Unveiling the Patterns of Mortality*

A Comparative Analysis of Poisson and Negative Binomial Models in Predicting Causes of Death in Alberta

Yiyi Yao Zixi Song Pu Yuan

March 11, 2024

This study examined mortality data from Alberta, between 2001 and 2022, using general linear models to find the pattern of top causes of deaths over years. The study indicated that, compared to the Poisson model, the Negative Binomial model better represents the variability in death rates, particularly in the face of overdispersion. The application of more relevant models helps to simulate mortality data in the present as well as predict future data. These findings are useful for improving public health and interventions, as well as addressing healthcare resource allocation problems and developing dynamic policies in response to Alberta's changing mortality environment.

1 Introduction

In the fields of public health and epidemiology, understanding mortality patterns is essential for effective healthcare planning, policy-making, and resource allocation. Alberta, Canada, with its diverse population and healthcare needs, provides a unique opportunity to study these patterns. Comprehensive analyses of mortality data can provide insight into the major causes of death, trends over time, and the impact of public health interventions. However, traditional modeling often fails to account for the complexity and variability inherent in mortality data, such as the overdispersion of deaths among different causes. This shortcoming highlights the need for a more accurate approach to modeling mortality data to inform healthcare strategies and interventions.

To address this need, this study utilizes mortality data for Alberta for the years 2001-2022 and performs a comparative analysis of Poisson and Negative Binomial models. The goals of the study were to identify and visualize the leading causes of death in the region, determine trends

*Code and data are available at: https://github.com/Yaoee111/mortality_alberta.git. SSRP replication available at: https://doi.org/10.17605/OSF.IO/QDWBR.

over time, and assess the impact of various diseases and conditions on mortality. Using both statistical models, the study seeks to determine which model better describes the variability and distribution of mortality data, particularly in the context of overdispersion.

Our estimand is the annual change in mortality rates for the 5 most common causes of death in Alberta from 2001 to 2022, with a special focus on observing the effect of the COVID-19 pandemic on these rates after 2019.

The results of the study suggest that the Negative Binomial model is a better fit for Alberta mortality data and is effective in reflecting observed changes in the number of deaths. In addition, it highlights the most common causes of death in the province, providing a comprehensive understanding of public health issues in Alberta. These findings will be critical in guiding future public health initiatives, interventions, and decision-making aimed at reducing mortality and meeting the health needs of Alberta residents.

The rest of the paper is organized as follows: The Section 2 describes the data, variables, and methods used in the study and the rationale for the selection of this dataset. The data is presented through graphs. The Section 3 describes how Poisson and Negative Binomial models were constructed and predictions were made. The Section 4 presents the results of the analysis in a table. The Section 5 provides an in-depth discussion of our findings and reflections on the research process. Finally, the Section A adds details of the models.

We use the statistical programming language `R` (R Core Team 2023). In the data analysis and visualization process, we also made use of the following `R` packages: `readr` (Wickham, Hester, and Bryan 2024), `knitr` (Xie 2018), `kableExtra` (Zhu 2021), `ggplot2` (Wickham 2021a), `stringr` (Wickham 2021b), `janitor` (Firke 2021), `rstanarm` (Team 2021), `modelsummary` (Arel-Bundock 2022), `broom` (Robinson 2014), `broom.mixed` (Bolker and Robinson 2022), `magrittr` (Bache and Wickham 2014), `parameters` (Lüdecke et al. 2020), `tidyverse` (Wickham, Vaughan, and Girlich 2024), `dplyr` (Wickham et al. 2022), `patchwork` (Pedersen 2019), `bayesplot` (Gabry and Mahr 2024), and `loo` (Vehtari, Gelman, and Gabry 2016).

2 Data

2.1 Overview of the dataset

This dataset is from Open Data Alberta and documents the leading causes of death over several years. The broader context of this dataset is public health and epidemiology. This dataset contains several key variables, including 'Calendar Year', 'Cause', and 'Total Deaths'. And this dataset is recorded from 2001 to 2022.

