Temporal Trends in Heart Disease and Diabetes Mortality in Alberta: A comparison between Poisson and Negative Binomial techniques*

Alexander Sun

April 1, 2024

In this study, we analyzed mortality data from Alberta, Canada, focusing on the effects of heart disease and diabetes from 2016 to 2021, amidst increasing fast food consumption. Our findings, using Poisson and negative binomial regression models, indicate that while heart disease is positively associated with mortality rates, diabetes, and other cardiovascular conditions are linked to lower mortality rates relative to the baseline, suggesting effective management of these conditions over time. This research illustrates the complex relationship between diseases and health outcomes, highlighting balance between medical advancement and deteriorating dietary habits. By shedding light on these temporal trends, our study contributes to a broader understanding of how diet influences the prevalence and mortality rates of heart disease and diabetes, reinforcing the importance of dietary education and regulation in public health initiatives.

Table of contents

---

*https://github.com/alexandersunliang/Temporal-Trends-in-Heart-Disease-and-Diabetes-Mortality-in-Alberta.git

1 Introduction

In recent decades, the prevalence of heart disease and diabetes has surged globally, a trend paralleled by the increasing consumption of fast food (Tsao et al. 2023). This paper focuses on Alberta, Canada, where these health challenges have become particularly pronounced. While studies have linked fast food consumption to various health outcomes, few have directly examined how mortality rates from heart disease and diabetes have changed within this region. This gap in research motivates our study, which aims to analyze trends in these mortality rates and discuss their potential association with the rise in fast food consumption, despite the absence of direct consumption data.

Using mortality data from the Alberta government, spanning two decades, we applied both Poisson and negative binomial regression models to analyze changes in heart disease and diabetes mortality rates (Government 2015). Our findings reveal significant trends in these rates, with notable effects that correspond to periods of reported national and global rises in fast food consumption and COVID-19. While direct causation cannot be established due to the lack of specific consumption data, the correlations emphasize the potential health impacts of dietary habits.

The importance of this research lies in its contribution to the ongoing dialogue about public health strategies aimed at combating heart disease and diabetes. By highlighting the temporal association between increased mortality rates and the era of rising fast food consumption, this study emphasizes the need for targeted public health interventions and policies. Furthermore, by focusing on the past five years, we can further investigate the potential relationship between COVID-19 and heart-related diseases/diabetes.

This paper is organized as follows: Following the introduction in Section 1, Section 2 provides an overview of our dataset, variables, and a general analysis of the data, establishing the theoretical foundation for the study. Section 3 describes the model, including the rationale behind the choice of negative binomial regression. The fourth section, Section 4 presents our findings, detailing the effects of our various illnesses on mortality rates. Finally, Section 5 discusses the implications of these findings for public health policy and suggests directions for future research. For this study, the estimand is the effect of heart diseases and diabetes on mortality rates.

Our goal is to provide an insight into how large macro-events from the past few years may have changed the mortality rates of diseases correlated with unhealthy eating.

2 Data

2.1 Data source

This analysis will be carried out in R (R Core Team 2023) using packages tidyverse (Wickham et al. 2019), rstanarm (Goodrich et al. 2022a), dplyr (Wickham et al. 2023), ggplot2 (Wickham 2016), knitr (Xie 2014), gridextra(Auguie 2017), bayesplot(Gabry et al. 2019), and modelsummary (Arel-Bundock 2022). The data set used in this paper is called *Leading causes of death* and was collected from the Alberta Provincial Government (Government 2015). The data set consists of a ranking of the 30 most common causes of death each year in Alberta. The government collected this data using death certificates from all over the province, making the information reasonably reliable and accurate. The data covers the last two decades, but for our research purpose we will focus on the last five years.

2.2 Broader Context of the Dataset

The availability of detailed public health data, such as the mortality statistics from Alberta, is crucial for the formulation of informed public health policies and strategies. Within the broader Canadian context, Alberta's commitment to data transparency enables a deeper analysis of health trends and outcomes, serving as a model for other provinces and territories. The analysis of mortality data plays a pivotal role in identifying health trends, assessing the burden of diseases, and planning public health interventions. By focusing on specific causes of death, researchers and policymakers can tailor strategies to target the underlying factors contributing to these trends, ultimately aiming to improve health outcomes and reduce preventable deaths.

