Respiratory-Related Mortality Rates Show A Positive Correlation With Increasing Air Pollution*

Based on Data Collected From The Province Of Alberta

Vanshika Vanshika Shivank Goel Navya Hooda

March 16, 2024

Air pollution's severe health impact on human respiratory and cardiac systems is a growing public health issue. The fine particulate matter (PM2.5) arise a primary concern due to its deep lung penetration and systemic effects. This research uses data from Alberta, Canada, to study the connection between PM2.5 levels as measured by the Air Quality Health Index (AQHI), and the number of deaths from certain heart and lung diseases. We utilized statistical models, and identified a strong association between increased PM2.5 levels and higher mortality from these diseases. These findings suggest that there is an immediate need for health policies that address air quality improvements to safeguard public health for Alberta and beyond.

1 Introduction

The impact of air pollution on human health has increasingly become a global concern, with respiratory illnesses being a significant consequence of poor air quality. The National Library of Medicine reports that approximately 4 million people die prematurely each year from chronic respiratory diseases linked to air pollution Med (2020). The World Health Organization also suggests that air pollution poses a risk for all-cause mortality and specific diseases, especially the exposure to air pollution is strongly linked with outcomes such as stroke, ischemic heart disease, chronic obstructive pulmonary disease, lung cancer, pneumonia, and cataracts Organization (2024).

The pollutants most significantly concerning for public health, are particulate matter (PM), carbon monoxide (CO), ozone (O3), nitrogen dioxide (NO2), and sulfur dioxide (SO2). Of

*Code and data are available at: https://github.com/shivankgoel003/Mortality-in-Alberta.

these, fine particulate matter is particularly worrisome due to its ability to penetrate the lungs deeply, enter the bloodstream, and reach organs, leading to systemic damage to tissues and cells.

Specifically, PM2.5 is of greater concern due to its harmful properties and dangers. PM2.5 is airborne particulate matter (PM2.5) and is not a single pollutant but rather a mixture of many chemical species and aerosols. PM2.5 is associated with the greatest proportion of adverse health effects related to air pollution, both in the United States and worldwide. Long-term (months to years) exposure to PM2.5 has been linked to premature death, particularly in people who have chronic heart or lung diseases, and reduced lung function growth in children Board (2024).

In this paper, we analyze data from Alberta, Canada, and aim to explore the estimand of the number of deaths that can be correlated to the Air Quality Health Index (AQHI) and PM2.5 quantities. Specifically, we assess how AQHI relates to the prevalence of respiratory and cardiac illnesses in this specific region, focusing on four major types: Chronic Obstructive Pulmonary Disease, Ischemic Heart Disease, Acute Myocardial Infarction, and Lung Cancer. We use the mortality rate in Alberta data to select respiratory-related illnesses and air quality health index data for Alberta through their provincial open data portal. To analyze respiratory illnesses, we use air pollution measured through the AQHI and specifically the particulate matter 2.5 (PM2.5) air pollutant as a predictor of mortality related to respiratory illnesses. In total, we will be assessing the overall relationship between AQHI and respiratory illnesses, the relationship between PM2.5 and respiratory illnesses, as well as the relevance of PM2.5 levels in lung and heart disease values as a means to draw any notable significance of PM2.5 in different types of illnesses.

Using negative binomial regression, this study seeks to uncover trends and correlations between AQHI and the prevalence of respiratory illnesses. Negative binomial regression is a type of generalized linear model (GLM) that is specifically designed to model count data. In our paper, we are using it to predict the number of deaths, based on the year and illness, for the pollutants present in the air. By analyzing this relationship, we provide valuable insights that can inform policymakers, healthcare professionals, and the public about the impact of air pollution on respiratory health in Alberta. This research aims to contribute to a better understanding of the health effects of air pollution and to support the development of targeted strategies for air quality improvement and public health protection in Alberta. Our analytical framework explores whether the variables assessed seem to correlate, if at all, to the total deaths in respiratory-related illnesses through 2012-2022.

This paper is organized as follows: In the Data section, we outline the sources of three different datasets, detail the data-cleaning processes applied to each dataset, and describe any data merging procedures used to prepare the data for input into various models. The Results section focuses on analyzing trends and correlations between AQHI, PM2.5, and respiratory and cardiac illnesses. Additionally, we discuss the trends and patterns identified by our model, along with the correlation analysis between these variables. In the Discussion section, we

present our overall findings, discuss any biases and weaknesses in the data that may have influenced these findings, and explain our approach to analyzing these limitations.

Overall, this research has the potential to inform public health strategies and interventions aimed at reducing respiratory illnesses and improving overall health in Alberta and beyond.

2 Data

2.1 Data Source and Collection:

The study relies on datasets obtained from the provincial open databases of Alberta, accessible through the official website government (2024). Three essential datasets were employed to extract relevant variables for analysis, with the goal of revealing the association between air quality and mortality rates in Alberta. The analysis begins with the leading causes of death dataset for Alberta, sourced from the provincial open data portal government (2023b). This dataset provides insights into mortality rates associated with various illnesses, facilitating the examination of trends related to respiratory and heart-related illnesses. To explore potential correlations between air quality and mortality rates, the study incorporates the Air Quality Health Index (AQHI) dataset for Alberta, sourced from the provincial open data portal government (2023a). This dataset offers entensive information on the AQHI across different municipalities in Alberta over multiple years. Additionally, the study utilizes PM2.5 air pollutant concentration level data sourced from Alberta's official resources Alberta Government (2023). This dataset provides detailed information on the concentration levels of PM2.5 pollutants over several years, offering valuable insights into air quality trends. The following subsections outline the sources, collection methodologies, and data-cleaning procedures implemented to ensure the accuracy and reliability of the datasets used in the analysis. This meticulous approach ensures that the data is prepared for thorough analysis, facilitating the exploration of correlations between air quality indicators and mortality rates in Alberta.

Leading Causes of Death in Alberta Data: The disease data is found from the government of Alberta's open data portal government (2024), and was last updated on September 22, 2023 and continues to be updated annually. This dataset encompasses mortality data related to the top 30 common causes of death. It reports on types of diseases, causes of death, mortality denoted by total death counts, and ranking for 2000-2022. Due to our focus on respiratory illnesses, in the leading cause of death dataset, we grouped diseases by categories. Our category of focus included filtering on illnesses like acute myocardial infarction, malignant neoplasms of the trachea, bronchus, and lung, other chronic obstructive pulmonary disease, and all other forms of chronic ischemic heart disease as seen in Figure 1. Leading causes of death are measured and ranked by the top 30 most common death causes each specific year. The causes of death are classified based on the International Classification of Diseases 10th Edition.