2.2 Variable Examination

Calendar Year: This numeric variable records the year in which the death occurred (from 2001 to 2022). It is essential for trend analysis over time. Cause: This is a categorical variable that records the medical cause of death. It provides insight into the general health issues that lead to death. Total Deaths: This is a numeric variable indicating the number of deaths attributed to each cause in a given year. This indicator is essential for assessing the impact of each health problem.

2.3 Rationale for Selection

While other datasets on mortality exist, the Open Data Alberta was chosen because of its comprehensive coverage of the province. It provides a detailed breakdown of causes of death and the dataset is official and reliable. Other datasets were considered, but they lacked a specific regional focus or did not provide the same level of detail over an equivalent period.

2.4 Data cleaning and preparation

Cleaning focused on simplifying the dataset to allow for a more streamlined examination. The main cleaning step consisted of removing rows that were not relevant to our analysis. This process did not construct new variables but rather refined the dataset to the most relevant parts.

2.5 Analysis and Insights

Graphs, including time-series plot of Total Deaths (Figure 1) and bar charts of Deaths by Causes (Figure 2), illustrate mortality trends and leading causes of death in Alberta. Through the time series plot, we can conclude that Total Deaths shows a steady upward trend with the increase in years.Noteworthy, after 2019, a certain condition caused a spike in mortality, which we will know later that it is because of Covid-19. The bar chart shows the total number of deaths attribute to each cause. For the sake of presentation, we have listed only the top 20 here. Summary statistics tables provide a overview of the data, highlighting key figures such as leading causes of death and number of deaths.

3 Model

The goal of our modeling work is twofold: Firstly, find out which model, Poisson or Negative Binomial, better explains the number of deaths from different causes in Alberta. Our first aim is to pick a model that fits our data well, showing us how deaths change with different causes.