2.3 Variables

The dataset comprises several key variables, central to this study's focus on heart disease and diabetes mortality rates:

Causes of Death: Specifically, the dataset categorizes mortality into detailed causes, including:

- **All Other Forms of Ischemic Heart Disease**: This category encompasses various conditions related to reduced blood flow to the heart muscle, excluding acute myocardial infarction.
- **Acute Myocardial Infarction (Heart Attack)**: Fatalities resulting directly from heart attack incidents.
- **Atherosclerotic Cardiovascular Disease**: Deaths caused by atherosclerosis, a condition characterized by the hardening and narrowing of the arteries due to plaque buildup, leading to cardiovascular problems.

Table 1: Yearly death counts for various heart diseases and diabetes

Year	Acute Myocardial Infarction	All Other Forms of Ischemic Heart Disease	Atherosclerotic Cardiovascular Disease	Diabetes Mellitus	Congestive Heart Failure
2016	1102	1626	885	502	352
2017	1028	1678	817	584	374
2018	1071	1788	630	577	347
2019	1061	1886	745	569	430
2020	1067	1897	678	743	387
2021	1075	1939	463	728	403

- **Diabetes Mellitus**: Mortality attributed to complications arising from diabetes, a chronic condition affecting blood sugar regulation.
- **Congestive Heart Failure**: Deaths resulting from the heart's inability to pump blood effectively, often a consequence of other heart conditions.

These five causes of death were chosen due to their correlation with unhealthy diets. We collected the recorded total deaths in Alberta from each of the above variables from 2016-2021 inclusive. This is shown in Table 1.

2.4 Data Preparation and Cleaning

The dataset was filtered to isolate deaths attributed to our causes of interest: acute myocardial infarction, all other forms of ischemic heart disease, atherosclerotic cardiovascular disease, diabetes mellitus, and congestive heart failure. This selection was crucial to align our study with its objectives, ensuring a focused examination of these specific health outcomes. Subsequently, we removed all missing values from the dataset. The final data points were compiled for use in our analysis later in the paper.

2.5 Potential Weaknesses in Data Collection

The data on Alberta's mortality rates may not fully capture all deaths, particularly those that occur unofficially or outside medical facilities. For example, deaths related to illegal activities may not be associated with a death certificate. This reliance on death/medical records could lead to an underestimation of actual mortality rates, affecting the dataset's completeness and the study's findings. Furthermore, causes of death that are interconnected with other symptoms may complicate which precise cause should be listed under the death certificate.

2.6 Preliminary Observations and Exploration

Our initial analysis revealed several notable observations. First, the trend analysis suggested a correlation between the years and mortality rates for specific causes of death, hinting at the possible influence of external factors such as healthcare policies or changes in societal health behaviors. Since we are looking at a five-year span from 2016 to 2021, the onset of the

COVID-19 pandemic introduced a plethora of health-related policies and regulations that may significantly impact our data. In Figure 1, we plotted the total deaths from each of the causes to visualize if there were any spikes in the data.
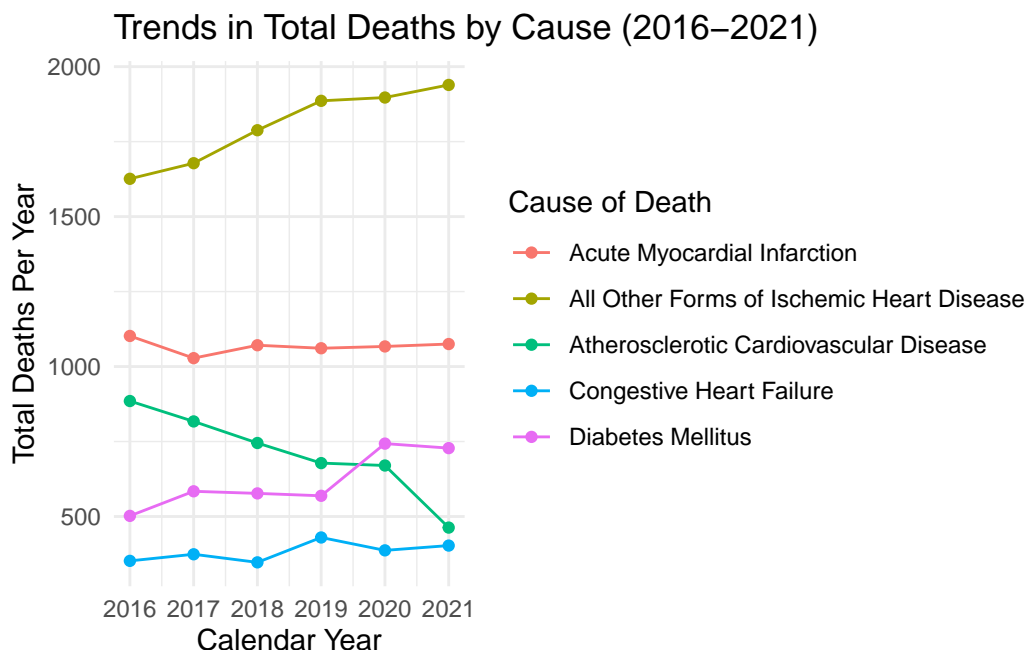

