My title*

My subtitle if needed

Yanyu Wu Ziyi Liu Hechen Zhang

March 17, 2024

We examined the distribution of the top 30 causes of death for each year between 2001 and 2022 in Alberta, Canada. The conifer model and negative binomial regression were used to analyse the long-term leading causes of death and the sudden emergence of specific causes of death. According to our findings, negative binomial regression improves our ability to predict outcomes when the data is too spread out by fitting the data more accurately. The results could not only help policymakers design more effective preventive measures to reduce mortality from these conditions, but also help researchers and policymakers make more precise decisions.

Table of contents

*Code and data are available at: https://github.com/HechenZ123/Cause-of-Deaths-in-Alberta.git

1 Introduction

The mortality rate, often referred to as the death rate, represents an approximation of the fraction of a population that dies within a given time frame (Porta 2014). Mortality rates can serve as a crucial indicator of a population's health status, and it also reveals the impact of diseases and other health-related issues over a period of time. This paper explores the leading causes of death in Alberta for crafting effective public health strategies and policies and understanding the most significant health threats affecting a population for researchers(Alberta 2015). By identifying the main causes of mortality, health authorities can prioritise research funding towards diseases and conditions that have the highest impact on community health and lifespan (Vargas et al. 2019).

As discussed in the data section, we used data from Service Alberta (Alberta 2015) on the leading causes of deaths, in which the five most significant causes in 2022 were analysed. These five causes are Organic dementia, Other causes not clearly defined, COVID-19, and Cancers of the trachea, bronchus, and lungs. It was noted that, among the examples mentioned, the negative binomial regression is more accurate compared to the Poisson model, while Poisson regression is prone to errors.

1.1 Importing Important Packages.

In this analysis, we employ a range of R (R Core Team 2023) packages tailored for data cleaning, transformation, analysis, and reporting. `Tidyverse` by Wickham et al. (2019) is used for data wrangling, `janitor` package by Firke (2021) is used for data cleaning operations, and `knitr` by Xie (2021) for data presentation in data tables.The following code section aims at importing the important packages that are essential for examining the missing values in the data set.We run the model in R R Core Team (2023) using the `rstanarm` package of Goodrich et al. (2022). We use the default priors from `rstanarm`. For comprehensive mixed effects model analysis, we leverage the `broom.mixed` package (Bolker and Robinson 2022), which extends the `broom` package functionalities to mixed models, facilitating the extraction, tidying, and representation of model outputs. Furthermore, the `modelsummary` package (Arel-Bundock 2022) provides tools for creating customizable summary tables of model results, enhancing the interpretability and dissemination of statistical findings. By calculating the LOO-CV scores for different models with `loo` (Yao et al. 2017), we could compare them based on their out-of-sample predictive accuracy. Lower values of LOOIC indicate better model performance. The following code sections aim to import these crucial packages, essential for conducting a thorough analysis and addressing the research questions at hand, while ensuring data integrity and transparent reporting of results.

2 Data

Our data is of leading causes of death (Figure 1), from Alberta (2015).

2.1 Data Sources

Our research team conducted an in-depth analysis of mortality trends in Alberta, choosing a dataset published on the Alberta Open Government Portal as the basis for their study. This dataset was made available to the public beginning in 2015 and was last updated in 2023, and is designed to promote public engagement by providing transparent access to key health statistics.This dataset provides a comprehensive view of deaths across the province, covering the top 30 leading causes of death in each annual count. This dataset was chosen because it provides key statistics on the number of deaths and mortality rates. These statistics cover data from chronic diseases to unintended injuries and even public health emergencies, allowing us to accurately track trends in key causes of death. We chose to rely on this dataset primarily because it is sourced from the official government of Alberta, which provides an authoritative data base for our study. The endorsement of a government department ensures the accuracy of the data, which is a key prerequisite for conducting scientific research. In addition, the high degree of fit of the dataset with our research objectives was an important reason for its selection. Our study aimed to gain insight into the key factors that influence mortality in Alberta and to explore possible public health interventions. By analysing this data, we are able to identify priority areas for health interventions and provide a scientific basis for policy makers.