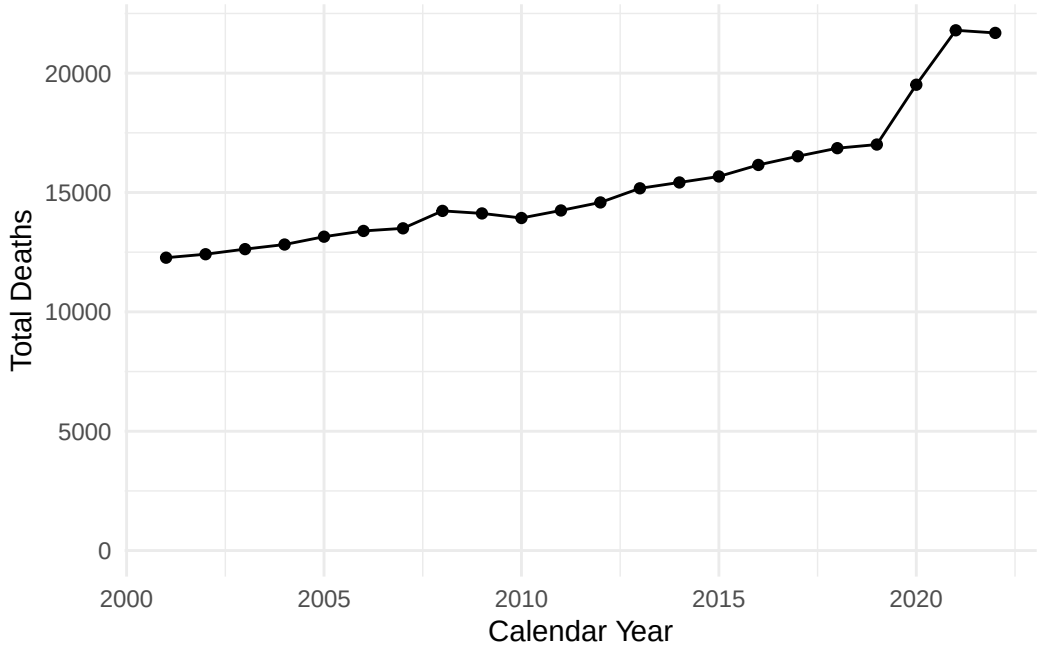

Figure 1: Time Series Plot of Total Deaths Over Time

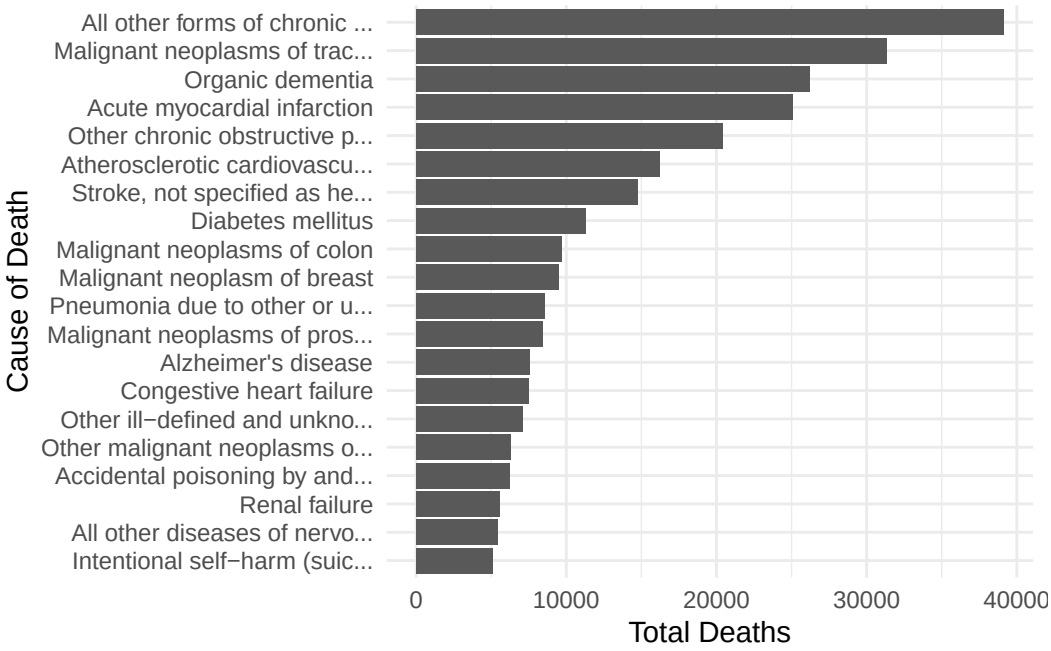

Figure 2: Bar Plot of Deaths by Top 20 Causes

The checks we did show that the Negative Binomial model works better because it can handle more variation in the data.

Our second aim is to figure out what influences how many people die from these causes. By using the better model, we can see which factors are most important in affecting mortality rates. This helps public health officials know where to focus their efforts to prevent deaths, making our communities healthier.

Background details are included in Appendix A.1.

3.1 Model set-up

To make it simpler to create a linear model to analyze the death data, we selected the top 10 causes of death for the year 2022. To present this data more directly, we created tables and figures. Table 1 shows the Top 10 causes of death in Alberta in 2022.

Table 1: Top-ten causes of death in Alberta in 2022

Year	Cause	Ranking	Deaths	Years
2022	Organic dementia	1	2,377	22
2022	All other forms of chronic …	2	2,098	22
2022	Other ill-defined and unkno…	3	1,714	4
2022	COVID-19, virus identified	4	1,547	3
2022	Malignant neoplasms of trac…	5	1,523	22
2022	Acute myocardial infarction	6	1,240	22
2022	Accidental poisoning by and…	7	1,200	10
2022	Other chronic obstructive p…	8	1,183	22
2022	Diabetes mellitus	9	730	22
2022	Stroke, not specified as he…	10	650	22

We then created line graphs of the top five causes of death in 2022, showing how they performed from 2001 to 2022 (Figure 3).