Figure 1: Trends in Total Deaths per Year by Cause from 2016-2021

From Figure 1, we see that although most causes kept their total death count per year roughly the same, Diabetes saw a noticeable spike form 2020 onwards and Atheroscleratic cardiovascular disease saw a significant decrease.

3 Model

In the realm of statistical modeling for count data, both the Poisson and Negative Binomial models are popular choices, each with its assumptions, benefits, and drawbacks. These models serve to elucidate the relationship between a set of predictors—in this case, the causes of death—and the count response variable, which is the total number of deaths per year.

The Poisson regression model is based on the assumption that the count of events follows a Poisson distribution, meaning its mean is equal to its variance. This assumption is suitable for datasets where the occurrence of events is rare and independent over a fixed period or space, such as the number of deaths in a small population or rare diseases. The simplicity of the Poisson model lies in its single parameter, making it computationally efficient and easy to interpret. However, this simplicity also constitutes a limitation, particularly when it comes to

overdispersion—a scenario often encountered in real-world data where the variance exceeds the mean due to heterogeneity in the data. When overdispersion is present, the Poisson model can underestimate the standard errors of the estimated coefficients, leading to an overstatement of the statistical significance of predictors.

In contrast, the Negative Binomial model adds an extra parameter to account for overdispersion, offering a more flexible fit for count data that exhibit greater variability. This model is especially apt when the data are counts of events that can occur more frequently than a rare event and when these counts have high variance. It can handle data from populations with heterogeneity that the Poisson model cannot accommodate. This flexibility comes with a trade-off in terms of increased model complexity and computational demand. The added dispersion parameter means that the Negative Binomial model has one more degree of freedom than the Poisson model, which can result in a better fit but also requires careful interpretation of the additional parameter.

For the current paper focusing on mortality rates due to diabetes and heart disease, it is imperative to choose a model that accurately reflects the underlying data distribution. If the mortality counts are subject to overdispersion, the Negative Binomial model would likely provide a more accurate representation of the data, enabling more reliable statistical inferences. For instance, if certain causes of death are more prevalent or exhibit more variability from year to year, the Negative Binomial model's capacity to incorporate this variability would make it a superior choice over the Poisson model. Ultimately, the selection between the two models should be guided by diagnostic checks, such as the dispersion parameter and goodness-of-fit tests.

3.1 Poisson Model

Listed below is what an example Poisson model would look like if applied to our scenario.

$$y_i|\lambda_i \sim \text{Poisson}(\lambda_i)$$

$$\log(\lambda_i) = \alpha + \beta_1 \times \text{AMI}_i + \beta_2 \times \text{OtherIHD}_i + \beta_3 \times \text{ASCVD}_i + \beta_4 \times \text{Diabetes}_i + \beta_5 \times \text{CHF}_i$$

where:

- y_i is the count of deaths due to the i-th cause.
- λ_i is the expected count of deaths for the i-th cause.
- α, β_1, β_2, β_3, β_4, and β_5 are the model parameters to be estimated.
- AMI, OtherIHD, ASCVD, Diabetes, and CHF are abbreviations of our previously selected variables.