2.2 Features

Order by total number of deaths and a ranking of the top 30 causes of death in Alberta each year.Our table lists the top eight causes of death in Alberta in 2022. Examine each variable in detail: *Year*: This denotes the data gathering year, which for all entries is 2022. *Cause*: This represents the medical condition or event that led to death. The causes listed are shown in Figure 1: `Organic dementia` `All other forms of chronic...`, `Other ill-defined and unknown...`, `COVID-19, virus identified`, `Malignant neoplasms of the trachea, bronchus, and lung`, `Acute myocardial infarction`, `Accidental poisoning by and...`, `Other chronic obstructive pulmonary diseases`, *Ranking*: This is a ranking by the number of deaths caused by each disease, with 1 being the highest. *Deaths*: The number of deaths attributed to each cause. *Years*: Indicates the number of years in which data was collected for that reason.

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

Figure 1: Top-teight causes of death in Alberta in 2022

2.3 Data Method

In the preliminary phase, we worked on organizing and harmonizing the naming of variables in the dataset to enhance its readability and efficiency of analysis. In addition, we summarized the data in order to quantify the incidence of various causes of death, thus making the dataset more streamlined for more in-depth analysis. To avoid missing years in the dataset from affecting the results of the analysis, we decided to delete all records containing these missing values.We then prioritized the top 8 causes of death by 2022 based on their number of deaths. Particular attention was paid to the top 5 leading causes of death, which is at a critical time when the world is facing a pandemic virus, and the data allow us to understand the impact of the new crown on mortality.Our mortality data are consistent with the characteristic of occurring independently with equal means and variances. These mortality trends are fitted by Poisson and negative binomial statistical models because of the suitability of both models for working with count data. A method for assessing the effect of predictor variables on mortality changes is provided, with exponential coefficients that visually represent the change in mortality risk that may result from each unit change in a predictor variable.

Table 1: Top Five Causes of Death in Alberta for 2022

CauseOfDeath
Organic dementia
All other forms of chronic ischemic heart disease
Other ill-defined and unknown causes of mortality
COVID-19, virus identified
Malignant neoplasms of trachea, bronchus and lung

We can also see that in the data table for 2022, COVID-19 ranks fourth, with 1,547 deaths. Alberta's cause-of-death data proves that the coronavirus pandemic is one of the most serious public health crises of the early 21st century, with global implications. This figure not only speaks to the outbreak's death rate, but also suggests that the region's health care system may be under strain. The COVID-19 pandemic has disrupted the distribution of deaths from previously common causes, which may include some chronic diseases that have long held the top spot, such as cancer and heart disease, even though it was not the leading cause of death in the years shown in the table.

3 Model

3.1 Model setup:

In analyzing the association between the total number of deaths and significant causes of death, we used two different regression models to predict, namely the Poisson distribution and the binomial distribution. Both the Poisson and binomial distribution models are regression models used to analyze count data, primarily for finding the relationship between independent variables (predictors) and dependent variables (count outcomes). In regression models, the independent variables are independent, while the dependent variables depend on the independent variables. In our model, the independent variables are the different causes of death mentioned in the table, such as COVID-19, malignant neoplasms, organic dementia, etc. These variables are used to predict or explain the number of deaths. The dependent variable is the number of deaths in Alberta from 2001 to 2022. This variable is the response variable in the model, and its count is the result to be predicted or explained.

Using the Poisson distribution has different advantages and disadvantages in this study. The Poisson model is simple in form, with fewer parameters, making it easy to understand and explain, and because of its simplicity, this model has a faster computation speed and is easy to implement. The Poisson distribution has a solid theoretical foundation for modeling the number of times an event occurs. It assumes that the average frequency of events occurring over a certain time or space is constant, and the occurrence of individual events is independent. At the same time, the Poisson distribution is a regression model specifically used for analyzing count data (where the response variable is non-negative). In our study, our response variable is the number of deaths, which is definitely a positive number, making the Poisson distribution suitable for our research. However, it also has some shortcomings, such as the problem of overdispersion; if the actual data show that the variance is greater than the mean, the Poisson model may not be applicable, and the model's predictive effect will worsen.

Regarding the negative binomial distribution, its advantage is that it includes an additional parameter to model overdispersion, making it more suitable for data where the variance is greater than the mean. Compared to the Poisson model, the negative binomial model offers more flexibility to fit various data, especially those that exhibit significant overdispersion or clustering. The negative binomial model includes an extra parameter to account for overdispersion, allowing the variance to exceed the mean. In our study, this model also attempts to link death counts with various causes but is more flexible in handling data variability. However, compared to the Poisson distribution, the negative binomial model is more complex, involving more parameters, which may lead to difficulties in model explanation and communication. The negative binomial model typically requires more computational resources, so it may not be as efficient as the Poisson model on large datasets.