For creating the Poisson and Negative Binomial models, We can introduce some math functions. The Poisson regression model is expressed mathematically as

$$\log(\lambda_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$$

where λ_i is the expected number of deaths (the mean of the Poisson distribution) for the i-th observation, X_{1i}, … , X_{pi} are the explanatory variables (e.g., cause of death), and β_0, β_1, … , β_p are the coefficients to be estimated.

The Negative Binomial regression model is expressed mathematically as:

$$\log(\mu_i) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$$

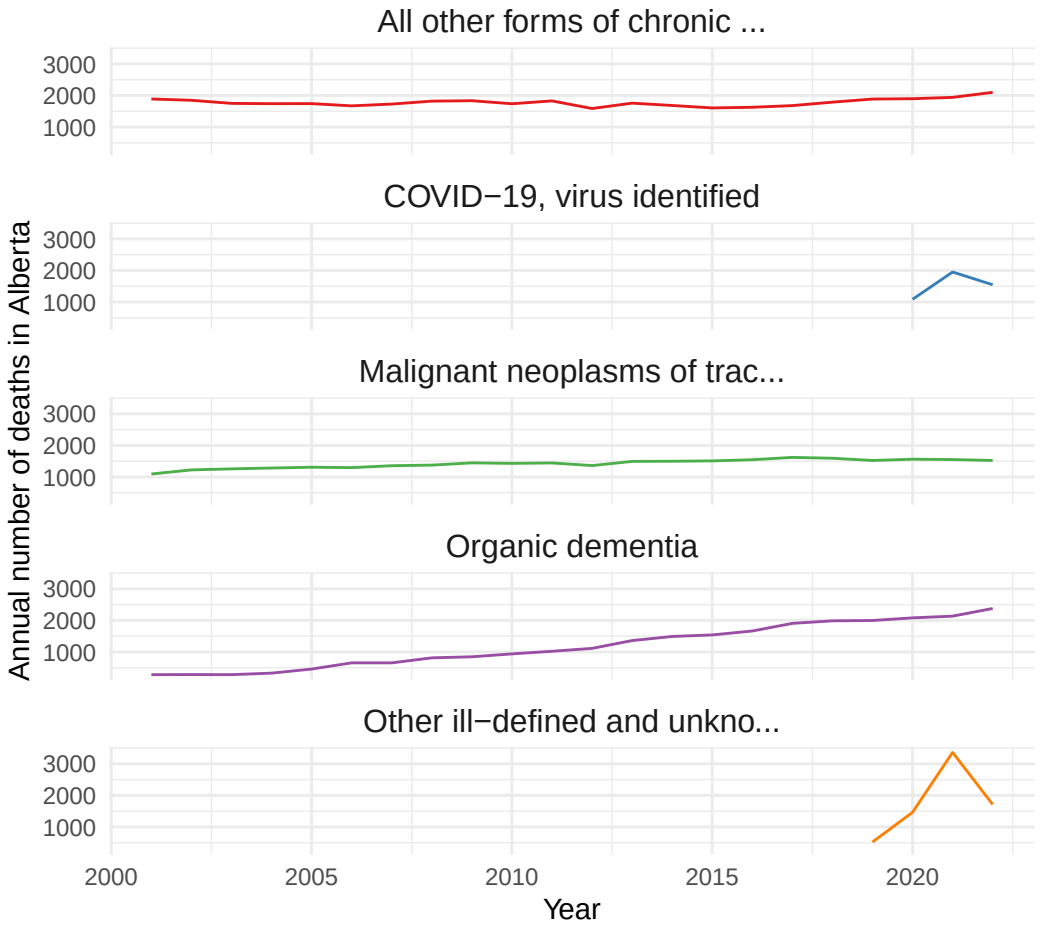

Figure 3: Annual number of deaths for the top-five causes in 2022, since 2001, for Alberta, Canada

And the variance is given by:

$$Var(Y_i) = \mu_i + \alpha\mu_i^2$$

where μ_i is the mean count for the i-th observation, and α is the overdispersion parameter. The variance is no longer assumed to be equal to the mean but grows with the mean, adjusted by α.