3.2 Negative Binomial Model

To address potential overdispersion in our data, we also applied the Negative Binomial regression model. This model extends the Poisson by introducing an extra parameter to account for the overdispersion, offering a more flexible approach to fit our data. The formulation of the Negative Binomial model is:

$$y_i | \mu_i, \phi \sim \text{NegBin}(\mu_i, \phi) \tag{1}$$
$$\mu_i = \exp(\alpha + \beta x_i) \tag{2}$$
$$\alpha \sim \text{Normal}(0, 2.5) \tag{3}$$
$$\beta \sim \text{Normal}(0, 2.5) \tag{4}$$
$$\phi \sim \text{Exponential}(1) \tag{5}$$

In this model, y_i denotes the total number of deaths per year, with μ_i being the expected count adjusted for overdispersion through the dispersion parameter θ. $\beta\{0\}$ *represents the intercept and $\beta\{i\}$* represents the effects of a unique cause of death. The value and sign of β denote whether having that disease/sickness increase or decrease the death count of that year and by how much.

Given the observed overdispersion in our dataset, the Negative Binomial model is anticipated to offer a more accurate and reliable fit compared to the Poisson model. By incorporating the extra dispersion parameter, it allows us to better capture the variability in death counts across different causes, providing a nuanced understanding of how each cause contributes to overall mortality. In applying these models, we aim to discern the relative impact of specified causes of death on the total number of deaths, while also accounting for the distributional characteristics of our count data. Through this comparative analysis, we seek to identify the most suitable model for our dataset, thereby enhancing the reliability of our findings and conclusions.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Goodrich et al. (2022b). We use the default priors from `rstanarm`.

3.3 Model justification

Since we have two potential methods of modeling, we can directly compare the two to decide which regression to run. In the previous section, we hypothesized that the Poisson model will lead to more inaccurate results due to overdispersion of data. We can see if the assumption that variance is equal to the mean is true for our y_i value for total death count in that given year.

Table 3: Regression coefficients for Poisson and Negative Binomial

	Poisson	Negative Binomial
(Intercept)	6.973	6.976
		(0.124)
causeAll Other Forms of Ischemic Heart Disease	0.524	0.528
		(0.173)
causeAtherosclerotic Cardiovascular Disease	-0.408	-0.414
		(0.168)
causeCongestive Heart Failure	-1.028	-1.023
		(0.177)
causeDiabetes Mellitus	-0.548	-0.547
		(0.176)
Num.Obs.	30	30
Log.Lik.	-274.969	-196.422
ELPD	-304.8	-198.2
ELPD s.e.	60.9	2.9
LOOIC	609.7	396.5
LOOIC s.e.	121.7	5.8
WAIC	608.3	396.4
RMSE	89.92	90.10

Table 2: Summary Statistics of the Number of Yearly Deaths by Cause

Minimum	Mean	Maximum	Standard Deviation	Variance	Count
347	915.7333	1939	511.7452	261883.1	30

From Table 2, we see that the variance is significantly off of the mean. Therefore, the primary assumption for applying a Poisson model where the variance is equal to the mean does not hold. From this alone, we can reasonably conclude that a Negative Binomial model most likely fits our purpose of estimating the total death count per year better due to accounting for overdispersion with the μ_i variable.

4 Results

Table 3 presents a comparative summary of the regression results obtained from the Poisson and Negative Binomial models. The comparison is based on a dataset of mortality rates due to various causes of heart disease and diabetes. The table provides estimates of the regression

coefficients, along with their standard errors (in parentheses), for each cause of death, as well as several model diagnostics.

Intercept: Both models show a large positive intercept, implying a high baseline number of deaths when all other variables are at zero. This suggests that there are other factors contributing to mortality rates that are not included in the model.

Coefficient Estimates: The coefficients represent the log change in the expected count of total deaths for each cause of death. For instance, "All Other Forms of Ischemic Heart Disease" has a positive coefficient in both models, suggesting a higher mortality rate associated with this cause. "Atherosclerotic Cardiovascular Disease" and "Diabetes Mellitus" have negative coefficients, indicating a lower count of deaths relative to the baseline for these causes.