In our data situation, the negative binomial model seems more appropriate than the Poisson model, possibly due to overdispersion in the data. Therefore, in analyzing causes of death (such as COVID-19, malignant neoplasms, organic dementia, etc.), choosing the negative binomial model may provide more accurate estimates.

Poisson model:

$$
\begin{aligned}
y_i|\lambda_i &\sim \text{Poisson}(\lambda_i) && (1)\\
log(\lambda_i) &= \beta_0 + \beta_1 \cdot x_i && (2)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (3)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (4)\\
& && (5)
\end{aligned}
$$

Negative binomial model:

$$
\begin{aligned}
y_i|\lambda_i, \theta &\sim \text{NegativeBinomial}(\mu_i, \theta) && (6)\\
log(\mu_i) &= \beta_0 + \beta_1 \cdot x_i && (7)\\
\beta_0 &\sim \text{Normal}(0, 2.5) && (8)\\
\beta_1 &\sim \text{Normal}(0, 2.5) && (9)\\
& && (10)
\end{aligned}
$$

From this formula, we can consider y_i to be the total number of deaths, with the i-th observation representing the cause of death. β_0 represents the intercept, which is the expected log count of the response variable when all predictor variables are held at zero. The intercepts of the two models are similar. Each cause of death has a corresponding coefficient β_i, indicating the change in the expected log count of deaths for each unit increase in the predictor variable (i.e., the presence of the cause of death). A negative coefficient suggests that the presence of that cause is associated with a decrease in the death count relative to the baseline level. The intercept β_0 and the slope β_i both follow a prior distribution that is normally distributed with a mean of 0 and a standard deviation of 2.5.

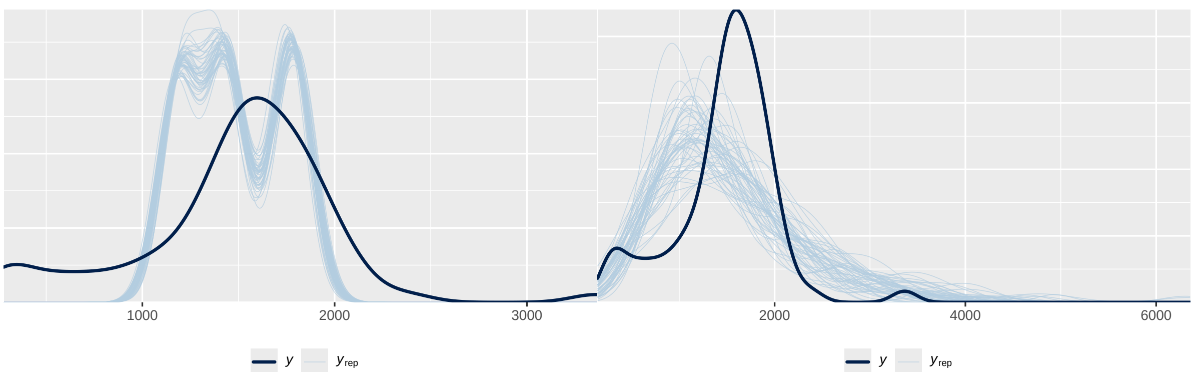
3.2 Model justification

Based on the models discussed above, it is challenging to determine whether each cause of death has a positive or negative correlation with the total number of deaths, as the leading cause of death changes each year. The specific impact of these changes must be determined by the coefficient of each leading cause of death to decide whether the correlation is positive or negative. According to Figure 3, from 2001 to 2022, organic dementia shows an upward trend, indicating that organic dementia is positively correlated with the total number of deaths. In the upcoming model results section, we will elaborate in detail on the association between each cause of death and the total number of deaths.

Table 2: Comparison of Mean and Variance of Total Deaths in Alberta from 2001 to 2022

Mean	Variance
1483.411	270830.3

Based on our discussion in the model set-up section, the negative binomial distribution seems more suitable for our research objectives than the Poisson distribution. The Poisson model assumes that the mean and variance are equal, but in the real world, they are rarely equal. As Table 2 shown below indicates, the mean and variance are not equal. The negative binomial distribution incorporates a dispersion parameter, making the predictions more accurate.



(a) Poisson model

(b) Negative binomial model

Figure 2: Comparing posterior prediction checks for Poisson and negative binomial models

We created Figure 2 that includes both the Poisson and negative binomial models. For the Poisson model, the actual observed values y are represented in black. The predicted values y_{Rep} are shown in light blue, which are replicated samples generated based on the posterior distribution of model parameters.

