Mortality in Alberta: Is the leading cause of death in the 2000s still terrifying?*

Shiji Zhang, Zongcheng Cao

March 16, 2024

Abstract

Medical technology continues to advance. The leading causes of death in the 2000s appear to pose a decreasing threat to individuals. In this report, we selected 2003 as representative of the 2000s and chose the top five causes of death for that year. Using negative binomial and Poisson models, we studied the impact of these causes on mortality over the past 20 years and used regression models to predict future trends. The research findings suggest that the progress in medical technology has not stopped malignant neoplasms and chronic diseases from harming humans. Governments should urge people to pay attention to health issues and raise awareness of disease prevention.

Table of contents

*Code and data are available at: https://github.com/Northboi/The-Mortality-in-Alberta

1 Intoduction

With the advancement of medicine, it seems that the threat of various diseases to human mortality is decreasing, especially those common ailments that have existed since ancient times (Montero et al. 2024). Our report focuses on studying the top 5 causes of mortality in the 2000s, represented by the year 2003. The reason for selecting 2003 as a representative year is because of the major infectious disease event of the 2000s, namely SARS (Costa, Moreli, and Saivish 2020). We found that in the 2000s, the leading causes of death were various chronic diseases, with malignant neoplasms and acute myocardial infarction also being significant contributors to mortality during this period.

In this report, we first introduced the source of our dataset and provided an overview of each variable in the dataset through tables. We displayed the top 15 causes of death in Alberta province in 2003 and selected the top five causes of death as the estimand to explore their associated risks. In the model section, we established linear regression models using these five causes as predictors and employed negative binomial and Poisson methods to study their effects on the number of deaths. Through analysis, we found that among these five causes, chronic diseases had the greatest impact on the number of deaths, while the impact of stroke appeared to be the smallest. We then used line graphs to illustrate the annual number of deaths caused by these causes from 2001 to 2022, showing the variation in the number of deaths caused by each cause. We found that the number of deaths caused by these diseases did not significantly decrease over the span of 20 years, and in fact, the number of deaths caused by chronic diseases and malignant neoplasms showed an increasing trend, which aligned with our conclusions in the model section.

Through our report, we aim to raise awareness of whether medical advancements truly have significant effects on reducing mortality. We also hope that our analysis can contribute to medical policies and healthcare environments in Alberta and even Canada by highlighting which diseases are most likely to threaten our lives. This can ultimately enhance people's awareness of health issues and disease prevention.

2 Data

2.1 Backgrounds

The open dataset used for this analysis is named "Leading Causes of Death" from the government of Alberta. The dataset gives information about the 30 leading causes of mortality in the province of Alberta, and provides a ranking of these death causes by the total number of deaths. The causes of death used in the dataset is based on the International Classification of Diseases 10th Edition. The dataset included complete information from 2001 to 2022, and it is updated annually from 2001, the last update happened in September, 2023. Therefore, ever since the year of 2001, 30 leading reasons of death will be added to the dataset every year, with the number of deaths caused by there reasons, and there are chances for a cause to exist in the ranking for multiple years. A variable "n" in the dataset will count the number of times for each specific cause of death to enter the ranking.

Data cleaning process is applied after loading the data. Since the dataset contains an observation NA and three (blank) in the year of 2014, 2015 and 2018, which can affect the process and result of this analysis. Therefore, all the observations with NA or (blank) are deleted from the dataset before any analysis. The reason we do not use imputation or any other way to fix the missing data is because it appears in the 'cause' column, which is a 'string' column. Therefore, we are not going to use imputations or other methods and we deleted the rows of these missing data instead, and there is no data wasted since the missing data appears in a 'string' column.

Our report's original dataset is sourced from the Open Government program of Alberta province(2024). And the R packages we used for programming this report by R language(R Core Team 2020) include following packages:tidyverse(Wickham et al. 2019), boot(Canty and Ripley 2022), broom.mixed(Bolker et al. 2021), collapse(Krantz 2021), dataverse(Leeper 2021), gutenbergr(Robinson 2021), janitor(Firke 2021), knitr(Xie 2021), marginaleffects(Arel-Bundock 2022a), modelsummary(Arel-Bundock 2022b), rstanarm(Team 2022), ggplot2(Wickham 2016), lubridate(Garrett Grolemund 2021), kableExtra(Zhu et al. 2024) and gridExtra(Auguie and Antonov 2017).

2.2 Variables

For the variable, there are 5 of them in the original dataset.