3.2 Model justification

In choosing between the Poisson and Negative Binomial models for analyzing mortality data, we consider the nature of our data. The Poisson model is straightforward and assumes that the average number of deaths is the same across all groups we study. It's a good basic model if our data are simple and don't vary too much. However, real-world data often don't follow these simple patterns. For example, the number of deaths from one cause might be more unpredictable than from another, leading to "overdispersion".

This is where the Negative Binomial model works better. It adds an extra feature to handle overdispersion, which makes it more flexible and usually a better fit for real-life data like ours. While it's a bit more complex to use, this model gives us more reliable results when our data show a lot of variability.

In short, if our mortality data are pretty consistent and don't vary much, the Poisson model could work well. But if we find that the number of deaths fluctuates a lot or is unpredictable, the Negative Binomial model is likely the better choice for accurate analysis.

4 Results

Our results are summarized in Table 2

To analyze the Poisson and Negative Binomial models, we imply the function $\log(E(y)) = \beta_0 + \beta_1 X_{1i} + ... + \beta_p X_{pi}$. The coefficients represent the β_p values in the equation, and the X_{pi} variables are the different causes of death you've included in the model.

4.1 Poisson Model

In the Poisson regression, the coefficients can be interpreted as the log change in the expected count of deaths for a one-unit increase in the predictor variable, holding other variables constant. For instance:

The intercept (7.484) indicates the log of expected deaths when all predictor variables are zero. Since this isn't meaningful for categorical variables like cause of death, it's better understood as the log of expected deaths for the reference category.

Table 2: Modeling the most prevalent cause of deaths in Alberta, 2001-2022

	Poisson	Negative binomial
(Intercept)	7.484	7.487
		(0.094)
causeCOVID-19, virus identified	−0.153	−0.147
		(0.256)
causeMalignant neoplasms of trac...	−0.223	−0.224
		(0.128)
causeOrganic dementia	−0.401	−0.403
		(0.129)
causeOther ill-defined and unkno...	−0.008	0.004
		(0.240)
Num.Obs.	73	73
Log.Lik.	−6421.556	−565.324
ELPD	−6751.2	−570.5
ELPD s.e.	1420.4	6.4
LOOIC	13 502.4	1141.0
LOOIC s.e.	2840.8	12.7
WAIC	14 398.1	1140.4
RMSE	457.92	457.98

Table 3: Leave-One-Out Cross-Validation

	elpd_diff	se_diff
cause_of_death_alberta_neg_binomial	0.0	0.0
cause_of_death_alberta_poisson	-6180.7	1414.4

The coefficient for "COVID-19, virus identified" (-0.153) suggests that, when comparing to the reference category (likely "no cause" or an omitted baseline cause), the expected log count of deaths decreases by 0.153, implying fewer deaths attributed to COVID-19 relative to the baseline.

To interpret these effects on the original count scale, we exponentiate the coefficients: For COVID-19, $e^(- 0.153) \approx 0.86$, meaning the expected count of deaths is multiplied by about 0.86 for COVID-19 relative to the reference category, indicating a decrease.

4.2 Negative Binomial Model

The Negative Binomial model accounts for overdispersion and has similar interpretation for the coefficients. However, it provides a potentially better fit if the variance of the death counts is significantly greater than the mean.

The coefficients in the Negative Binomial model are nearly identical to those in the Poisson model, which indicates that both models provide a consistent interpretation of how the causes of death relate to mortality rates.

4.3 LOO

The Negative Binomial model demonstrated a better fit over the Poisson model as indicated by the Leave-One-Out (LOO) Cross-Validation results (Table 3). In this case, we find that the Negative Binomial model is a better fit than the Poisson because ELPD is larger.