It can be seen that there is a significant overlap between many of the predictive distributions and the actual observed values, but at higher values, the range of the predictive distribution becomes very wide, indicating that the model may not fit well for high-count values.

For the negative binomial models, the actual observed values y are also represented in black, and the predicted values yrep are likewise shown in light blue, generated based on the posterior distribution of the model parameters. Compared to the Poisson model, the predictive distributions generated by the negative binomial model appear to be more concentrated, meaning they are more closely clustered around the actual observed values at higher counts, indicating that the negative binomial model may be more suitable for these data. From the figure, it can be observed that the predictive distributions of the negative binomial model match the actual observed values better than those of the Poisson model. In the Poisson model's figure, the predicted values align well with the actual values near the peak in the middle but are wider in the tail distribution, which may mean that the Poisson model struggles with extreme values. In contrast, in the figure for the negative binomial model, the predicted values seem to follow the actual observed values more closely across the entire range, typically indicating it can better handle the overdispersion of the data.

Table 3: LOO-CV Model Comparison: ELPD Difference and SE Difference

	Difference in Expected Log Predictive Density (ELPD Difference)	Standard Error of ELPD Difference (SE Difference)
Negative Binomial Model	0.00	0.000
Poisson Model	-6160.56	1412.084

To demonstrate that the negative binomial model is a better fit, we compared the Expected Log Predictive Density (ELPD) difference and the Standard Error (SE) difference between these two models in Table 3. The ELPD difference column compares the difference in expected log pointwise predictive density between the two models—the negative binomial model and the Poisson model. ELPD is an indicator of model predictive performance, with higher ELPD values typically indicating better model predictive performance. For the negative binomial model (listed as cause_of_death_alberta_neg_binomial), the ELPD difference is 0.0. For the Poisson model (listed as cause_of_death_alberta_poisson), the ELPD difference is -6160.6, meaning that, relative to the negative binomial model, the Poisson model's predictive performance is worse.

The SE difference represents the uncertainty in the estimation of the ELPD difference. A smaller SE difference indicates that the estimation of the ELPD difference is more precise; a larger SE difference indicates that the estimation of the ELPD difference is less stable and has more uncertainty. In other words, it tells us the reliability of the difference in predictive performance between the models. The SE difference is 1412.1, which is a relatively large value, indicating considerable uncertainty in the estimation of the predictive performance difference between the negative binomial model and the Poisson model. Nonetheless, the significant difference in ELPD (-6160.6) compared to its SE suggests that this difference is statistically significant. In other words, even considering the uncertainty in the ELPD difference, the negative binomial model's predictive performance is significantly better than that of the Poisson model, supporting our previous statements.

4 Results

4.1 data result

Figure 3 shows the annual number of deaths from the top five causes of death in Alberta in 2022 and traces the trend from 2000 to 2022. In this case, the number of deaths from all other forms of chronic ischemic heart disease shows a relatively stable trend line with a slight increase over time. This could mean that the number of deaths from chronic ischemic heart disease remains high, and this slight increasing trend could be related to population growth and aging, as ischemic heart disease is more commonly seen in older adults. Beginning in 2020, a trend line in COVID-19 deaths emerges abruptly, coinciding with a pandemic outbreak. This trend line, although short-lived, shows a sharp rise in deaths due to the virus, which is consistent with what happened during the pandemic. Factors such as vaccination and treatment measures have led to a significant reduction in this trend over time. For deaths from malignant tumors of the trachea, bronchus, and lungs, the trend line is relatively flat, suggesting that the number of deaths from these cancers has remained consistent over the years without significant increases or decreases. This stable trend may indicate that effective treatments have been able to counterbalance the effects of the increase in cases or that the promotion of regular medical check-ups has helped to reduce mortality from these diseases. The trend line for dementia deaths shows a gradual increase in deaths over time, which may be related to the growth of the older population. The increase in dementia-related deaths may be due to a number of factors, including increased life expectancy, which has led to a higher proportion of older people in the population, and the lack of effective treatments to prevent the progression of the disease. For the "other unspecified and unknown causes" category, there is a sharp increase from 2018 to 2019 and a rapid decline from 2021 to 2022. This increase coincided with the period of the COVID-19 pandemic, which we hypothesize may have been due to the fact that the new coronavirus and its variants were not immediately recognized or adequately documented at that time, resulting in many deaths that could not be accurately attributed to known causes. In addition, the healthcare system may have been under heavy pressure during the New Crown epidemic and lacked the resources to conduct thorough investigations, which in turn led to a large number of deaths being categorized as unknown causes.