Standard Errors: The standard errors provide an indication of the precision of the coefficient estimates. The Negative Binomial model typically shows larger standard errors than the Poisson, reflecting additional uncertainty due to modeling overdispersion.

Num.Obs. indicates the number of observations used in both models, which is the same, providing a consistent basis for comparison.

Log.Lik. (Log-Likelihood) measures the fit of the model to the data, with higher values indicating a better fit. The Negative Binomial model shows a less negative log-likelihood, suggesting a better fit to the data compared to the Poisson model.

ELPD (Expected Log Predictive Density) and its standard error are measures for model comparison and predictive performance, where higher values indicate a better model. The Negative Binomial model has a higher ELPD, indicating better predictive performance.

LOOIC (Leave-One-Out Information Criterion) is similar to AIC but based on leave-one-out cross-validation; lower values indicate a better model fit. The Negative Binomial model has a lower LOOIC, suggesting a better fit.

WAIC (Watanabe-Akaike Information Criterion) is another model selection criterion that is fully Bayesian and accounts for model complexity; like AIC and LOOIC, lower is better. Again, the Negative Binomial model performs better.

RMSE (Root Mean Square Error) reflects the model's accuracy in predicting the number of deaths, with lower values indicating a more accurate model. Both models show similar RMSE, indicating comparable predictive accuracy.

From the table above, we have the following calculated regression models using both Poisson and Negative Binomial:

Poisson Regression Model:

$$\log(\lambda_i) = 6.973 + 0.524 \cdot X_{\text{Ischemic Heart Disease}} - 0.408 \cdot X_{\text{Cardiovascular Disease}} - 1.028 \cdot X_{\text{Heart Failure}} - 0.548 \cdot X_{\text{Diabetes}} \tag{6}$$

Negative Binomial Regression Model:

$$\log(\mu_i) = 6.976 + 0.528 \cdot X_{\text{Ischemic Heart Disease}} - 0.414 \cdot X_{\text{Cardiovascular Disease}} - 1.023 \cdot X_{\text{Heart Failure}} - 0.547 \cdot X_{\text{Diabetes}}$$

(7)

5 Discussion

5.1 Interpretation of the regression models

Our regression analysis revealed several intriguing patterns about the relationship between various health conditions and mortality rates. In interpreting the models, it is crucial to consider the baseline level of deaths, represented by the intercept, which is the expected number of deaths when the incidence of specific conditions is zero. For the Poisson and Negative Binomial models, the intercepts were approximately 6.973 and 6.976, respectively. When exponentiated, these values suggest a high baseline number of deaths, which serves as a point of comparison for the other coefficients in the model. It is important to clarify that a positive coefficient for a health condition does not imply that the presence of the condition increases one's likelihood of survival. Instead, it reflects the relative risk of mortality associated with that condition compared to the baseline established by the intercept. For example, the positive coefficient for "All Other Forms of Ischemic Heart Disease" does not indicate that the disease leads to improved survival but rather that it is associated with a higher mortality rate relative to the baseline level of deaths captured by the intercept.

The negative coefficients for conditions like "Atherosclerotic Cardiovascular Disease" and "Diabetes Mellitus" in both models indicate that, relative to the intercept, these conditions are associated with a lower mortality rate. This seemingly paradoxical result could be due to effective management and treatment strategies that improve survival rates for these conditions. Moreover, it might reflect the relative impact of these conditions compared to more lethal alternatives that define the intercept's baseline mortality. "Congestive Heart Failure" showed a negative association with mortality in the model. This finding suggests that, while heart failure is a serious condition, its presence as a recorded cause of death has a lower relative risk compared to the baseline mortality rate, which could include deaths from more acute or untreated conditions.

The differences in coefficients between the two models can provide insights into model fit and appropriateness. The Negative Binomial model generally yielded a better fit to our data, indicated by less negative log-likelihood, higher expected log predictive density (ELPD), and lower information criteria (LOOIC and WAIC) scores. This suggests that it more accurately accounts for variability in mortality data, potentially due the variance in the data being higher than the mean, which is common in count data like mortality statistics.