- Calendar Year: A string variable, indicates the specific year of each cause of death was recorded in the ranking.
- Cause: A string variable, provides the name of each reason for death, if a leading cause entered the ranking for more than one year, then there will be multiple records with the same "Cause" variable.
- Ranking: An integer variable from 1 to 30, demonstrates the place of the cause of death in the ranking for a specific year.
- Total Deaths: An integer variable,
- n: An integer variable, counts the number of times for a specific cause of death enter the ranking, which is equivalent to the number of times for a specific cause of death appears in the dataset until the latest update.

The (Table 1) shows the types and descriptions for these variables.

Table 1: Variable Description for raw data

Column	Type	Description
calendar year	num	Indicating the specific year of each cause of death was recorded in the ranking.
causes	str	Providing the name of each reason for death. If a leading cause entered the ranking for more than one year, then there will be multiple records with the same 'Cause' variable.
ranking	num	Demonstrating the place of the cause of death in the ranking for a specific year.
total deaths	num	Counting the number of times a specific cause of death entered the ranking, which is equivalent to the number of times a specific cause of death appears in the dataset until the latest update.
n	num	Counting the number of times a specific cause of death entered the ranking, which is equivalent to the number of times a specific cause of death appears in the dataset until the latest update.

3 Model

3.1 Model Choices

As reviewing the processed dataset, the leading causes of death in the year 2003 were extracted from the dataset, as table 2 shows.

To explore the relationship between these factors with the expected value of the total death caused by these factors, two regressions will be constructed around the total death and the top five death factors, which are Poisson regression and negative binomial regression. Specifically, the five factors are "All other forms of chronic ischemic heart disease", "Malignant neoplasms of trachea, bronchus and lung", "Acute myocardial infarction", "Stroke, not specified as hemorrhage or infarction", "Atherosclerotic cardiovascular disease, so described" in ascending order.

count data refers to the frequency of an event occurring within a given time or space. As discussed in the previous part, the dependent variable in this regression, which is the "Total Deaths" variable, is a counting type of data since it counts the number of deaths caused by the corresponding factor. Specifically, count data such as the "Total Deaths" variables do not contain any negative numbers. Under such circumstances, a particular form of distribution can yield an effective model, namely Poisson distribution. Poisson distribution "takes on a probability value only for nonnegative integers; this characteristic of the Poisson distribution makes it an excellent choice for modeling count outcomes, which only take on integer values of 0 or greater"(Coxe, West, and Aiken 2009). Therefore, In deciding the appropriate model for this analysis, Poisson regression emerges as a fitting model given the dataset and statistical requirements, because "the simplest such regression model for counted data is Poisson regression" (Gardner, Mulvey, and Shaw 1995).

Even though Poisson Regression has several advantages for constructing regression for this dataset, on the other hand, Poisson regression "is denned by a highly restrictive model for the variance of the dependent variable, and badly misleading conclusions might be drawn if, as is likely, the data are inconsistent with this model"(Gardner, Mulvey, and Shaw 1995). To prevent any mistakes in the Poisson regression, and contrast the results with it, another regression will be used in this analysis along with the Poisson regression, the negative binomial regression. Similar to Poisson regression, negative binomial regression is a fitting model for count data, but with more advantages. "The negative binomial can be viewed as a form of Poisson regression that includes a random component reflecting the uncertainty about the true rates at which events occur for individual cases"(Gardner, Mulvey, and Shaw 1995). Thus, both Poisson regression and negative binomial regression will be applied to the dataset, and there will be a comparison between the two.

Table 2: Top 15 causes of death in Alberta in 2003 .

Year	Cause	Ranking	Deaths	Years
2003	All other forms of chronic …	1	1,749	22
2003	Malignant neoplasms of trac…	2	1,257	22
2003	Acute myocardial infarction	3	1,242	22
2003	Stroke, not specified as he…	4	760	22
2003	Atherosclerotic cardiovascu…	5	709	22
2003	Other chronic obstructive p…	6	670	22
2003	Diabetes mellitus	7	455	22
2003	Malignant neoplasm of breast	8	419	22
2003	Other malignant neoplasms o…	9	394	16
2003	Malignant neoplasms of colon	10	364	22
2003	Malignant neoplasms of pros…	11	362	22

Table 2: Top 15 causes of death in Alberta in 2003 .

