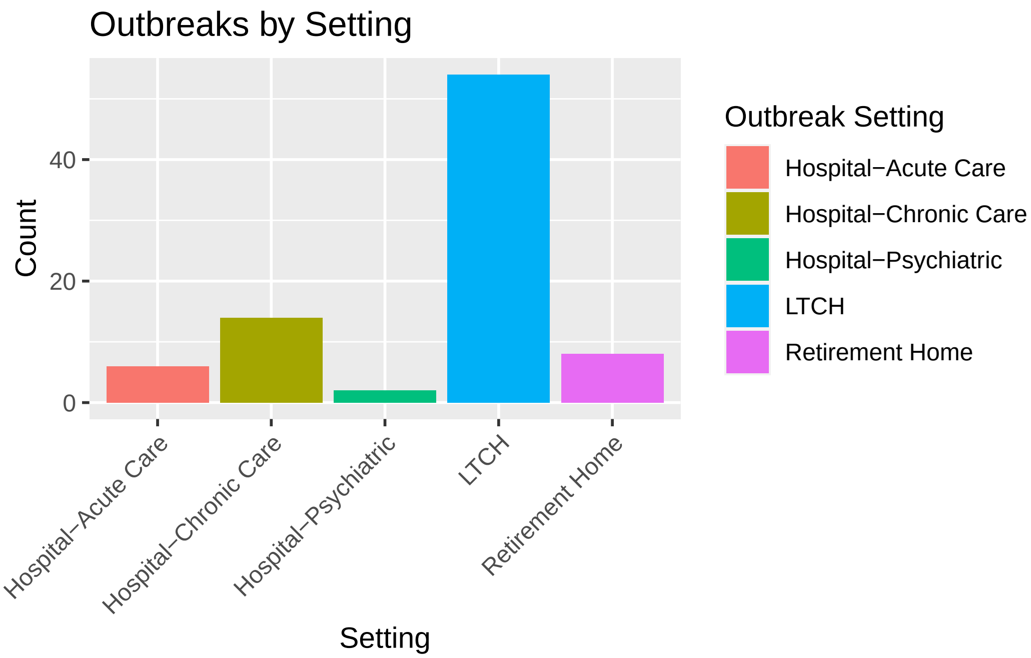
Table 1: Outbreaks in Toronto Healthcare Institutions - First 10 Rows

Institution_Name	Causative_Agent_1	Date_Outbreak_Began	Active
Willow Nursing Home	COVID-19	2023-06-01	N
Willow Hospital	Influenza A	2023-09-07	Y
Elm Nursing Home	Norovirus	2023-11-24	N
Birch Care Center	Influenza B	2023-07-30	N
Elm Care Center	Norovirus	2023-06-04	Y
Cedar Nursing Home	Influenza A	2023-06-08	N
Cedar Hospital	Norovirus	2023-06-11	N
Birch Hospital	Rhinovirus	2023-06-16	N
Elm Care Center	Influenza B	2023-06-12	Y
Oak Nursing Home	Norovirus	2023-05-03	N

understanding and value to the raw data, culminating in a data-driven foundation for strategic actions.