AQHI Data: The second dataset we used is the air quality health index (AQHI) dataset found at the government of Alberta's open data portal government (2024). This dataset contains

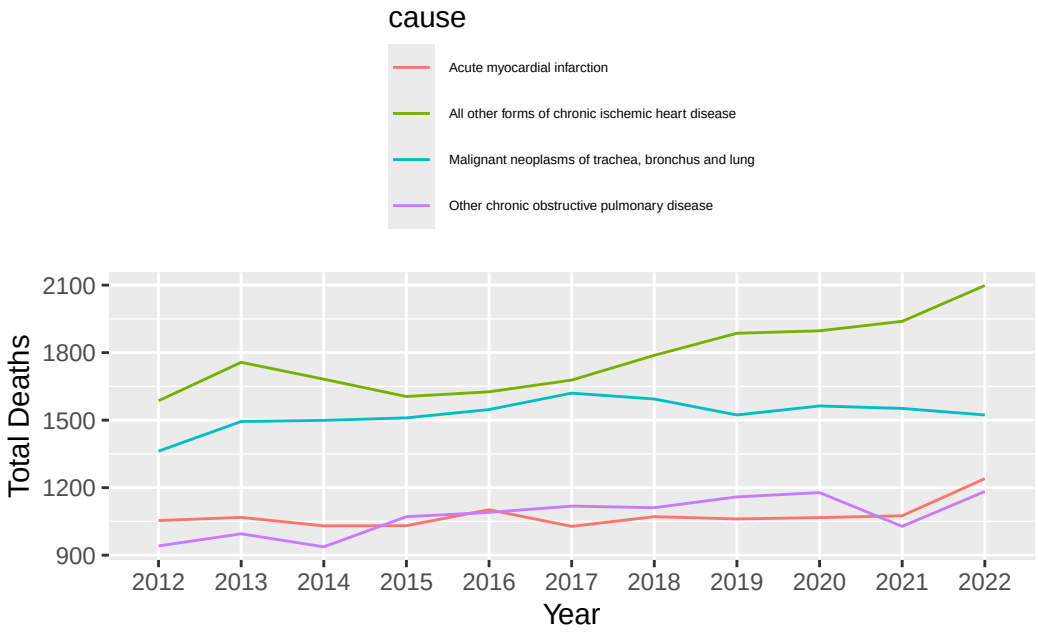


Figure 1: Total Deaths by Year for Each Cause

AQHI by municipality for the years 2012-2022 and reports air quality health index, and health risk both quantitatively and qualitatively. To use the AQHI dataset we employed simple data-cleaning practice to maintain descriptive variable names and readability. The data is measured by the percentage of hours for each year at a given air quality level, by municipality. The Air Quality Health Index is calculated based on the relative risks of a combination of common air pollutants that is known to harm human health. These pollutants are ozone (O3) at ground level, particulate matter (PM2.5), and nitrogen dioxide (NO2). Risks are defined as follows: 1-3 High Quality; 4-6 Moderate Quality; 7-9 Low Quality; 10+ Very Low Quality.

PM2.5 Data: We used the PM2.5 data set retrieved from government (2024) which was last updated in April 2023. It reports on average PM2.5 concentration levels through the years 2000-2021 using a provincial average, the 10th percentile quantities, and the 90th percentile quantities, with a focus on 8 municipalities Edmonton, Fort McMurray, Grande Prairie, Lethbridge, Medicine Hat, and Red Deer respectively and lastly reports the Canadian Ambient Air Quality Standard (CAAQS) value. The trends of the PM2.5 concentration levels over the years can be seen in Figure 2.

The Alberta Air Zone report Environment and Areas (2021), which is linked to our dataset, provides a detailed explanation of the measurement and processing of the PM2.5 quantity. Alberta Air Zones divides Alberta into six air zones which are aligned with Alberta's Land-use Framework regional boundaries. Ambient air quality in Alberta is monitored at continuous air monitoring stations located within these air zones. PM2.5 quantities are taken throughout

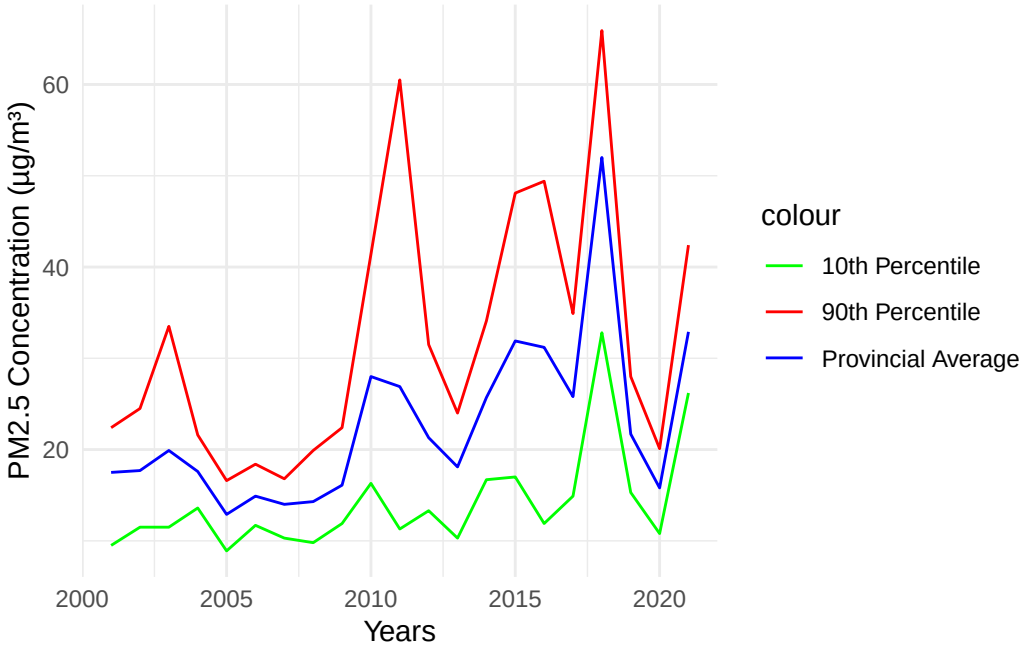


Figure 2: Annual Trends of PM2.5 Concentrations in Alberta

these stations across Alberta, and they measure the quantities in µg (micrograms per cubic meter of air).

2.2 Data Cleaning

We used R (R Core Team 2023) for data cleaning and processing, utilizing packages like tidyverse (Wickham et al. 2019) for data manipulation and janitor (Firke 2023) for cleaning column names. Other packages used includes `ggplot2` (Wickham 2016), `dplyr` (Wickham et al. 2023), `readr` (Wickham, Hester, and Bryan 2024), `tibble` (Müller and Wickham 2023), `janitor` (Firke 2023),`reshape2` (Wickham 2007), `knitr` (Xie 2023), `ggbeeswarm` (Clarke, Sherrill-Mix, and Dawson 2023), `ggrepel` (Slowikowski 2024), `kableExtra`(Zhu 2024), `readxl`(Wickham and Bryan 2023), `MASS`(Venables and Ripley 2002), `rstanarm`(Goodrich et al. 2022), `modelsummary`(Arel-Bundock 2022) and `here` (Müller 2020).