11

5.2 Model Comparison

The log-likelihood values provide an indication of how well the model fits the data. For the Poisson model, the log-likelihood is -274.969, while for the Negative Binomial model, it is considerably higher (less negative) at -196.422. This substantial difference suggests that the Negative Binomial model provides a much better fit to the data compared to the Poisson model. A higher log-likelihood (closer to zero) implies that the model's estimated probabilities of the observed outcomes are higher, meaning that the model is more likely to produce the observed data. Furthermore, the LOOIC values are another essential tool for model comparison. They offer a balance between model fit and complexity, penalizing for overfitting. For the Poisson model, the LOOIC is 609.7, with a standard error of 121.7. In contrast, the Negative Binomial model has a LOOIC of 396.5, with a much smaller standard error of 5.8. The lower LOOIC value for the Negative Binomial model, combined with its lower standard error, indicates a model that is more efficient in explaining the data without overfitting.

The graphs presented in Figure 2 illustrate the posterior predictive checks for the Poisson and Negative Binomial models applied to the mortality data concerning various causes of death. For both graphs, the x-axis represents the count of total deaths, and there are two sets of lines: one for the observed counts (denoted by y) and another for the predicted counts from the respective models (denoted by y_rep). The lines for predicted counts represent different simulated datasets generated by the model, reflecting its probabilistic predictions.

In the Poisson Model graph, the overlapping lines suggest that while the model captures the general trend of the observed data, it may not encapsulate the full variability, as indicated by the divergence between the observed and predicted counts particularly in regions of higher death counts. The model seems to generally underpredict the total death count per year when the data peaks which could be caused by the difference in variance and mean.

Conversely, in the Negative Binomial Model graph, the predicted counts lines appear to more closely follow the pattern of the observed data, including the peaks and valleys, which indicates a better fit. This model's ability to account for overdispersion with its μ_i term likely contributes to its more accurate representation of the variability in the data. The thicker lines indicating the observed counts y remain consistent across both models, serving as a benchmark for evaluating the predictive accuracy. The spread and concentration of the lighter lines (the predictive simulations y_rep) around these observed counts provide visual insight into the model's performance.

The Negative Binomial model appears to yield a better fit to the data, particularly for larger counts, which suggests it might be the more appropriate model for analyzing the impact of diabetes and heart disease on mortality rates within the studied population. This would be consistent with the expectation that mortality data, influenced by complex and varied factors, would exhibit overdispersion—an assumption better accommodated by the Negative Binomial model than by the Poisson model.
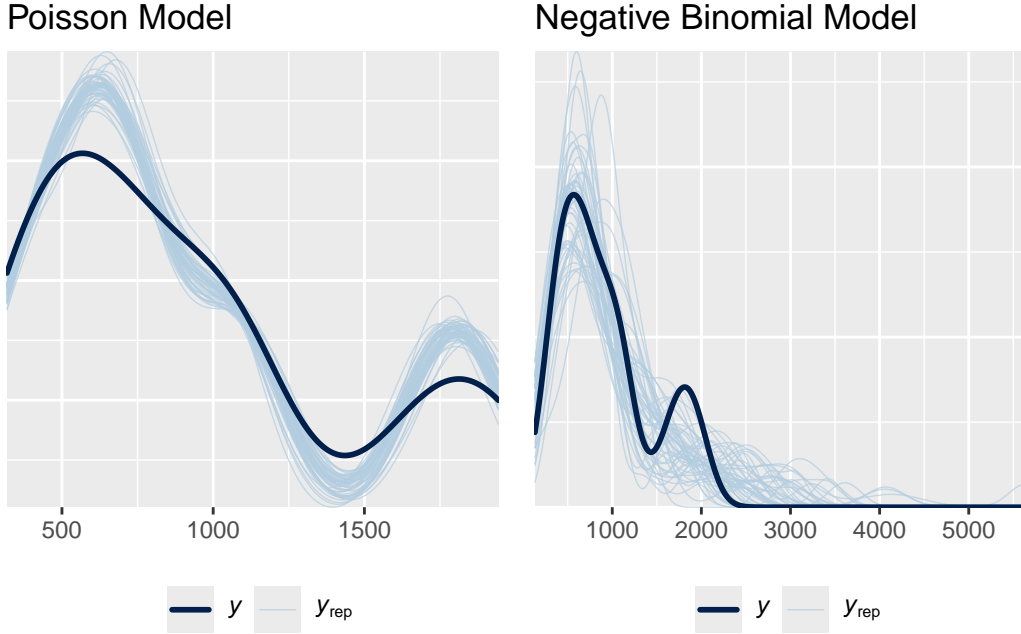
Figure 2: Comparison of Poisson and Negative Binomial Models