Year	Cause	Ranking	Deaths	Years
2003	Alzheimer's disease	12	343	22
2003	Pneumonia due to other or u...	13	328	22
2003	Malignant neoplasms ofpancreas	14	294	10
2003	Organic dementia	15	281	22

3.2 Poisson model

Table 3: The description for each predictor in the model

Predictors	Descriptions
x1	All other forms of chronic ischemic heart disease
x2	Malignant neoplasms of trachea, bronchus and lung
x3	Atherosclerotic cardiovascular disease, so described
x4	Stroke, not specified as hemorrhage or infarction

Poisson regression is employed in the dataset to construct a model in the first place,This Equation 1 shows the relationship between the expectation of total deaths and the probability of each cause occurring. and is in the linear regression form.

$$log(E(y)) = 7.037 + 0.446x_1 + 0.223x_2 - 0.436x_3 - 0.531x_4 \tag{1}$$

To investigate the impact of a specific cause on total deaths individually, we exponentiate both sides of the equation with base e and make it the Equation 2.

$$E(y) = e^{7.037+0.446x_1+0.223x_2-0.436x_3-0.531x_4} \tag{2}$$

This analysis investigates how the top five causes of death of 2003 affects the expectation value of total deaths. y represents the dependent variable, which is the Total Death Variable, Given that there are five causes, the effect of the "Acute myocardial infarction" factor will be included in the intercept, while the four factors represent the following terms with dummy variables. As the (Table 3) shows, x1 represents "All other forms of chronic ischemic heart disease", x2 represents "Malignant neoplasms of trachea, bronchus and lung", x3 represents "Atherosclerotic cardiovascular disease, so described", and x4 represents "Stroke, not specified as hemorrhage or infarction". Although x1, x2, x3, and x4 represent different causes, the common interpretation of x is the probability of occurrence of a particular cause. When we want to investigate how many total deaths a specific cause would lead to, we simply set the probability of occurrence of that cause (x1) to 1, while setting the probabilities of all other causes (x2, x3, x4) to 0. In this way, the resulting E(y) represents the number of total deaths when only that particular cause occurs.

For example, if a research aims to replicate the expected total death associated with one of the factors, such as "Malignant neoplasms of trachea, bronchus and lung", then let the other dummy variables equal to 0, so x1=0, x3=0, x4=0, and x2 =1. The expression for the total deaths will be Equation 3.

$$E(y) = e^{7.037+0.223*1} \tag{3}$$

Table 4: Modeling the most prevalent cause of deaths of year 2003 in Alberta, 2001-2020.

	Poisson	Negative binomial
(Intercept)	7.037	7.038
		(0.039)
causeAll other forms of...	0.446	0.446
		(0.054)
causeMalignant neoplas...	0.223	0.223
		(0.053)
causeAtherosclero...	−0.436	−0.436
		(0.054)
causeStroke...	−0.531	−0.532
		(0.054)
Num.Obs.	110	110
Log.Lik.	−1080.186	−698.855
ELPD	−1110.0	−701.6
ELPD s.e.	88.7	4.6
LOOIC	2220.0	1403.2
LOOIC s.e.	177.3	9.2
WAIC	2220.5	1403.1
RMSE	109.33	109.33

After calculating this equation, the expected total death caused by "Malignant neoplasms of trachea, bronchus and lung" in 2003 can be replicated.

3.3 Negative Binomial model

As described above, to ensure the accuracy, a negative binomial model is constructed along with the Poisson regression model. And Equation 4 shows the relationship between the expectation of total deaths and the probability of each cause occurring, and is in the linear regression form.

$$log(E(y)) = 7.038 + 0.446x_1 + 0.223x_2 - 0.436x_3 - 0.532x_4 \tag{4}$$

Samely, we exponentiate both sides of the equation with base e and make it the Equation 5

$$E(y) = e^{77.038+0.446x_1+0.223x_2-0.436x_3-0.532x_4} \tag{5}$$

Similar to the Poisson model, the y indicates the Total Death Variable, which means that E(y) is the expected total deaths. Moreover, within the five major causes, the impact of "Acute myocardial infarction" will be incorporated into the intercept.As the (Table 3) shows, the four remaining factors will be denoted by dummy variables: "All other forms of chronic ischemic heart disease" is symbolized by x1, "Malignant neoplasms of trachea, bronchus and lung" symbolized as x2, "Atherosclerotic cardiovascular disease, so described" by x3 and "Stroke, not specified as hemorrhage or infarction" by x4.

With identical steps as the Poisson regression, if a research is interested to replicate the expected total death caused by a factor such as "All other forms of chronic ischemic heart disease" in 2003, then the expression can be modified by letting the other dummy variables equal to 0, so x2=0, x3=0, x4=0 and letting x1=1. The expression will be Equation 6

$$E(y) = e^{7.038+0.446*1} \tag{6}$$

Then, the expected total death caused by "All other forms of chronic ischemic heart disease" in 2003 in Alberta can be replicated.

3.4 Poisson vs Negative Binomial

Based on the information above, we have created a figure of the Poisson model and negative binomial model as shown in the (Figure 1).On the graph, there are two lines: a solid black line representing the observed data values and a set of lighter blue lines representing the values obtained through modeling for prediction. First, we can observe that in both the Poisson and negative binomial models, there is not much difference between the lighter blue lines and the black line at any given number of deaths. There are no apparent residual patterns reflected in these two graphs. This indicates that both models are at least suitable for our data.