4.2 model result

Table 4 shows the four most significant causes of death in Alberta from 2001-2022: organic dementia, COVID-19, malignant neoplasms of the trachea, bronchus, and lung, and other ill-defined and unknown causes of mortality. The coefficients indicate the relationship between the total number of deaths and the causes of death listed in the table.the model estimates the coefficients of each variable, quantifying the expected change of one unit change in the predictor variable (specific cause of death) against the response variable (in this case, the number of deaths), while holding all other variables constant. For predictive modeling, the fit of the model to historical data is crucial because it tells us how well the model explains

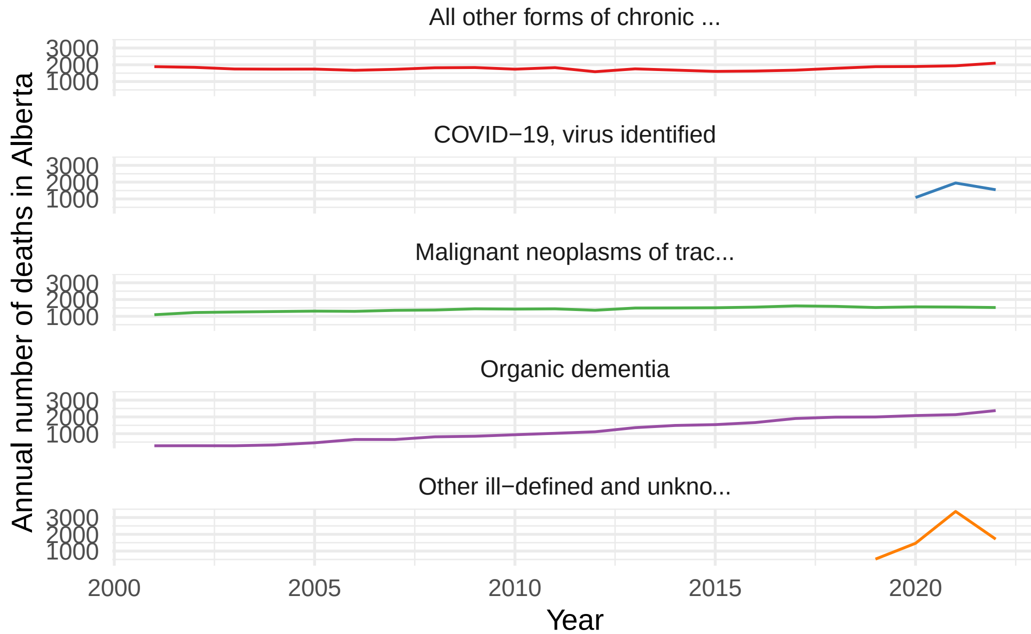


Figure 3: Annual number of deaths for the top-five causes in 2022, since 2001, for Alberta, Canada

changes in the data. The better the fit, the more confident we can use the model to predict the future.

4.2.1 Poisson distribution model

In the Poisson model, the expected logarithmic function $log(\mu_i)$ is defined as:

$$log(\lambda) = 7.484 - 0.152 \cdot COVID - 0.223 \cdot MaligantNeoplasms - 0.400 \cdot OrganicDementia - 0.007 \cdot UnknownCauses$$

The intercept represents the expected log count of deaths when all causes are absent. In the Poisson model, the baseline number of deaths equals 7.484. For COVID-19, the coefficient is -0.152, indicating that the presence of this cause is associated with a slight decrease in the total death count. Moreover, for malignant neoplasms with a coefficient of -0.223, suggesting a slightly larger decrease in total death counts when malignant neoplasms are a cause. In terms of organic dementia, a coefficient of -0.400 suggests a more substantial decrease in total death counts with organic dementia as a cause. For other ill-defined and unknown causes, the coefficient is -0.007, indicating almost no change in the total death count when these are listed as the cause.

Table 4: Modeling the most prevalent cause of deaths in Alberta, 2001-2022

	Poisson	Negative binomial
(Intercept)	7.484	7.482
		(0.093)
causeCOVID-19, virus identified	−0.152	−0.129
		(0.262)
causeMalignant neoplasms of trac...	−0.223	−0.220
		(0.131)
causeOrganic dementia	−0.400	−0.396
		(0.131)
causeOther ill-defined and unkno...	−0.007	0.017
		(0.241)
Num.Obs.	73	73
Log.Lik.	−6421.556	−565.317
ELPD	−6731.0	−570.5
ELPD s.e.	1418.0	6.3
LOOIC	13 462.1	1140.9
LOOIC s.e.	2836.0	12.6
WAIC	14 288.6	1140.4
RMSE	457.92	458.07

4.2.2 Negative binomial distribution model

In the negative binomial model, the expected logarithmic count is defined as:

$$log(\mu_i) = 7.482 - 0.129 \cdot COVID - 0.220 \cdot MaligantNeoplasms - 0.396 \cdot OrganicDementia + 0.017 \cdot UnknownCauses$$

For a negative binomial model, the baseline number of deaths when all causes are absent equals 7.482. For the cause of Covid, each additional count is correlated with a 0.129 decrease in the log count of deaths. This also suggests that areas with one more unit of covid-19 reported deaths are expected to see a decrease in the total number of deaths (logarithmic scale). In addition, for malignant neoplasms, the coefficient indicated each additional count of deaths due to this cause is associated with a 0.22 decrease in the log count of total deaths. Similarly, the coefficient for cause of organic dementia is equal to -0.396, this indicates that for each additional count of deaths due to this cause is associated with a 0.396 decrease in the log count of total deaths. Lastly, the coefficient for other ill-defined and unknown causes is small and positive which suggest that an increase in the number of deaths due to these causes are slightly associated with an increase in the log count of total deaths.

Both models' coefficients represent that as each of these causes of death become a more prevalent reason for death, the total number of deaths decreases in relation to the baseline level set by the intercept, except for "Other ill-defined and unknown causes," which shows no significant effect.

5 Discussion

5.1 Addressing Public Health Challenges in Alberta: Strategies for Health Policy and Social Regulation

Based on the top five causes of death in Alberta, it is imperative that we consider initiatives to improve health policy and social regulation to address health challenges, prioritizing public health issues. Looking at the data, given the significant impact of Covid-19 on mortality rates, it is crucial to continue efforts to implement public health interventions. This includes increasing mass vaccination activities, continuing to promote mask-wearing and social distancing measures, and enhancing testing and contact tracing capabilities. Furthermore, it is essential to ensure that the healthcare system has sufficient capacity to handle an increase in cases. It is worth noting that in the coming years, due to the passage of several years since the virus initially emerged, Covid-19 may not continue to be such a significant cause of death, as the virus gradually becomes less virulent or severe (Talic et al. 2021). Implementing policies for cancer prevention and control can help reduce mortality from malignant neoplasms of the trachea, bronchus, and lung. This may include implementing tobacco control measures, such as increasing tobacco product taxes, comprehensive smoking cessation programs, and restricting tobacco advertising and promotion. Additionally, promoting healthy lifestyles, early cancer screening programs, and providing high-quality cancer treatment services are also crucial (Eastman 2023).

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

References

Alberta, Service. 2015. "Leading Causes of Death." *Leading Causes of Death - Open Government*. https://open.alberta.ca/dataset/leading-causes-of-death/resource/3e241965-fee3-400e-9652-07cfbf0c0bda.

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.

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

Eastman, Peggy. 2023. "NCI Releases New National Cancer Plan to Realize Vision of Cancer Moonshot." LWW.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://github.com/sfirke/janitor.

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

Porta, Miquel. 2014. *A Dictionary of Epidemiology*. Oxford university press.

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

Talic, Stella, Shivangi Shah, Holly Wild, Danijela Gasevic, Ashika Maharaj, Zanfina Ademi, Xue Li, et al. 2021. "Effectiveness of Public Health Measures in Reducing the Incidence of Covid-19, SARS-CoV-2 Transmission, and Covid-19 Mortality: Systematic Review and Meta-Analysis." *Bmj* 375.

Vargas, Ashley J, Sheri D Schully, Jennifer Villani, Luis Ganoza Caballero, and David M Murray. 2019. "Assessment of Prevention Research Measuring Leading Risk Factors and Causes of Mortality and Disability Supported by the US National Institutes of Health." *JAMA Network Open* 2 (11): e1914718–18.

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.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

Yao, Yuling, Aki Vehtari, Daniel Simpson, and Andrew Gelman. 2017. "Using Stacking to Average Bayesian Predictive Distributions." *Bayesian Analysis*. https://doi.org/10.1214/17-BA1091.