The raw air quality data were preprocessed to remove inconsistencies and irrelevant information as shown in Table 1. Specifically, we filtered the dataset to include observations from the years 2012 to 2021, which are relevant to our analysis. Additionally, we merged this dataset with additional information on peak pollution levels for analysis.

Table 1: Sample of Cleaned State Score Data

municipality	year	air_quality_health_index	health_risk	original_value
Red Deer	2012	1	High Quality	0.08876
Fort Saskatchewan	2012	1	High Quality	0.14230
Edmonton	2012	1	High Quality	0.03852
Lethbridge	2012	1	High Quality	0.03033
Caroline	2012	1	High Quality	0.14105
Medicine Hat	2012	1	High Quality	0.01899

Similar to the previous datasets, the raw mortality data(deaths-leading-causes data) underwent cleaning procedures to focus on specific causes relevant to our analysis. We filtered the dataset to include observations up to 2021 and merged it with additional information on air quality for correlation analysis as showed in Table 2.

Table 2: Sample of Cleaned State Score Data

year	cause	ranking	total_deaths
2012	All other forms of chronic ischemic heart disease	1	1586
2012	Malignant neoplasms of trachea, bronchus and lung	2	1362
2012	Acute myocardial infarction	4	1054
2012	Other chronic obstructive pulmonary disease	5	941
2013	All other forms of chronic ischemic heart disease	1	1757
2013	Malignant neoplasms of trachea, bronchus and lung	2	1494

The PM2.5 Concentration data was cleaned and merged with other datasets. The Provincial Average of PM2.5 concentration per year is added it to the dataset of mortality to get better results.

2.3 Data Modifications

In this study, we constructed unique datasets by thoughtfully selecting and merging data from the Government of Alberta's open data portal, Alberta.ca, spanning the years 2012 to 2022. Our process involved merging variables from various datasets to create specific datasets tailored for model building and analysis. One such dataset, 'cleaned_chart_data,' was created by merging variables such as causes of death, total deaths, provincial average PM2.5 levels, and CAAQS. This dataset was designed to facilitate our analysis of any significant correlations between these variables.

Additionally, we derived two other datasets, 'merged_data' and 'merged_heart_data,' by merging variables related to heart disease numbers, lung disease numbers, and provincial average PM2.5 values from year 2001 to 2021. These datasets were instrumental in examining the impact of PM2.5 on each type of illness, as previously discussed.A snapshot of the 'merged_data" for the lung disease is referenced in Table 3.

Overall, our methodology ensured the creation of merged datasets that allowed for a detailed investigation into the relationships between PM2.5 levels and various health outcomes in Alberta.

Table 3: Sample of Cleaned State Score Data

year	total_deaths	provincial_average
2001	1759	17.5
2002	1904	17.7
2003	1927	19.9
2004	1961	17.6
2005	2071	12.9
2006	2080	14.9

3 Model

3.1 Model Set-up

In our analysis, we implemented a series of regression models to study the impact of air pollution on various health outcomes, especially focusing on respiratory and cardiac illnesses in Alberta, Canada. The models used include three variations of negative binomial regression and one Poisson regression model.

We used **Negative Binomial Regression (NBR)** to study how air pollution affects deaths related to breathing problems. NBR is extremely useful for handling overdispersed count data, where the variance exceeds the mean UCLA (2021). This method is good for analyzing data where the numbers (like the total number of deaths) vary more than usual. In many real-world scenarios, these numbers can be more unpredictable than what simpler methods like the Poisson distribution can handle. Negative Binomial Regression works better in these kinds of situations since it accounts for this extra unpredictability. It gives us easy-to-understand results that show how average air pollution in a province **(provincial_average)** can affect the number of deaths **(total_deaths)**.

Our first model compares different causes of mortality and their corresponding death count. The second model investigates the relationship between heart-related causes of death and PM2.5 levels. And lastly, the third model analyzes the connection between lung-related causes of death and air quality. To build our model for heart and lung-related causes of death, we identified four key diseases from the dataset and grouped them into two broader categories: heart diseases and lung diseases.

The diseases and their grouping are as follows:

Heart Diseases:

Ischemic Heart Disease (All other forms of chronic ischemic heart disease)

Heart Attack (Acute Myocardial Infraction)

Lung Diseases:

Trachea/Bronchus/Lung Cancer (Malignant neoplasms of trachea, bronchus, and lung)

COPD (Other chronic obstructive pulmonary disease)

In our approach, the total deaths from Ischemic Heart Disease and Heart Attack were summed up to create the dataset for heart diseases, and similarly, data from Trachea/Bronchus/Lung Cancer and COPD were combined for lung diseases.

Details of Regression Models

The general form of our negative binomial model is represented as follows.

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

1. **Modelling the count data**:

$y_i|\mu_i, \phi \sim NegBin(\mu_i, \phi)$

Here, y_i represents the count of deaths for each health outcome (like heart or lung disease). The model assumes that the count data follow a Negative Binomial distribution.

2. **Link function and predictors**

$\mu_i = exp(\alpha + \beta x_i)$

Link Function – This is the link function used for the negative binomial regression. By default, when we specify dist = negbin, the log link function is assumed (and does not need to be specified) UCLA (2021). In this equation μ_i is the expected count of deaths, which we model as an exponential function of the predictors. This ensures that our predictions for the count data are always positive. The predictor (x_i) in our models include average air pollution levels (provincial_average).

3. **Coefficient**

Coefficients are the estimated negative binomial regression coefficients for the model. We can interpret the negative binomial regression coefficient as follows: for a one unit change in the predictor variable, the log of expected counts of the response variable changes by the respective regression coefficient, given the other predictor variables in the model are held constant UCLA (2021).

$\alpha \sim$ Normal(0,2.5): The intercept α has a normal prior, indicating about our initial assumption about the baseline level of health outcomes in the absence of predictors.

$\beta \sim$ Normal(0,2.5): The coefficients β for our predictors also follow a normal distribution. These coefficients tells us about the relationship strength between air pollution and the health outcomes.

4. **Dispersion parameter**

As the dispersion parameter gets larger and larger, the variance converges to the same value as the mean, and the negative binomial converges to a Poisson distribution Clay Ford (2023) ϕ is modeled using an exponential distribution. It accounts for extra variation in the count data that is not explained by the Poisson model alone.

3.2 Model Summary and Model Results

Model 1: Comparative Analysis of Mortality Causes

Interpretation of table results from Table 4

Table 4 presents the results from our first model, and compares various causes of mortality using both Poisson and negative binomial regression approaches. This table helps in understanding the relationship between air pollution and mortality due to different causes.