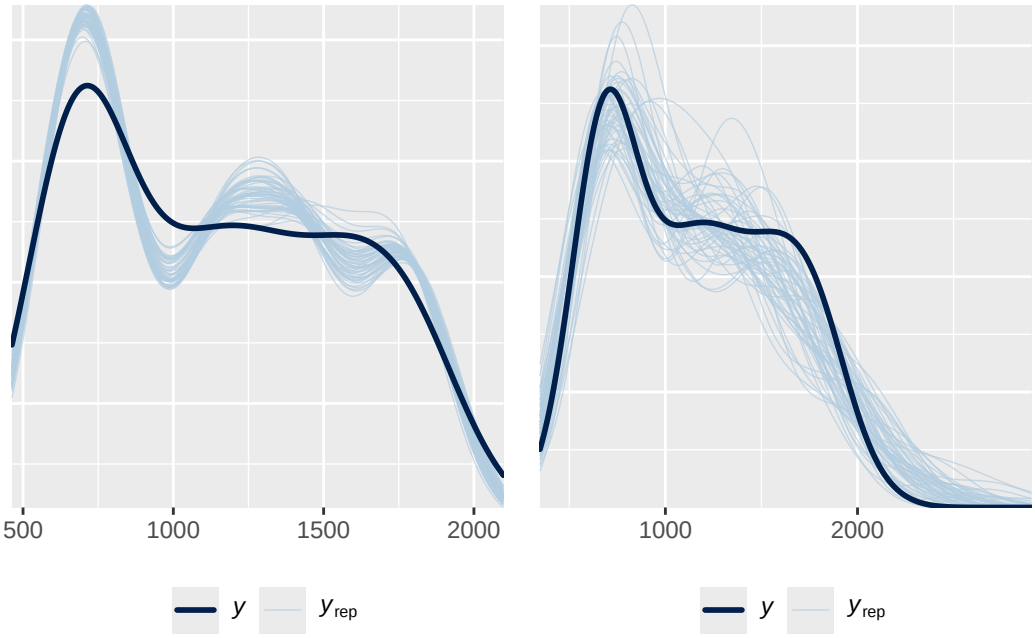

Figure 1: Comparing posterior prediction checks for Poisson and negative binomial models

However, upon closer analysis, it becomes apparent that the Poisson model does not predict the data as precisely as the negative binomial model. In the (Figure 1), the Poisson model, the upper one, shows that the lighter blue lines representing predicted data roughly coincide with the solid black line representing observed data at the beginning and end. However, in the middle part, specifically in the range of deaths from approximately 600 to 1750, the predicted data's lighter blue lines exhibit some fluctuations, despite still following a general similar trend as the observed data represented by the black line.

In contrast, the negative binomial model depicted at the bottom of the (Figure 1) does not exhibit big fluctuations. The lighter blue lines representing predicted data overall closely overlap with the solid black line representing observed data. It is worth noting that the negative binomial model has a larger spread compared to the Poisson model. This indicates that the negative binomial model

takes into account more uncertainty and variability in predictions. Consequently, the predictions from the negative binomial model may be more accurate than those from the Poisson model.

To better compare the accuracy of the Poisson and negative binomial models in Alberta's cause of death dataset, we utilize the (Table 5) to compare the ELPD difference and SE difference of the two models. ELPD stands for Expected Log Predictive Density and its value reflects the likelihood of generating new data under the current model. In other words, it indicates to what extent a model considers the variability in the data. When a model has a larger ELPD difference, it means that the model's predictive accuracy is higher. SE stands for Standard Error. When ELPD and SE are presented together, the magnitude of the SE difference can be understood as how much the ELPD value might vary if the model is applied to other different datasets. Therefore, if a model has a smaller SE difference, it demonstrates better stability in prediction.

Table 5: ELPD and SE difference in Negative Binomial and Poisson models.

	ELPD difference	SE difference
Alberta CoD under Neg. Binomial	0.0000	0.00000
Alberta CoD under Poisson	-408.4195	85.72368

The (Table 5) shows that the negative binomial model has a larger SE difference compared to the Poisson model. This is because the negative binomial model can better handle overdispersion, which aligns with our previous conclusion that the negative binomial model has a larger spread. On the other hand, the Poisson model has a smaller SE difference, indicating that its ELPD measurement uncertainty is smaller. The negative binomial model can better accommodate the variability in the data without seemingly increasing prediction uncertainty. Therefore, this further illustrates that in this scenario, the negative binomial model is more suitable for prediction compared to the Poisson model.