5.3 Impact of COVID-19

The onset of the COVID-19 pandemic in late 2019 brought unprecedented changes to global health trends. Its direct impact on mortality rates would likely be reflected in the data for the years 2020 and 2021. If COVID-19 is included as a variable in the regression model, we might observe a significant coefficient associated with it, reflecting its substantial effect on mortality rates. Moreover, the pandemic could indirectly affect the coefficients of other variables. Healthcare systems were overwhelmed, which could have led to an increase in mortality from other causes due to resource reallocation. The model's intercept in the years of the pandemic might capture a higher baseline mortality rate due to the surge in deaths associated with COVID-19. Although illnesses directly related to COVID-19 were not included in the regression models, it is entirely possible that patients that contracted COVID-19 had weaker immune systems, and as a result increased the mortality even though their cause of death was not COVID-19. From Figure 1, we see that there is a significant uptick in Diabetes and Ischemic Heart Disease following 2019 when the pandemic began. This may be the result of either a lack of available hospital care due to medical services being overloaded with the pandemic, an increase in fast food consumption due to quarantine, or other causes brought about by COVID.

5.4 Weaknesses and next steps

A notable weakness of the study lies in its indirect measure of fast food consumption. The lack of direct dietary data necessitates reliance on broader consumption trends as a proxy, potentially obscuring the nuanced relationship between individual dietary habits and health outcomes. This gap highlights the need for comprehensive datasets that encompass dietary patterns, enabling a more granular analysis of the impact of specific dietary choices on health.

Furthermore, the regression analysis, though insightful, is constrained by the scope of variables included. The absence of factors such as socioeconomic status, physical activity levels, and access to healthcare could result in an incomplete picture of the determinants of mortality rates. The coefficients indicating a decrease in mortality associated with certain health conditions may reflect advancements in treatment and management rather than a true decrease in risk, underscoring the importance of contextualizing statistical findings within the broader landscape of public health research.

Another potential limitation is the assumption of linearity in the relationship between the predictor variables and the log count of deaths. Real-world data, especially in the context of health outcomes, often exhibit complex, non-linear interactions that could be better captured through alternative modeling approaches. Exploring models that accommodate such non-linearities could provide deeper insights into the dynamics at play.

In terms of next steps, this research paves the way for several avenues of inquiry. Foremost among these is the integration of detailed dietary data into the analysis. Collaborations with public health agencies to access or collect granular consumption data could significantly enhance the study's ability to draw direct correlations between diet and mortality outcomes. Additionally, expanding the model to include a wider array of variables, such as lifestyle factors and healthcare access, would allow for a more comprehensive understanding of the influences on mortality rates.

Adopting a longitudinal approach to examine changes in individual health outcomes over time, potentially through cohort studies, could also address some of the limitations inherent in cross-sectional data analysis. Such studies would provide a richer dataset for examining the impact of fast food consumption on health, accounting for changes in dietary habits, healthcare practices, and disease prevalence over time.

Ultimately, this research contributes to an essential dialogue on the intersection of diet, public health, and mortality. By spotlighting the need for more nuanced data and sophisticated analytical techniques, it calls for a concerted effort among researchers, policymakers, and public health practitioners to address the complex challenges at the heart of diet-related health outcomes.

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.

Auguie, Baptiste. 2017. *gridExtra: Miscellaneous Functions for "Grid" Graphics.* https://CRAN.R-project.org/package=gridExtra.

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. 2022a. "Rstanarm: Bayesian Applied Regression Modeling via Stan." https://mc-stan.org/rstanarm/.

———. 2022b. "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.

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

Tsao, Connie W, Aaron W Aday, Zaid I Almarzooq, Cheryl A M Anderson, Pankaj Arora, Christy L Avery, Carissa M Baker-Smith, et al. 2023. "Heart Disease and Stroke Statistics-2023 Update: A Report from the American Heart Association." *Circulation* 147 (8): e93–621.

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.

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.