5 Discussion

5.1 Things we have done in this paper

In this paper, we analyzed Alberta's mortality data to see which statistical model, Poisson or Negative Binomial (NB), fits best. Our analysis started with the Poisson model, which is straightforward but assumes that the average number of deaths is the same as the variability

in those numbers. However, our data showed more variability than the Poisson model could handle, a condition known as overdispersion.

This led us to the NB model, which is designed to deal with overdispersion by adding an extra parameter. Through visual checks and Leave-One-Out Cross-Validation (LOO), we found that the NB model provided a better fit for our data, capturing the true variability in mortality rates more accurately than the Poisson model.

Ultimately, our work demonstrates that the NB model is more suitable for analyzing mortality data, offering clearer insights into the causes of death in Alberta. This finding is crucial for public health planning and interventions.

5.2 Something that we learn about the world

One of the findings of our study is that mortality rates increased substantially after 2019, largely due to the Covid-19 epidemic. Based on our analysis of the post-2019 data, the emergence of COVID-19 had a clear impact on mortality. Worldwide epidemics have a corresponding impact on local mortality rates. The dramatic increase in mortality following a pandemic highlights the profound impact of infectious diseases on global health systems and society as a whole.

Because pandemics like Covid-19 typically emerge suddenly, general models have difficulty predicting such diseases. Therefore, we need more accurate models to simulate and predict the data.

These data reflect the importance of a rapid public health response to emerging infectious disease threats. The ability to rapidly identify, contain and slow the spread of such diseases can significantly alter their impact on mortality. This highlights the need for establishing adequate monitoring systems and global information sharing, as well as early detection of disease outbreaks.

5.3 Another thing that we learn about the world

Using Leave-One-Out Cross-Validation, we found that the Negative Binomial model is a better fit for our data on causes of death in Alberta than the Poisson model. This suggests that the NB model, which accounts for overdispersion is more aligned with the real-world complexity and variability observed in mortality data.

Through this study, we learn that statistical modeling, when carefully applied, can offer profound insights into real-world phenomena, such as mortality patterns. The ability of the NB model to better capture the variability in the data underscores the importance of choosing appropriate models. This not only enhances our understanding of mortality but also points to the broader lesson that in the face of complex data, models that offer flexibility and can accommodate extra variability are crucial for deriving accurate and meaningful insights. This

approach can be extended beyond mortality data to various fields where understanding patterns and making predictions are vital, illustrating the universal value of statistical modeling in deciphering the complexities of the world around us.

5.4 Weaknesses

Our study, looking at death trends in Alberta and how COVID-19 affected these, has been informative. But, like all research, it has its limits.

First off, we're working with past data that others collected. This means we might not have all the details we wish we had, like the exact reasons behind each death. Sometimes, how deaths are recorded can mix up different health issues, which might confuse our results.

We mainly used two types of math models to understand the data. These models are great tools, but they're not perfect. They work under certain rules that might not fit every situation perfectly. For instance, we chose one model over another because it handled the data's ups and downs better. Yet, there could be other reasons for these ups and downs that we didn't get into.

Also, our study didn't dive deep into how people's living situations, access to healthcare, or personal habits might influence these death trends. These are big factors, but they need more complex studies to really understand.

Lastly, what we found out about Alberta might not be true for other places. Every area has its own health challenges and ways of handling them. Plus, the big impact we saw from COVID-19 is just one part of the bigger picture of health over the years.

So, while we learned a lot, there's still much more to explore. Future studies could look into these other factors and use different methods to get a fuller picture of what's going on.

5.5 Next steps

After looking at death trends in Alberta, here's how we can learn more and better tackle health issues:

1. Get More Information: We need to look at more things like people's living situations, if they can get to a doctor easily, and their lifestyle choices. This will help us understand why some health problems happen more often.
2. Try New Ways of Studying Data: We could use newer, smarter methods to study the information. This might help us see patterns and connections we missed before.
3. Look Closer at Certain Groups: It's important to study different groups of people or areas in more detail. This can show us who might need more help with their health.