4 Result

4.1 Replicate result

After analyzing the critical numerics of the negative binomial and Poisson models, we now utilize the formulas of both models to replicate the total deaths caused by the top five causes of death in 2003. (Table 6) displays the predicted values for each cause under each model based on formulas 1 and 2. By comparing these predicted values with the actual deaths caused by these causes in 2003 as shown in (Table 2), we found that the predictions from the negative binomial model seem to be more accurate only in the case of malignant neoplasms deaths (1420 is closer to 1257 total deaths than 1422). However, for the other four causes, the Poisson model exhibits slightly higher prediction accuracy.

This seems contrary to the conclusions drawn from (Figure 1) and (Table 5). Previously, we concluded that the negative binomial model had a larger elpd difference and a smaller SE difference, indicating it should be more accurate in predictions compared to the Poisson model. In fact, our previous conclusion was not incorrect. Although the negative binomial model with larger elpd and smaller SE may have better predictive capabilities when facing new data, it does not necessarily mean it has better replicative abilities. In this scenario, we attempted to replicate the deaths caused by the top five ranked causes in 2003 using formulas 1 and 2. The Poisson model demonstrates better replicative abilities on the existing current dataset. On the other hand, the negative binomial

Table 6: The replicated total deaths of top 5 causes of death in Alberta in 2003

Cause	Poisson	Negative.Binomial
Chronic	1777.5655	1779.3439
Neoplasms	1422.2565	1420.8350
Cardiovascular	735.8307	735.8307
Stroke	669.1445	668.4757
Myocardial Infarction	1137.9686	1136.8312

model, with its larger spread and consideration of more instability and variability, may need to predict when facing new data to showcase its advantages.

4.2 The trend in reality

Through our analysis in the model section using negative binomial and Poisson methods and according to (Table 4), we found that chronic diseases and Malignant neoplasms made the greatest contribution to the number of deaths each year in 2003 and throughout the entire 2000s. In other words, they may not have had the highest mortality rates, but they were the leading causes of death in Alberta province during that time. To further investigate the top five causes of death in 2003, we utilized line graphs to observe the trends in the number of deaths caused by these five causes from 2001 to 2022.

The (Figure 2) consists of five different color of lines, each displaying the trend of deaths caused by different causes over the years. The x-axis represents the years from 2001, the start of the dataset records, to 2022, the end of the dataset records. The y-axis represents the number of deaths attributed to each cause in Alberta province each year. The y-axis ranges from a minimum of 500 deaths to a maximum of 2000 deaths.

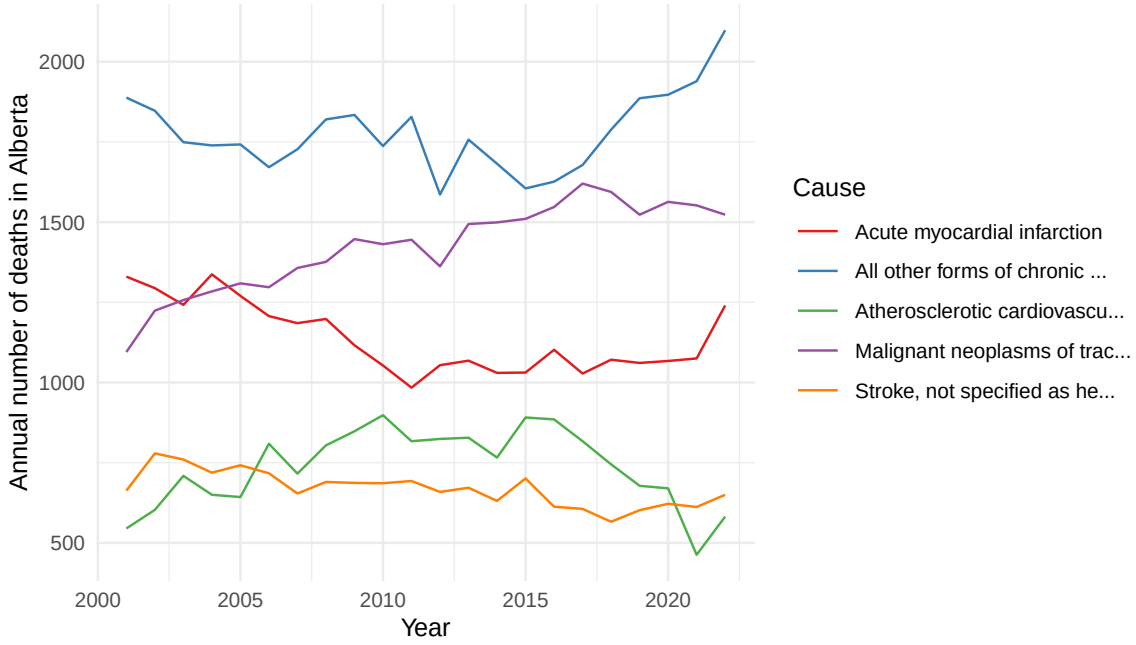

Figure 2: Annual number of deaths for the top-five causes in 2003, since 2001, for Alberta, Canada

When observing these five line graphs, we can see that the overall trends of two of them are upward. Firstly, in 2001, the first year of data collection, the number of deaths attributed to chronic diseases in Alberta province did not exceed 2000 people. Even in 2003, the year we selected, the number of deaths due to chronic diseases was only 1749 people. However, by the year 2022, we can see that the number of deaths due to chronic diseases in Alberta province has exceeded 2000 people. Similarly, Malignant neoplasms have also shown an upward trend over the past twenty years. Starting from 2001, the number of deaths due to Malignant neoplasms in Alberta province was only around 1200 people. By 2003, the number of deaths had slightly increased to 1257 people. However, by 2022, the number of deaths had exceeded 1500 people.

Apart from chronic diseases and Malignant neoplasms, the other three causes seem to have a general downward trend of varying sizes over the past twenty years. acute myocardial infarction had over 1250 deaths in 2001, which decreased slightly to 1242 deaths by 2003. By 2022, the number of deaths due to acute myocardial infarction remained below 1250. Atherosclerotic cardiovascular disease had around 500 deaths in the year data collection began in Alberta province, but the number of deaths increased to 709 in 2003. Interestingly, there seemed to be a slight upward trend in the number of deaths due to atherosclerotic cardiovascular disease from 2005 to 2015, although it did not exceed 1000 deaths per year. By around 2021, the number of deaths due to this disease dropped below 500, but in 2022, it rebounded to over 500 deaths. As for stroke, the number of deaths in Alberta province has remained relatively stable, ranging from about 500 to 750 deaths per year. Stroke appeared to peak in 2003 with 760 deaths, which allowed it to enter the top five causes of death that year, which seems to be an uncommon occurrence over the 20-year period.

Overall, the top five causes of death in Alberta in 2003 have not shown a general downward trend over these 20 years, especially the two leading causes of death, chronic diseases and malignant neoplasms, which remain much higher in fatalities than the other three causes and are positioned high in figure 2. What is more concerning is that the annual number of deaths from chronic diseases and malignant neoplasms in Alberta seems to have maintained an overall upward trend over these 20 years.

5 Discussion

5.1 Findings

After modeling and visualizing the Poisson and Negative Binomial models for the top five causes of death in Alberta in 2003, we found that there has been no significant improvement in the top five causes of death throughout the 2000s, even until 2022. In fact, the top two causes of death in 2003, chronic diseases and malignant neoplasms, have shown a worsening trend, with the number of deaths increasing each year. This challenges the notion that advancements in medical technology have reduced mortality rates from diseases in Alberta province (Montero et al. 2024).According to (Figure 2), the number of deaths attributed to gout surpassed those related to cardiovascular diseases around 2020 in Alberta province, and it continues to show an accelerating upward trend. We believe that relevant governments and organizations should enhance public awareness and preventive measures regarding gout to prevent further increases in the number of deaths caused by this condition.

In our created Poisson and Negative Binomial models, we arrived at an interesting conclusion. Although the Negative Binomial model exhibited larger ELPD difference and smaller SE difference compared to the Poisson model according to (Table 5) , this does not necessarily mean that the Negative Binomial model is more accurate in replicating total deaths for a specific year than the Poisson model. The advantage of the Negative Binomial model lies in its consideration of more variability and instability in the data, which may lead to higher accuracy in predicting new data.

However, when replicating existing data, the Poisson model appears to be more accurate than the Negative Binomial model. Therefore, if governments and organizations intend to model causes based on past total deaths, we recommend using the Poisson model. If they aim to predict future total deaths, we suggest using the Negative Binomial model.

5.2 Suggestions for the future

Seeing such a large number of deaths due to malignant neoplasms and chronic diseases in Alberta province over the past 20 years, we believe that the government and relevant institutions should take these diseases seriously. According to the Alberta's Tomorrow Project (ATP), a study published in 2021 found that chronic diseases continue to show an increasing trend among selected research populations in the province (Ye et al. 2021). Based on past data, chronic diseases have caused significant harm to people in Alberta. One particular aspect of chronic diseases is their difficulty to cure. Chronic diseases may recur, causing patients to suffer repeatedly. Conditions such as hypertension, depression, and diabetes are examples. These diseases not only cause physical suffering but also have a profound impact on mental health, especially conditions like depression (Ye et al. 2021), which can severely affect daily behavior and life. The ATP project found that depression is a recurrent disease clinically, especially among women.