1. Coefficients - As described above, the coefficients from these models (the β values) tell us how changes in the predictors (like provincial average PM2.5 levels) are associated with changes in the response variable (total death counts). A positive coefficient indicates that an increase in the predictor leads to an increase in the response, while a negative coefficient suggests the opposite.

Ischemic Heart Disease: Both Poisson and negative binomial models showed a coefficient of 0.510, suggesting a consistent and significant association between air pollution and mortality due to Ischemic Heart Disease. This positive coefficient highlights a serious public health concern, suggesting that as air pollution worsens, the risk of death due to Ischemic Heart Disease increases.

Trachea/Bronchus/Lung Cancer: Similarly, for Trachea/Bronchus/Lung Cancer, both models indicated a coefficient of 0.353, describing the impact of air pollution on these lung diseases. It is learned through this coefficient that increased levels of air pollution are associated with higher mortality rates from these types of lung cancer.

COPD: For COPD, the coefficient was minimal (0.004), which suggests that, there is a negligible direct association between air pollution and COPD mortality. It's important to note that this does not necessarily mean air pollution has no effect on COPD but rather, it might indicate that the relationship is more complex and possibly influenced by other variables not captured in the model.

2. Number of Observations (Num.Obs.):

"Num.Obs. 9048" indicates the total number of data points (observations) used in the model. A higher number of observations often produces better predictions and thus more accuaracy in results.

3. Log-Likelihood (Log.Lik.): The log-likelihood values (-69,064.136 for Poisson and -53,402.269 for negative binomial) measure how well the model fits the data. A higher (less negative) log-likelihood value generally indicates a better model fit IBM (2021). In this case, the higher log-likelihood for the negative binomial model suggests it fits the data better than the Poisson model, which is consistent with the expectation that the negative binomial model handles overdispersion more effectively. This result is also evident using posterior predictive checks Figure 7, to show that the negative binomial approach is a better choice for this circumstance.

4. Expected Log Predictive Density (ELPD): ELPD (-69,078.6 for Poisson and -53,405.5 for negative binomial) suggests that negative binomial regression may be better at predicting new data compared to the Poisson model, which truly matches our expectation Alexander (2024).

5. Leave-One-Out Cross-Validation Information Criterion (LOOIC): The negative binomial model's lower LOOIC suggests it is a better model in terms of prediction and handling the data complexity.

6. Root Mean Square Error (RMSE): RMSE (97.30 for both models) measures the model's prediction error. Root mean square indicates the quality of model fit. Lesser RMSE indicates better model fit and higher RMSE indicates poor model fit Sachid Deshmukh (2019). The fact that RMSE is the same for both models suggests similar predictive accuracy in terms of the magnitude of errors.

Model 2: Heart Disease and Air Quality

Model 2 discusses and provides results about the specific relationship between heart-related diseases and PM2.5 levels.

Interpretation of table results from Table 5:

Table 5 showcases the relationship between heart-related diseases and PM2.5 levels, utilizing our second regression model.

1. Intercept (8.02): The intercept value of 8.02 represents the log count of heart disease-related deaths when the PM2.5 level is at its baseline (zero). This high value suggests other influential factors for heart disease mortality beyond air pollution levels.

2. Provincial Average PM2.5 (Coefficient = 0.00): The coefficient for provincial average PM2.5 being 0.00 indicates no significant impact of PM2.5 levels on heart disease mortality within the analyzed range. This finding shows that heart disease is caused by many factors, so public health plans should use various methods to address it.

Table 4: Different causes of mortality and their death counts

	Poisson	Negative binomial
Ischemic Heart Disease	0.510	0.510
		(0.002)
Trachea/Bronchus/Lung Cancer	0.353	0.353
		(0.002)
COPD	0.004	0.004
		(0.002)
Num.Obs.	9048	9048
Log.Lik.	−69 064.136	−53 402.269
ELPD	−69 078.6	−53 405.5
ELPD s.e.	468.1	70.9
LOOIC	138 157.2	106 811.0
LOOIC s.e.	936.3	141.9
WAIC	138 157.2	106 811.0
RMSE	97.30	97.30

3. Number of Observations: The analysis is based on a smaller dataset of 21 observations to study particular relationship over the years between heart diseases and PM2.5.
4. Model Fit and Predictive Accuracy: Key statistical measures (Log-Likelihood, ELPD, LOOIC, RMSE), similar to model 1, would be helpful in edetermining the model's fit and predictive power. A higher Log-Likelihood, lower LOOIC, and comparable RMSE with other models would further validate the findings.

Model 3: Lung Disease and Air Quality

Interpretation of table results from Table 6

Table 6 focuses on the connection between lung-related diseases and air quality, as represented by PM2.5 levels.

1. Intercept (7.57): The intercept of 7.57, indicates the baseline level of lung disease mortality in the absence of PM2.5 contributions. This again suggest the influence of other variables in lung disease outcomes.
2. Provincial Average PM2.5 (Coefficient = 0.01): A small but positive coefficient for PM2.5 suggests a marginal increase in lung disease mortality with rising PM2.5 levels.
3. Number of Observations: As with the previous model 2, the number of observations were 21 to study particular relationship over the years between lung diseases and PM2.5.

Table 5: Heart related causes death count and the air quality

	heart causes model
(Intercept)	8.02
	(0.18)
provincial_average	0.00
	(0.01)
Num.Obs.	21
Log.Lik.	−163.733
ELPD	−164.5
ELPD s.e.	0.5
LOOIC	329.0
LOOIC s.e.	1.0
WAIC	328.9
RMSE	142.90

4. Model Fit and Predictive Accuracy: Evaluating this model using Log-Likelihood, ELPD, LOOIC, and RMSE is crucial. For instance, a higher Log-Likelihood and lower LOOIC compared to alternative models would indicate a better fit for the lung disease data.

4 Results

4.1 Correlation between AQHI and Causes/Death

Exploring the relationship between AQHI and total mortality measured by the correlation graph Figure 3 we notice there is a weak relationship between our chosen group of illnesses and AQHI.

The points representing mortality rates due to acute myocardial infarction seem relatively stable across the range of AQHI values. The fitted line is almost flat, suggesting that within the AQHI range provided, there is no strong relationship between AQHI and the mortality rates from this condition.

Similar to acute myocardial infarction, the mortality rates for 'All Other Forms of Chronic Ischemic Heart Disease' appear stable across the AQHI values. The regression line is again flat, indicating no significant trend that suggests an increase or decrease in mortality rates associated with changes in AQHI within the observed range.

For 'Malignant Neoplasms of Trachea, Bronchus, and Lung' the scatter plot shows a slight decrease in mortality rates with higher AQHI, as indicated by a gentle negative slope of the regression line. However, given the tight clustering of points, this trend might not be

Table 6: Lung related causes death count and the air quality

	Lung Causes model
(Intercept)	7.57
	(0.20)
provincial_average	0.01
	(0.01)
Num.Obs.	21
Log.Lik.	−160.858
ELPD	−161.6
ELPD s.e.	0.8
LOOIC	323.3
LOOIC s.e.	1.6
WAIC	323.2
RMSE	248.30

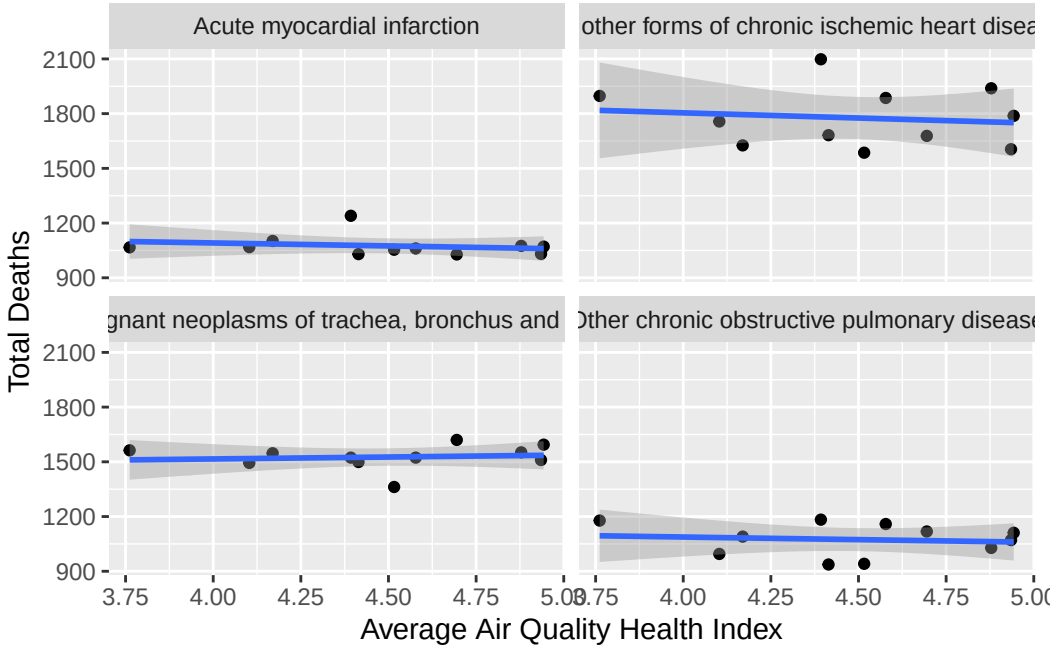


Figure 3: Mortality Rates vs. Air Quality Health Index

statistically significant. The data does not indicate a strong correlation between AQHI and mortality from these neoplasms within the given range.

The trend for COPD appears similar to that of malignant neoplasms with a slight negative slope. However, the range of mortality rates is relatively narrow, and the points are closely clustered, suggesting a weak relationship, if any, between AQHI and mortality from COPD.

While we observe a weak correlation between the AQHI and each type of illness, it is important to consider the trends in AQHI over the period from 2012 to 2022, as depicted in Figure 4. We notice a noticeable dip in AQHI around the COVID-19 years, but overall, the AQHI remains consistently below 5 on its 1-10 severity scale. This low peak in AQHI may not be severe enough to cause any correlation with the total mortalities of the respiratory and cardiac illnesses we are analyzing. Some of the weak relationships we assessed above can likely be attributed to this factor.

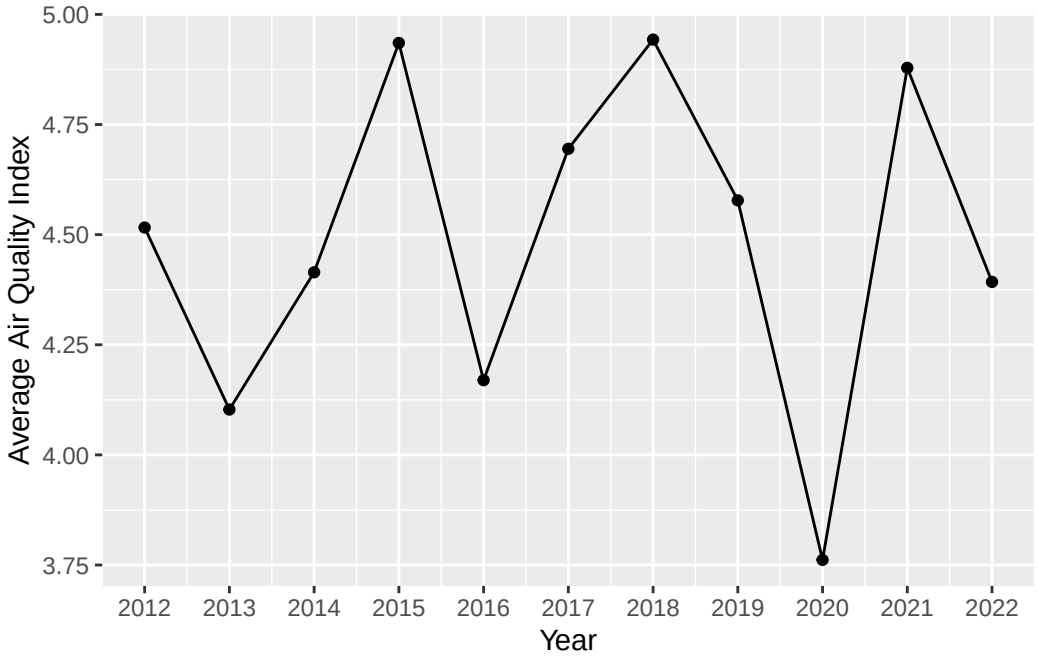


Figure 4: Average Air Quality Health Index by Year

4.2 Lung and Heart Related Mortality vs. PM2.5 Levels

To assess how PM2.5 affected lung disease and heart disease we interpret the the trends by creating 2 different correlation plots. In the context of lung diseases, PM2.5 is expected to be a significant factor, whereas it should not have a significant impact on heart diseases. In Figure 5, the lung-related causes graph, we observed a strong linear correlation between the

provincial average of PM2.5 and the total deaths from lung diseases. As the provincial average of PM2.5 levels increased, so did the total death count from lung diseases.

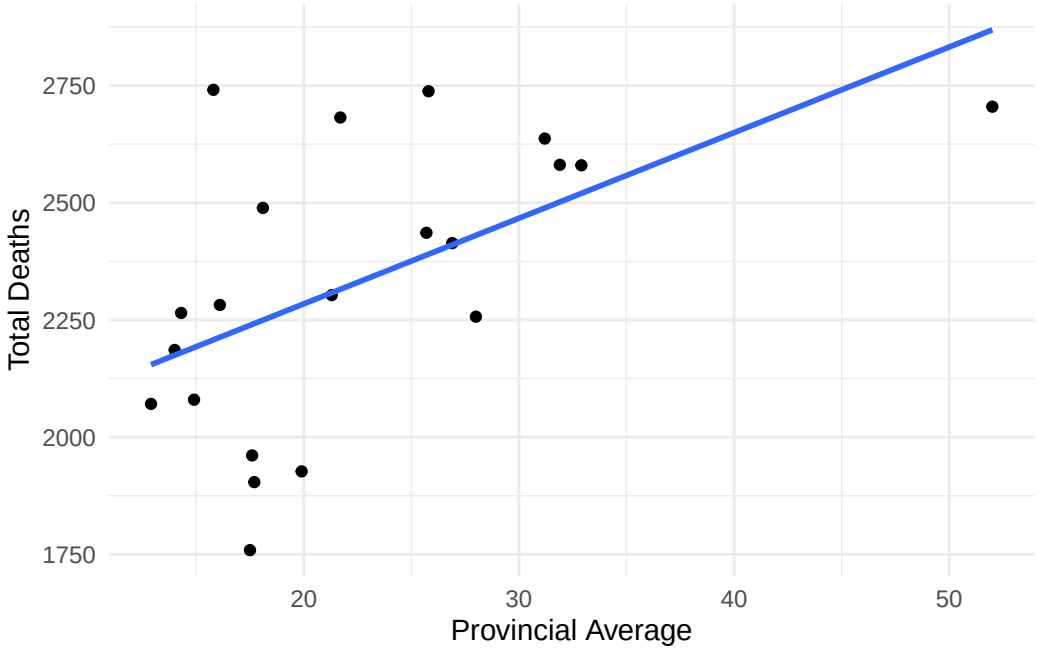


Figure 5: Lung Related Causes Mortality Rates vs Average PM2.5 Quantity

In contrast, the heart disease-related deaths showed only a slight correlation when compared to the same levels of PM2.5, as seen in Figure 6 . The values of PM2.5 did not appear to significantly affect heart-related diseases, as the increase in PM2.5 levels resulted in a weak positive linear relationship with the total heart disease-related deaths.

5 Discussion

5.1 Findings

Combining the insights from our models and the graphical data from the graphs for lung and heart-related mortality rates versus average PM2.5 levels, we can conclude the following overall findings from our analysis: AQHI did not have a significant correlation with respiratory and cardiac illness mortality. PM2.5 had a significant association with lung-related illness mortality. PM2.5 had a weak correlation to heart-related diseases and, overall did not correlate to its mortality numbers.

More specifically we discuss the model findings, possible reasons for our results, and discrepancies in our correlation analysis and model results below.

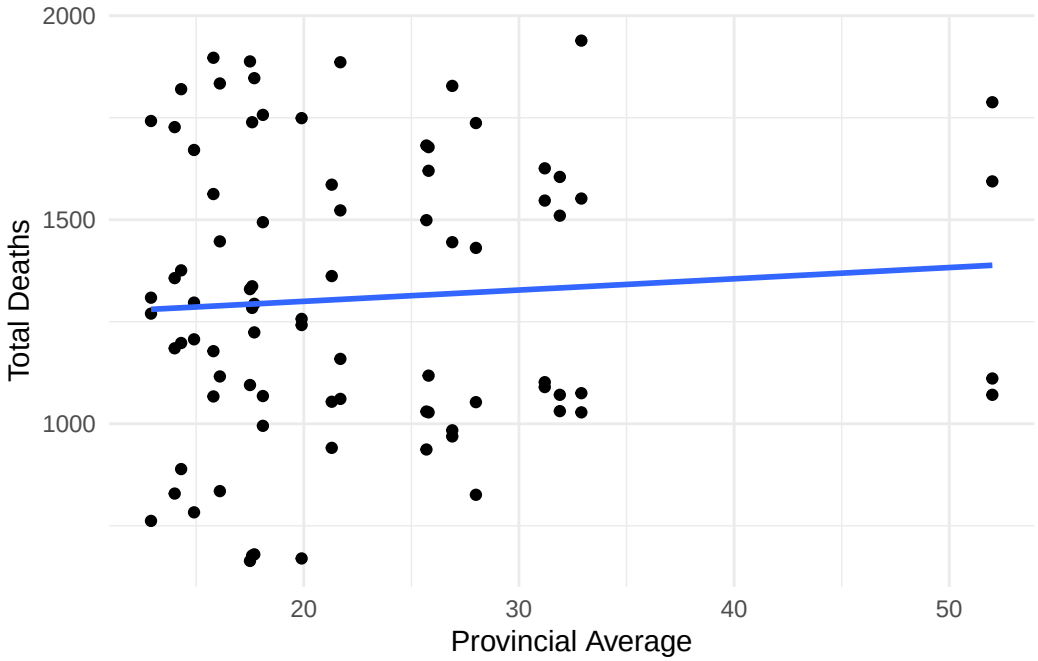


Figure 6: Heart Related Causes Mortality Rates vs Average PM2.5 Quantity

The PM2.5 regression models demonstrate that there is a statistically significant positive association between PM2.5 levels and mortality rates for Ischemic Heart Disease and Trachea/Bronchus/Lung Cancer, as indicated by positive coefficients. For COPD and other heart-related causes, the models suggested no direct association with air pollution within the data scope that was analyzed.

The correlation graph for lung-related causes and PM2.5 levels shows a slight upward trend, indicating a potential increase in mortality rates with higher PM2.5 levels. This corresponds with the positive coefficients found in the regression models, supporting the hypothesis that poor air quality can affect lung health.

However, the graph for heart-related causes shows a relatively flat trend, which suggests a weak relationship between PM2.5 levels and mortality rates. However, the regression model indicated a significant relationship for Ischemic Heart Disease, overall these results likely suggest that other factors could be influencing the mortality rates that are not captured in this analysis.

The differences in results between the model analysis and correlation graphs can be attributed to several factors. Firstly, the complexity of environmental health relationships means that simple linear representations may not capture all relationships without the benefit of controlling for other variables. The regression model's focus on specific parameters may also lead to discrepancies, as it may not fully account for all the variables at play. The nature of the

data and the limitations of the models themselves can contribute to discrepancies between the two analyses. Additionally, the models account for overdispersion and other confounding variables, providing a more informed understanding of the relationship between air pollution and mortality rates than the data visualizations alone.

Overall, the findings from the models, supported by the trends from the graphs, suggest that PM2.5 has a varying impact on different health outcomes. There is a clear indication that higher levels of PM2.5 are associated with increased mortality from lung cancer and Ischemic Heart Disease. Alternatively, the weak graphical correlation in the heart-related graph does not diminish the relevance of the model's findings but rather underscores the need for a thorough analysis that accounts for a variety of factors affecting heart health.

5.2 Bias

Our analysis is based on the datasets available from provincial open databases, potentially introducing selection bias as these datasets may not represent the entire population of Alberta accurately. Certain demographics or regions may be omitted from the analysis, leading to biased conclusions. Our study does not account for all potential variables that could influence the relationship between air quality and mortality rates, such as socioeconomic factors, healthcare access, lifestyle habits, and weather conditions. Since we are only analyzing a few variables to find correlations towards respiratory-related mortality, these variables could introduce bias into the analysis. The datasets used in the analysis may also suffer from sampling bias, particularly if they do not adequately capture the entire population or if certain years or regions are overrepresented compared to others. This could affect the reliability of the findings. As a result, the accuracy and integrity of the data heavily depend on the government of Alberta's data collection reporting and practice for the three datasets used in our study.

5.3 Limitations

In our study, there might be several limitations affecting our conclusions. First, between 2010 and 2017, older equipment at Alberta's monitoring stations was replaced with new monitoring equipment for PM2.5. These new instruments measured an additional portion (semi-volatile) of the PM2.5 mass not captured by older instruments. As a result, the older monitoring equipment used between 2000 and 2017 likely underreported concentrations of PM2.5 under some conditions, after 2010, the increase in PM2.5 concentrations may be a result of changes in monitoring equipment (PM2.5 reference). Overall, the concentrations measured with the new monitors may not be directly comparable with measurements from years in which older instruments were used, possibly hindering trends in our study by some amount.

Furthermore, our study's analysis is limited by the time frame of the available datasets, which may not capture long-term trends or changes in air quality and mortality rates over extended periods. As a result, our findings may not be valid if generalized to other regions with different

environmental and demographic characteristics. While the study aims to investigate the relationship between air quality and mortality rates, it may be challenging to establish causation based solely on observational data. The findings may indicate correlations between air quality and mortality rates but cannot definitively prove causality.

Lastly, our datasets used are of relatively small size, and this may have significantly impacted the statistical power of the analysis. With fewer data points, the ability to detect meaningful relationships or patterns between variables may have deterred our models from finding any underlying correlations that were not found with the current size. This limitation may have reduced the reliability of the findings we concluded.

Appendix

A Model details

A.1 Posterior predictive check

Posterior Predictive Checks for General Model (Death Causes and Total Deaths):

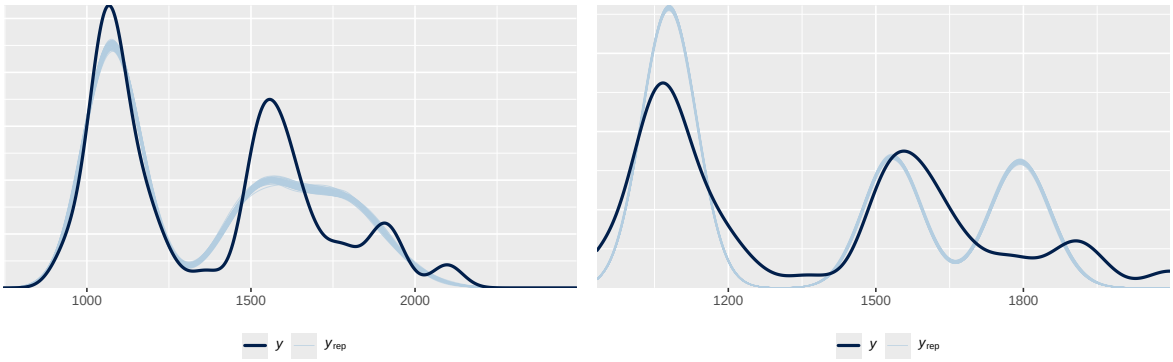
In Figure 7a (Posterior prediction check): This plot indicates how well the model's predictions align with the observed data. The overlap between the observed counts (y) and the posterior predictions (y_rep) suggests the model captures the underlying distribution of total death causes effectively.

In Figure 7b (Comparing the posterior with the prior): The plot displays a refinement from broad prior beliefs to more precise posterior beliefs after incorporating the data. The tighter posterior distribution signals that the data have had a substantial impact on the model's estimates.

Posterior Predictive Checks for Heart and Lung Models:

In Figure 8a (Posterior prediction check for the heart model): The graph shows the posterior predictions closely mirroring the observed mortality data, indicating that the heart model accurately reflects the real-world distribution of heart-related deaths.

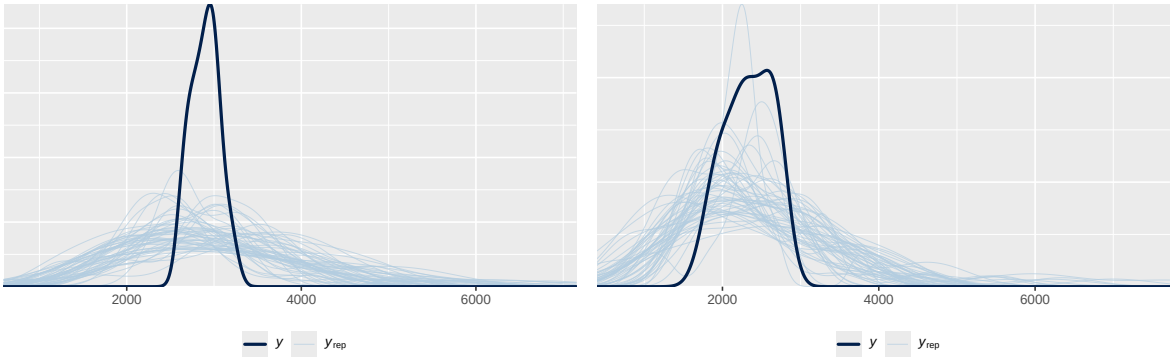
In Figure 8b (Posterior prediction check for the lung model): Similar to the heart model, this plot demonstrates that the lung model's predictions are consistent with the observed data, reaffirming the model's validity in describing lung-related mortality patterns.



(a) Posterior prediction check

(b) Comparing the posterior with the prior

Figure 7: Examining how the model fits, and is affected by, the data



(a) Posterior prediction check

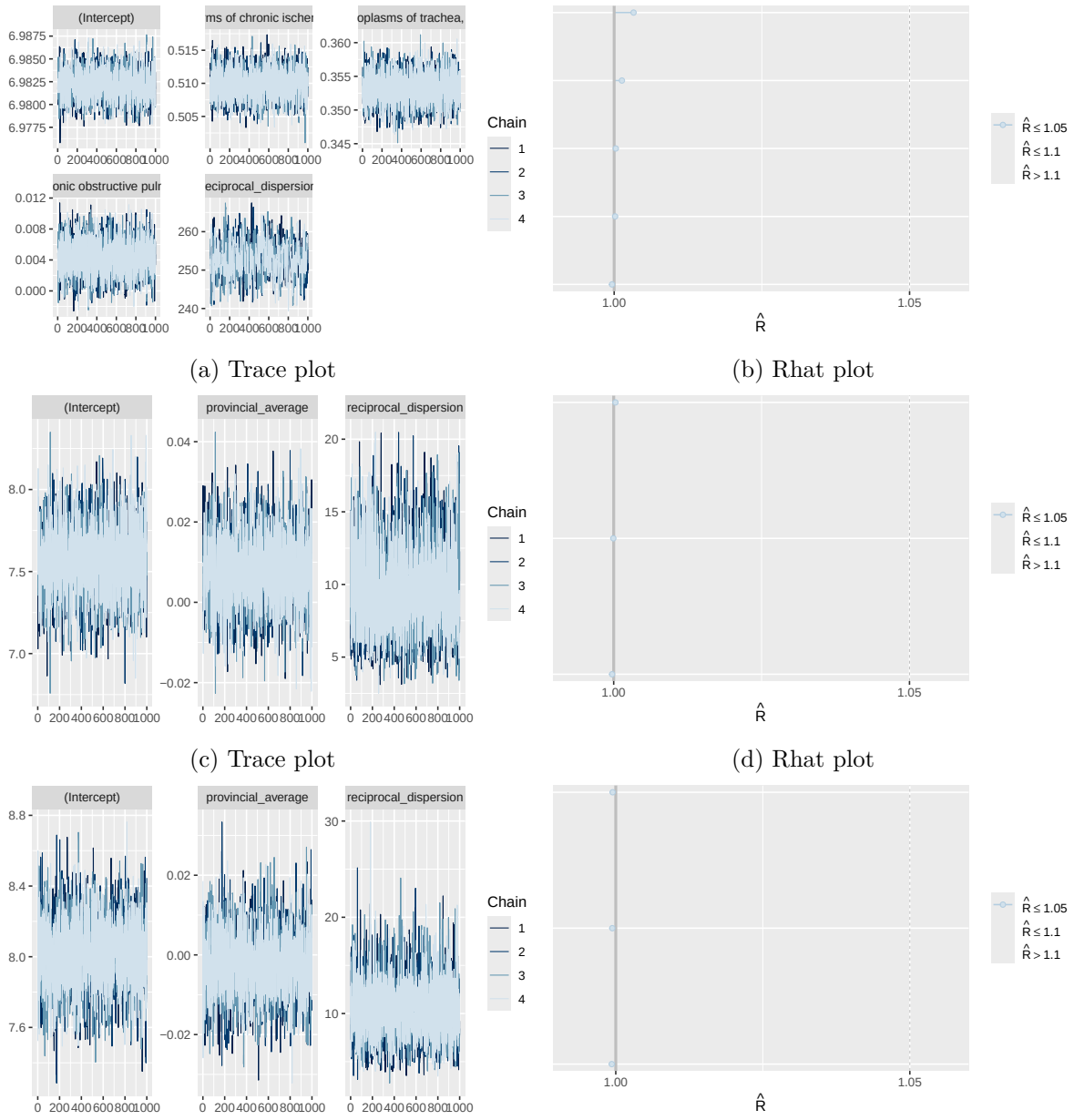
(b) Comparing the posterior with the prior

Figure 8: Examining how the model fits, and is affected by, the data

A.2 Diagnostics

Figure 9a: This trace plot shows the sampled values over iterations for various model parameters. The chains appear stable and well-mixed around a consistent mean, suggesting good convergence of the MCMC algorithm. Figure 9c: The trace plot for the lung model indicates that the chains for each parameter exhibit consistent behavior across iterations, without trending upwards or downwards, which implies convergence. Figure 9e: For the heart model, this trace plot displays a similar pattern of stability and suggests that the chains have converged, as evidenced by their hovering around a constant value. Rhat Plots:

Figure 9b: This Rhat plot displays values close to 1.00 for each chain, indicating that convergence has likely been achieved for the first model's parameters. Figure 9d: The Rhat values for the lung model are near the ideal of 1.00, suggesting that there is no evidence of non-convergence and that the model is well-fitted. Figure 9f: In the heart model's Rhat plot, the Rhat values are again close to 1.00, indicating successful convergence across the chains for all estimated parameters.



(a) Trace plot

(b) Rhat plot

(c) Trace plot

(d) Rhat plot

(e) Trace plot

(f) Rhat plot

Figure 9: Checking the convergence of the MCMC algorithm

References

Alberta Government. 2023. "Air Indicators – Fine Particulate Matter." https://www.alberta.ca/air-indicators-fine-particulate-matter.

Alexander, Rohan. 2024. "Telling Stories with Data." https://tellingstorieswithdata.com/13-ijaglm.html#negative-binomial-regression.

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.

Board, California Air Resources. 2024. "Inhalable Particulate Matter and Health (PM2.5 and PM10)." https://ww2.arb.ca.gov/resources/inhalable-particulate-matter-and-health.

Clarke, Erik, Scott Sherrill-Mix, and Charlotte Dawson. 2023. *Ggbeeswarm: Categorical Scatter (Violin Point) Plots*. https://github.com/eclarke/ggbeeswarm.

Clay Ford, University of Virginia Library, Statistical Research Consultant. 2023. "Getting Started with Negative Binomial Regression Modeling." https://library.virginia.edu/data/articles/getting-started-with-negative-binomial-regression-modeling#:~:text=As%20the%20dispersion%20parameter%20gets,converges%20to%20a%20Poisson%20distribution.&text=For%20each%20race%20the%20sample,It%20appears%20we%20have%20overdispersion.

Environment, Ministry of, and Protected Areas. 2021. "Status of Air Quality in Alberta." https://open.alberta.ca/dataset/9b00aab3-c37d-4080-854e-5f329c621b92/resource/057c65ac-7837-49bb-9528-38c2611540c4/download/epa-alberta-air-zones-report-2019-2021.pdf.

Firke, Sam. 2023. *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/.

government, Alberta. 2023a. "Air Quality Index by Municipality." https://open.alberta.ca/opendata/air-quality-index-by-municipality#detailed.

———. 2023b. "Leading Causes of Death." https://open.alberta.ca/opendata/leading-causes-of-death.

———. 2024. "Alberta." https://www.alberta.ca/.

IBM. 2021. "Goodness-of-Fit Statistics." https://www.ibm.com/docs/en/spss-modeler/saas?topic=uprasdrglm-goodness-fit-statistics.

Med, Lancet Respir. 2020. "Prevalence and Attributable Health Burden of Chronic Respiratory Diseases, 1990–2017: A Systematic Analysis for the Global Burden of Disease Study 2017." https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7284317/.

Müller, