Alberta Mortality 2000 to 2022*

My subtitle if needed

Yihang Cai

March 15, 2024

First sentence. Second sentence. Third sentence. Fourth sentence.

1 Introduction

The COVID-19 pandemic has profoundly impacted global health, becoming a prominent cause of mortality with over 19,700 Canadians dying to the virus in 2022, marking the highest annual death toll since the pandemic's onset (ICI.Radio-Canada.ca, n.d.). This stark statistic places COVID-19 among the leading causes of death in Canada since 2020. Beyond the pandemic, however, numerous other factors contribute to mortality rates, prompting an investigation into the primary causes of death preceding the emergence of COVID-19. This paper focuses on Alberta, a Canadian province that has maintained detailed records of mortality causes since 2001, offering a comprehensive dataset of the top 30 causes of death annually from 2001 to 2022.

This research endeavors to analyze the top five causes of death over the past 22 years in Alberta, quantifying their relative contributions to the overall mortality rate. To this end, we employ two statistical models: the Poisson distribution model and the Negative binomial distribution model. The choice of dual models allows for a nuanced analysis, accommodating the data's variability and overdispersion, with the ultimate goal of identifying the model that best combines accuracy and precision in reflecting the underlying trends.

Data for this study were sourced from the Alberta Government's open data portal, focusing on explaining the relationship between individual causes of death and the total mortality rate. This analysis seeks to determine the estimand, a statistical term denoting the quantity of interest that our models aim to estimate.

The structure of this paper is designed to guide the reader through our research journey, starting with Section Section 1, which introduces the study's context, objectives, and preliminary

*Code and data are available at: https://github.com/peachvegetable/Alberta-mortality

observations. Section Section 2 presents visualizations of the changing landscape of mortality causes from 2001 to 2022, accompanied by a detailed description of the dataset. Section Section 3 delves into the methodology, clarifying the statistical models employed in our analysis. The findings and their statistical examination are detailed in Section 4, while Section 5 discusses these results, reflecting on the study's limitations and proposing avenues for future research.

2 Data

Data for this study were prepared and analyzed using R (R Core Team 2023), grasping several packages including Tidyverse (Wickham et al. 2019) for data manipulation, ggplot2 (Wickham 2016) for visualization, Janitor (Firke 2023) for data cleaning, Readr (Wickham, Hester, and Bryan 2023) for data import, Dplyr (Wickham et al. 2023) for data manipulation, Knitr (Xie 2014) for dynamic reporting, Modelsummary (Arel-Bundock 2022) for summarizing model outputs, and Rstanarm (Goodrich et al. 2022) for Bayesian modeling.

Our analysis focuses on the Alberta Government's open data, encompassing annual records of the top 30 causes of death from 2001 to 2022 (Government 2015). This comprehensive dataset sheds light on mortality trends in Alberta, offering insights into the dominant health challenges before the onset of the COVID-19 pandemic. Such an extensive temporal range is crucial for identifying long-term trends and shifts in public health priorities.

The dataset's primary variables include 'calendar_year' (renamed to 'Yaer'), capturing the range from 2001 to 2022; 'cause' (renamed to 'Cause'), listing the top 30 distinct causes of death annually; and 'total_deaths' (renamed to 'Deaths'), representing the yearly mortality count per cause. Initial data cleaning involved condensing the lengthy cause-of-death labels to ensure clarity in visualization. Furthermore, our analysis strategically omits causes not consistently present over the 22-year span, such as COVID-19, to maintain a focus on long-term trends and ensure analytical consistency.

Table 1: Top 10 causes of deaths in 2022, Alberta

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

Table 1: Top 10 causes of deaths in 2022, Alberta

Year	Cause	Ranking	Deaths	Years
2022	Stroke, not specified as he...	10	650	22

Table 1 presents data from Alberta for the year 2022, focusing on the top 10 causes of death. Each row in the table lists a specific cause of death, its rank in terms of mortality for that year, the number of deaths attributed to that cause, and the number of years that particular cause has appeared in the top 10 list out of the 22-year period studied.

For instance, the top cause of death for 2022 is listed as "Organic dementia," which caused 2,377 deaths and has been among the top 10 causes for all 22 years analyzed. The fifth cause is "COVID-19, virus identified" with 1,547 deaths, but it has only been in the top 10 causes for 3 years, corresponding to the recent years of the pandemic.

We can use this table for quickly identifying the most significant health concerns in Alberta in 2022 and assessing their persistence over time.

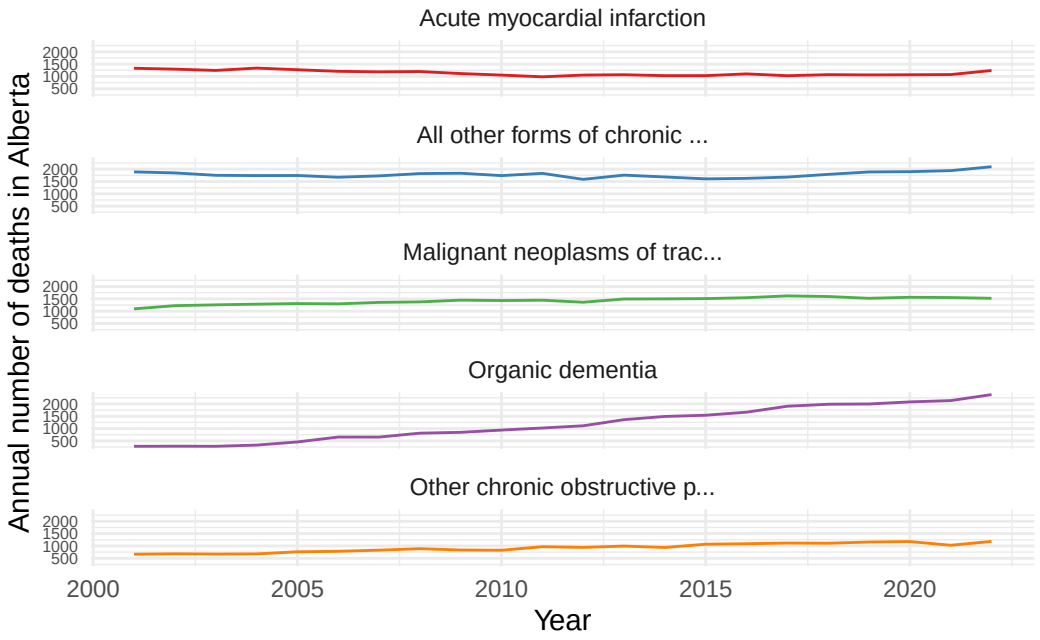


Figure 1: Top 5 causes of deaths from 2000 to 2022, for Alberta, Canada

Figure 1 shows a multi-line graph depicting the annual number of deaths in Alberta for various causes from 2001 to 2022. Acute Myocardial Infarction (Heart Attack): The line appears relatively flat, which indicates that the number of deaths from heart attacks has remained consistent over the 22-year period. There is no significant upward or downward trend, suggesting

stable incidence or improvements in treatment may be offsetting the risk factors in the population. All Other Forms of Chronic Diseases: This category shows a slight downward trend, indicating a small decrease in the number of deaths over time. This could be due to better management of chronic diseases or changes in diagnostic criteria and classification. Malignant Neoplasms of Trachea, Bronchus, and Lung (Lung Cancer): The line is fairly stable with a slight increase, suggesting a small rise in mortality from lung cancer over the years. This could reflect population growth, aging, or lifestyle factors influencing disease rates. Organic Dementia: There is a noticeable upward trend in deaths due to organic dementia. This is likely reflective of an aging population and increased diagnosis rates, as well as potentially increased prevalence of the disease. Other Chronic Obstructive Pulmonary Diseases (COPD): The trend for COPD is relatively stable with a very slight increase, indicating mortality rates from COPD have remained somewhat constant, potentially reflecting better treatments balancing out any increases in prevalence. When comparing the slopes of the lines, organic dementia shows the most significant increase, potentially indicating it is becoming a more prevalent cause of death. Acute myocardial infarction and COPD display relative stability compared to other causes, suggesting that the impact of these conditions on mortality rates has not changed markedly. The slight decline in 'All Other Forms of Chronic Diseases' could reflect effective public health interventions or improvements in healthcare delivery.

3 Model

3.1 Model set-up

Two models generated, one follows poisson distribution and another follows negative binomial distribution. We are interested in how total number of deaths differs by different causes of deaths.

Define y_i as the total number of deaths for the i-th observation. Then β_0 is the expected log count of total deaths when none of the causes in the model are present. It's the starting point of the model's prediction.

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) \\
\log(\lambda_i) &= \beta_0 + \beta_1 \times \text{cause}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

Define y_i as the total deaths for the i-th observation. θ is the additional parameter to model overdispersion. μ_i is the mean of the Negative Binomial distribution for the i-th observation, β_0 is the intercept, and β_1 represents the effect of each cause of death.