Among the issues worth paying attention to in young populations, besides depression, anxiety also deserves serious consideration (Roberts et al. 2015). The prevalence of social media in today's society may bring joy to young people, but it also causes unnecessary anxiety. For instance, browsing social networking sites can lead to feelings of inferiority due to others' seemingly glamorous lives, thus triggering anxiety. Academic pressure and potential family-related stressors are also underlying causes of anxiety in young individuals. If these psychological issues are not addressed and treated properly, they may lead to physical discomfort and even more serious consequences.

There is no doubt that we believe the government of Alberta or Canada needs to make efforts in addressing chronic diseases. Firstly, and most importantly, we believe the government should raise public awareness. More promotional activities should be conducted to increase people's understanding of chronic diseases. Utilizing articles like this report to warn people and make them aware that chronic diseases have posed a significant threat to people's health in Alberta over the past 20 years could be effective. In addition to raising public awareness, we believe the government should allocate more resources. For example, the government or relevant agencies can train and hire more professionals in preventive medicine, allowing citizens to quickly receive responses to inquiries about chronic diseases. This could be considered a cost-effective treatment method.

For diseases like depression and anxiety, which pose more psychological harm to people, government efforts are crucial. It's important to increase awareness of these illnesses, especially among parents. Many parents may not realize the severity of the problem when their children show signs of depression, potentially missing the optimal treatment window and leading to dire consequences. The government can also enact some regulations, such as requiring workplaces to provide health training and allowing employers to offer more flexible working conditions to employees. These improvements can have a substantial impact on alleviating depression and anxiety.

Moreover, "Malignant neoplasms of trachea, bronchus and lung" ranked second in the mortality rate of this study. Illnesses such as lung cancer and tracheal cancer are under this specific factor. Specifically, lung cancer is a commonly occurring condition that has the probability to be prevented from happening, and reduce the incidence rate. There are some causes that will lead to the development of lung cancer, such as smoking and radon, which the government has potential to take into regulation.

Smoking is a habitual action for a large amount of the population that may cause lung cancer.

Undoubtedly, the most efficient way of preventing lung cancer caused by smoking is to quit smoking, which hardly depends on individual discipline. Still, there exists some action for the government to reduce the risk of smoking. For example, "reducing cigarette smoking can be accomplished through preventing adolescent initiation of cigarette smoking or through facilitating cessation among adult smokers"(Burns 2000). The government has the ability to spread more information about the risk of smoking through social media, which can affect the younger generation to reduce new smokers. With "increasing rates of cessation in the population, particularly cessation at younger ages, will lead to a decline in lung cancer death rates, a phenomenon already evident for males in the U.S"(Burns 2000).

"Radon is a colorless, odorless, gaseous decay product of uranium found normally in soil. It can accumulate in enclosed spaces such as homes, schools, and workplaces."(Peterson et al. 2013). Along with smoking, radon is also a factor that has the risk of lung cancer. The exposure to radon has the risk of causing lung cancer. Particularly, The risk of radon is significant in Canada, "calculations suggest that a mean of 13.6 % or 847 lung cancer deaths in Ontario are due to radon"(Peterson et al. 2013). Guidelines already exist for the regulation of radon, such as the Canadian guideline which suggests the radon level to be below 200Bq/m^3, and the WHO guideline that indicates the radon level to be less than 100Bq/m^3. However, measurement of radon is not popularized in Canada. "If all homes above 200Bq/m^3, the current Canadian guideline, were remediated to background levels, it is estimated that 91 lung cancer deaths could be prevented each year, 233 if remediation was performed at 100Bq/m^3"(Peterson et al. 2013). Therefore, the government can reduce the risk of lung cancer by dedicating efforts to popularize the use of indoor radon measurement devices and increase the frequency of testing. With such actions, total deaths caused by "Malignant neoplasms of trachea, bronchus and lung" have a high chance to decline in the future.

5.3 Drawbacks

It's undeniable that this report still has some drawbacks in its design process, in terms of the assumptions and limitations of the model, as well as in aspects of missing data.Whether it's the Poisson or the negative binomial model, one of their assumptions emphasizes the independence of events. However, in the cause of death column of the mortality of Alberta dataset mentioned in this report, it seems that not every cause is necessarily independent of the others. In other words, the occurrence of one cause might affect the probability of another cause occurring. For example, a stroke might increase the likelihood of acute myocardial infarction(Muñiz 2012), indicating that the occurrences of each cause are not independent of one another. A limitation of the Poisson model is that its characteristic of simplifying real-world events might oversimplify them, thereby failing to consider many factors that exist in reality. For instance, the number of deaths due to certain seasonal causes might fluctuate with the seasons, a factor that the Poisson model might overlook.