4. Study Over Time: Following the same people or groups over years can tell us how health problems change and why.
5. Check If Health Actions Work: We should see if the steps we take to stop diseases, like COVID-19, really work. This means looking at how well vaccines or health advice work.
6. Compare Different Places: Seeing how different areas deal with health problems can teach us what works best.

A Appendix

A.1 Model details

A.1.1 Posterior predictive check

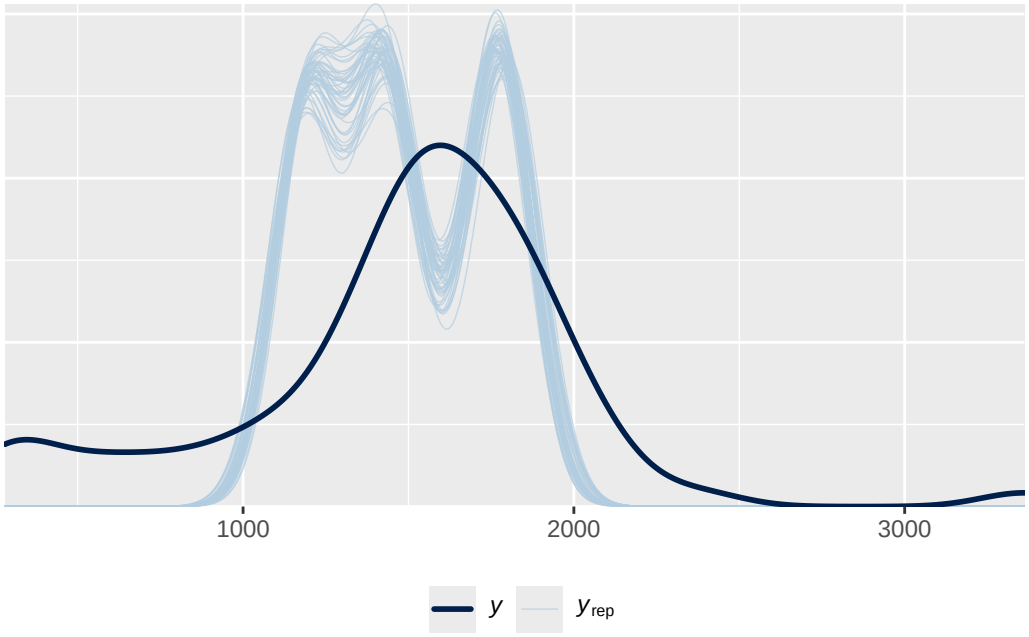

Figure 4: Posterior prediction checks for poisson model

In Figure 4, we implement a posterior predictive check for the Poisson model, which reveals noticeable discrepancies between the observed mortality counts and those predicted by the model. This suggests the presence of overdispersion in the data, indicating that the Poisson model may not adequately capture the variability in mortality rates across different causes of death in Alberta.

In Figure 5, we conduct a posterior predictive check for the Negative Binomial model, displaying a significantly better alignment between the observed data and the model's predictions. This improved fit highlights the Negative Binomial model's ability to account for overdispersion, making it a more accurate and reliable choice for analyzing Alberta's mortality data.

References

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.