3 Visualizing the Data and the Results

3.1 Distribution of Outbreaks by Setting

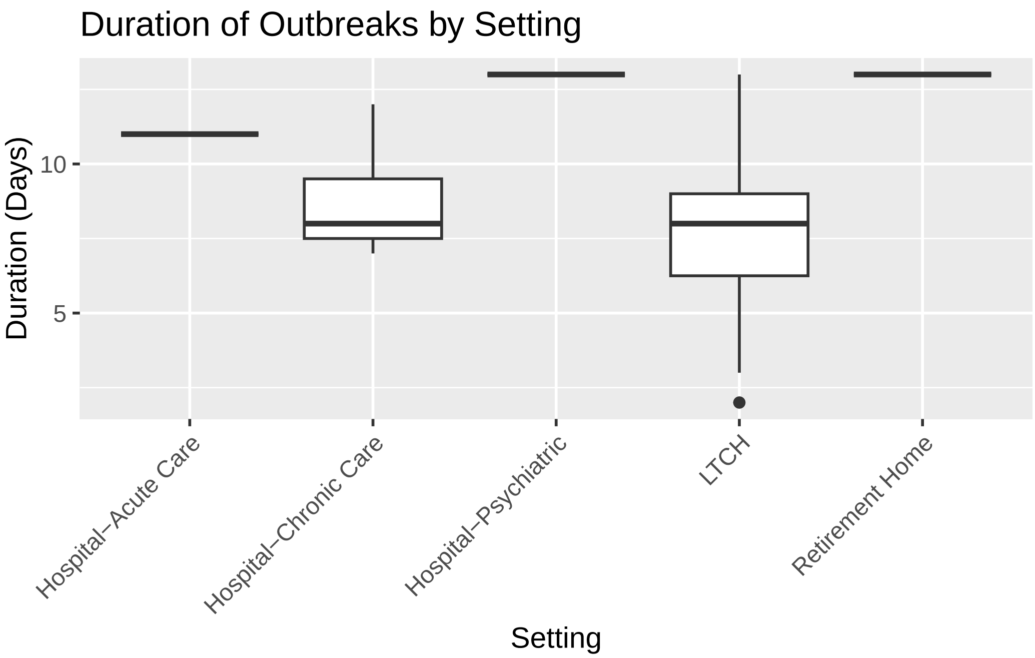


This chart is a pivotal visualization in understanding the landscape of outbreak occurrences across various environments or settings within the dataset. This chart categorizes and displays the frequency of outbreaks in distinct settings such as hospitals, long-term care homes, schools, and other community spaces. Each bar in the chart represents a different setting, with the height of the bar corresponding to the number of outbreaks recorded in that particular environment.

The primary objective of this visualization is to identify which settings are most susceptible to outbreaks, providing crucial insights for public health monitoring and intervention strategies. For example, a higher bar for long-term care homes might indicate a greater vulnerability in these facilities, necessitating targeted preventive measures. Conversely, shorter bars might suggest settings that are relatively less affected or better managed in terms of outbreak control.

By clearly showing where outbreaks are most prevalent, health authorities and policymakers can better understand where to focus their efforts, whether in bolstering prevention strategies, enhancing response protocols, or directing educational resources.

3.2 Duration Of Outbreaks By Setting



In this plot, each box represents the spread of outbreak durations within a specific setting, such as a hospital, school, or community center. The key elements of the box plot – the median (central line within the box), the interquartile range (IQR, the box itself), and potential outliers (represented as points outside the box's whiskers) – collectively offer a concise yet comprehensive view of how long outbreaks tend to last in each setting.

This plot is useful for identifying variations in outbreak durations across different environments. For instance, a longer IQR in one setting might indicate more variability in how long outbreaks last there, suggesting a need for more flexible response strategies. The median provides a quick reference to the typical duration of an outbreak in each setting, while outliers can signal exceptionally long or short outbreaks that may warrant further investigation.

4 Discussion

4.1 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.

Weaknesses and next steps should also be included.

Appendix

A Additional data details

References

Gebru, Timnit, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé Iii, and Kate Crawford. 2021. “Datasheets for Datasets.” *Communications of the ACM* 64 (12): 86–92.

Gelfand, Sharla. 2022. *Opendatatoronto: Access the City of Toronto Open Data Portal*. https://CRAN.R-project.org/package=opendatatoronto.

R Core Team. 2022. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Spinu, Vitalie, Garrett Grolemund, and Hadley Wickham. 2023. *Lubridate: Make Dealing with Dates a Little Easier*. https://lubridate.tidyverse.org.

Wickham, Hadley. 2023. *Tidyverse: Easily Install and Load the Tidyverse*. https://tidyverse.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.