In the data section, we have already mentioned that there are some missing data in the original mortality of Alberta dataset. Some are present in the form of NA, while three instances appear as (Blank) in the data for 2014, 2015, and 2018. Moreover, these data absences occur in the "cause" column, making it difficult for us to impute them, as we do not know which causes to attribute their ranking and total deaths to. Therefore, if we had a dataset with more complete data, we could conduct a more accurate analysis and modeling.

Reference

2024. *Alberta.ca*. https://www.alberta.ca/open-government-program.

Arel-Bundock, Vincent. 2022a. *Marginal Effects for Regression Models*. https://CRAN.R-project.org/package=marginaleffects.

———. 2022b. *Summary Tables and Plots for Statistical Models and Data: Beautiful, Customizable, and Publication-Ready*. https://CRAN.R-project.org/package=modelsummary.

Auguie, Baptiste, and Anton Antonov. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics*. https://cran.r-project.org/package=gridExtra.

Bolker, Ben et al. 2021. *Tidying Methods for Mixed Models*. https://CRAN.R-project.org/package=broom.mixed.

Burns, David M. 2000. "Primary Prevention, Smoking, and Smoking Cessation: Implications for Future Trends in Lung Cancer Prevention." *Cancer* 89 (S11): 2506–9.

Canty, Angelo, and Brian D. Ripley. 2022. *Bootstrap Functions (Originally by Angelo Canty for s)*. https://CRAN.R-project.org/package=boot.

Costa, Vivaldo Gomes da, Marcos Lázaro Moreli, and Marielena Vogel Saivish. 2020. "The Emergence of SARS, MERS and Novel SARS-2 Coronaviruses in the 21st Century." *Archives of Virology* 165 (7): 1517–26.

Coxe, Stefany, Stephen G West, and Leona S Aiken. 2009. "The Analysis of Count Data: A Gentle Introduction to Poisson Regression and Its Alternatives." *Journal of Personality Assessment* 91 (2): 121–36.

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

Gardner, William, Edward P Mulvey, and Esther C Shaw. 1995. "Regression Analyses of Counts and Rates: Poisson, Overdispersed Poisson, and Negative Binomial Models." *Psychological Bulletin* 118 (3): 392.

Garrett Grolemund, Hadley Wickham. 2021. *Make Dealing with Dates a Little Easier*. https://CRAN.R-project.org/package=lubridate.

Krantz, Sebastian. 2021. *Advanced and Fast Data Transformation*. https://CRAN.R-project.org/package=collapse.

Leeper, Thomas. 2021. *Client for Dataverse 4 Repositories*. https://CRAN.R-project.org/package=dataverse.

Montero, David A, Roberto M Vidal, Juliana Velasco, Leandro J Carreño, Juan P Torres, Angel A Oñate, and Miguel O'Ryan. 2024. "Two Centuries of Vaccination: Historical and Conceptual Approach and Future Perspectives." *Frontiers in Public Health* 11: 1326154.

Muñiz, Antonio E. 2012. "Myocardial Infarction and Stroke as the Presenting Symptoms of Acute Myeloid Leukemia." *The Journal of Emergency Medicine* 42 (6): 651–54.

Peterson, Emily, Amira Aker, JinHee Kim, Ye Li, Kevin Brand, and Ray Copes. 2013. "Lung Cancer Risk from Radon in Ontario, Canada: How Many Lung Cancers Can We Prevent?" *Cancer Causes & Control* 24: 2013–20.

R Core Team. 2020. *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. https://www.R-project.org/.

Roberts, KC, DP Rao, TL Bennett, Lidia Loukine, and GC Jayaraman. 2015. "Prevalence and Patterns of Chronic Disease Multimorbidity and Associated Determinants in Canada." *Health Promotion and Chronic Disease Prevention in Canada: Research, Policy and Practice* 35 (6): 87.

Robinson, David. 2021. *Download and Process Public Domain Works from Project Gutenberg*. https://CRAN.R-project.org/package=gutenbergr.

Team, Stan Development. 2022. *Bayesian Applied Regression Modeling via Stan*. https://CRAN.R-project.org/package=rstanarm.

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.

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

Ye, Ming, Jennifer E Vena, Jeffrey A Johnson, Grace Shen-Tu, and Dean T Eurich. 2021. "Chronic Disease Surveillance in Alberta's Tomorrow Project Using Administrative Health Data." *International Journal of Population Data Science* 6 (1).

Zhu, Hao et al. 2024. *kableExtra: Construct Complex Table with 'Kable' and Pipe Syntax*. https://CRAN.R-project.org/package=kableExtra.