Table 2: Showing the mean and variance of total deaths in each year from 2001 to 2022, in Alberta

Table 3: Mean and Variance of Total Number of Deaths Over 22 Years

Estimate	Variance
509.2	201414.1

$$
\begin{aligned}
y_i|\lambda_i, \theta &\sim \text{NegativeBinomial}(\mu_i, \theta) \\
\log(\mu_i) &= \beta_0 + \beta_1 \times \text{cause}_i \\
\beta_0 &\sim \text{Normal}(0, 2.5) \\
\beta_1 &\sim \text{Normal}(0, 2.5)
\end{aligned}
$$

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

We calculate the total number of deaths by the formula:

$$\text{total deaths} = e^{\beta_0 + \sum_{i=1}^{5} \beta_i X_i} \tag{1}$$

Where X_1, X_2, X_3, X_4, and X_5 represent $X_{\text{Acute myocardial infarction}}$, $X_{\text{All other forms of chronic}}$, $X_{\text{Malignant neoplasms}}$, $X_{\text{Organic dementia}}$, and $X_{\text{Other chronic}}$ respectively.

3.2 Model justification

The choice of the Poisson and Negative Binomial models stems from their suitability for count data. The Poisson model is a natural starting point for modeling count data, assuming that each event occurs independently and the mean rate of occurrence is constant.

However, the Poisson model's restrictive assumption that the mean equals the variance often does not hold in real-world data, prompting the use of the Negative Binomial model. The Negative Binomial model relaxes the equal mean-variance assumption by introducing a dispersion parameter, making it well-suited for modeling overdispersed count data that is common in mortality records.

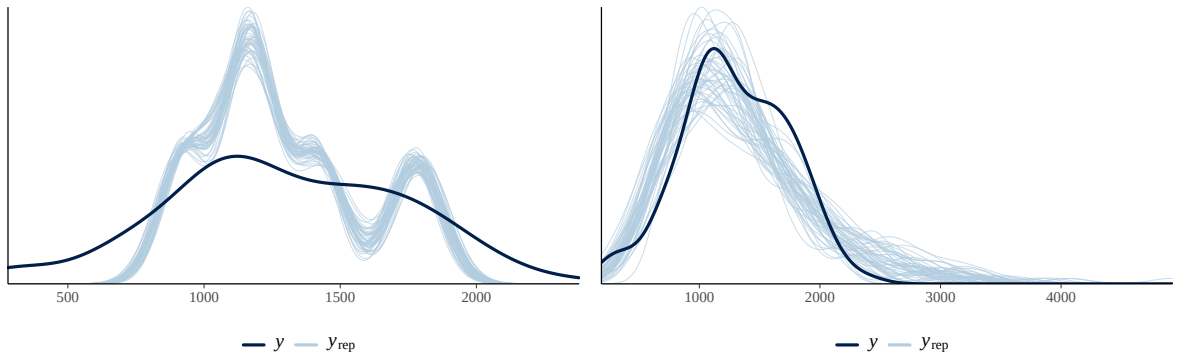
Table 2 illustrates a significant disparity between the mean(509.2) and the variance(201414.1) of total deaths over a span of 22 years, indicating the presence of overdispersion. Such a finding suggests that the Negative Binomial model may be more appropriate than the Poisson model for this dataset. The Poisson model operates under the assumption that the mean and variance are equal, an assumption not supported by our data. In contrast, the Negative

Table 4: LOO Comparison between Poisson and Negative Binomial Models

Table 5: LOO Comparison between Poisson and Negative Binomial Models

Model	ELPD diff	SE diff
Negative binomial model	0.000	0.000
Poisson model	-5091.179	1201.615

Binomial model can accommodate the observed overdispersion by allowing for a variance that is greater than the mean, thus providing a potentially better fit for the data. We also compare how the two models fit the real dataset using ppcheck in the bayesplot package (Gabry et al. 2019).



(a) Poisson model

(b) Negative binomial model

Figure 2: Comparing Poisson and negative binomial models

Figure 2 displayes a graphical comparison of posterior predictive checks for Poisson model and Negative binomial model. Figure 2a shows the fit for the Poisson model. The simulated "y_rep" lines show the spread and central tendency of the predictions that the Poisson model makes for the observed data. The replicated data lines seem to spread out and fluctuate more around the observed data, which suggests a lesser fit, particularly if the spread is wider and doesn't track the shape of the observed data closely. Figure 2b shows the fig for the Negative binomial model. The replicated data lines appear to follow the observed data line more closely, suggesting that the Negative binomial model captures the central tendency and variability of the observed data better than the Poisson model.

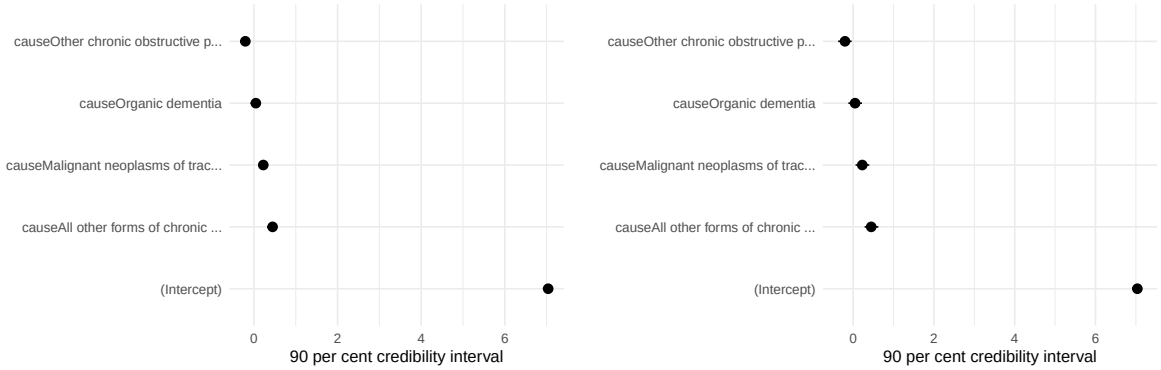
Table 4 shows a model comparison using the expected log pointwise predictive density (ELPD) for two models: a Poisson model and a Negative binomial model. Negative binomial model has an ELPD difference of 0.0, which serves as the baseline because it's the model with the higher (or equal) ELPD. Poisson_model shows an ELPD difference of -5091.2 compared to the Negative binomial model, with a standard error of the difference of 1201.6. The large negative

Table 6: Modeling the cause deaths in Alberta, 2000 - 2022

	Poisson	Negative binomial
(Intercept)	7.037	7.037
		(0.070)
causeAll other forms of chronic ...	0.446	0.448
		(0.101)
causeMalignant neoplasms of trac...	0.223	0.226
		(0.102)
causeOrganic dementia	0.046	0.048
		(0.101)
causeOther chronic obstructive p...	−0.206	−0.202
		(0.101)
Num.Obs.	110	110
Log.Lik.	−5718.182	−810.965
ELPD	−5906.6	−815.4
ELPD s.e.	1211.7	10.5
LOOIC	11 813.2	1630.9
LOOIC s.e.	2423.5	21.1
WAIC	11 965.6	1630.8
RMSE	325.38	325.38

ELPD difference suggests that the Negative binomial model has a much higher ELPD than the Poisson model, which means it is better at predicting new data according to this metric. The standard error of 1201.6 indicates the uncertainty around this ELPD difference estimate. A high standard error relative to the ELPD difference can indicate less confidence in the model comparison, but in this case, the magnitude of the difference (-5091.2) is much larger than the standard error, suggesting the result is quite robust. In summary, the table suggests that the Negative binomial model is significantly better at predicting the observed data than the Poisson model in this particular analysis.

Therefore, the visual evidence Figure 2 and the observed overdispersion beyond what the Poisson model can adequately handle Table 2, and the ELPD (**elpdtable?**) indicate that the Negative binomial model may be more appropriate for this dataset.



(a) Poisson model (b) Negative binomial model

Figure 3: 90% Confidence Interval

3.3 Model summary

4 Results

5 Discussion

5.1 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.

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

A Additional data details

B Model details

B.1 Posterior predictive check

B.2 Diagnostics

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.

Firke, Sam. 2023. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gabry, Jonah, Daniel Simpson, Aki Vehtari, Michael Betancourt, and Andrew Gelman. 2019. "Visualization in Bayesian Workflow." *J. R. Stat. Soc. A* 182: 389–402. https://doi.org/10.1111/rssa.12378.

Goodrich, Ben, Jonah Gabry, Imad Ali, and Sam Brilleman. 2022. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

Government, Alberta. 2015. *Leading Causes of Death*. https://open.alberta.ca/dataset/03339dc5-fb51-4552-97c7-853688fc428d/resource/1a10c821-7399-4d0f-95fb-f96728d01fae/download/deaths-leading-causes.xlsx.

ICI.Radio-Canada.ca, Zone Santé -. n.d. "Life Expectancy Fell in 2022 for 3rd Year in a Row: StatsCan: RCI." *Radio*. Radio-Canada.ca. https://ici.radio-canada.ca/rci/en/news/2030697/life-expectancy-fell-in-2022-for-3rd-year-in-a-row-statscan.

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

Wickham, Hadley. 2016. *Ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York. https://ggplot2.tidyverse.org.

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, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Jim Hester, and Jennifer Bryan. 2023. *Readr: Read Rectangular Text Data*. https://CRAN.R-project.org/package=readr.

Xie, Yihui. 2014. *Knitr: A Comprehensive Tool for Reproducible Research in R*. Edited by Victoria Stodden, Friedrich Leisch, and Roger D. Peng. Chapman; Hall/CRC. http://www.crcpress.com/product/isbn/9781466561595.