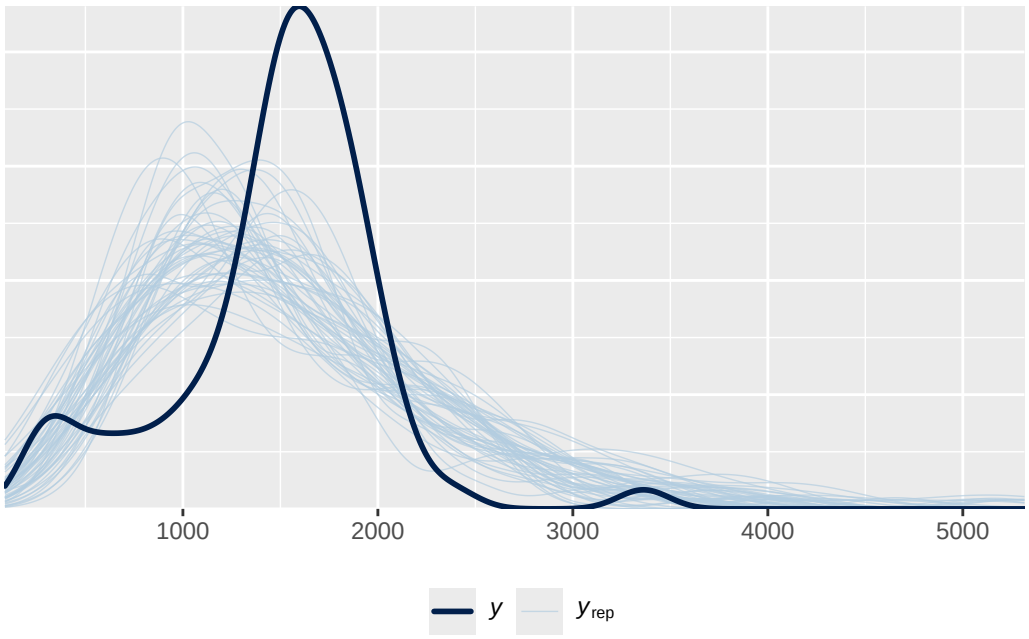

Figure 5: Posterior prediction checks for negative binomial model

Bache, Stefan Milton, and Hadley Wickham. 2014. "Magrittr: A Forward-Pipe Operator for r." *R Package Version* 1 (1).

Bolker, Ben, and David Robinson. 2022. *Broom.mixed: Tidying Methods for Mixed Models.*

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data.* https://CRAN.R-project.org/package=janitor.

Gabry, Jonah, and Tristan Mahr. 2024. "Bayesplot: Plotting for Bayesian Models." https://mc-stan.org/bayesplot/.

Lüdecke, Daniel, Mattan Ben-Shachar, Indrajeet Patil, and Dominique Makowski. 2020. *Extracting, Computing and Exploring the Parameters of Statistical Models Using R.* http://dx.doi.org/10.21105/joss.02445.

Pedersen, Thomas Lin. 2019. "Package 'Patchwork'." *R Package Http://CRAN. R-Project. Org/Package= Patchwork. Cran.*

R Core Team. 2023. *R: A Language and Environment for Statistical Computing.* Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Robinson, David. 2014. "Broom: An r Package for Converting Statistical Analysis Objects into Tidy Data Frames." *arXiv Preprint arXiv:1412.3565.*

Team, Stan Development. 2021. *Rstanarm: Bayesian Applied Regression Modeling via Stan.* https://mc-stan.org/rstanarm.

Vehtari, Aki, Andrew Gelman, and Jonah Gabry. 2016. "Loo: Efficient Leave-One-Out Cross-Validation and WAIC for Bayesian Models." https://CRAN.R-project.org/package=loo.

Wickham, Hadley. 2021a. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics.* https://CRAN.R-project.org/package=ggplot2.

———. 2021b. *Stringr: Simple, Consistent Wrappers for Common String Operations*. https://CRAN.R-project.org/package=stringr.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2022. *Dplyr: A Grammar of Data Manipulation*. https://github.com/tidyverse/dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2024. *Readr: Read Rectangular Text Data*. https://readr.tidyverse.org.

Wickham, Hadley, Davis Vaughan, and Maximilian Girlich. 2024. *Tidyr: Tidy Messy Data*. https://tidyr.tidyverse.org.

Xie, Yihui. 2018. "Knitr: A Comprehensive Tool for Reproducible Research in r." In *Implementing Reproducible Research*, 3–31. Chapman; Hall/CRC.

Zhu, Hao. 2021. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.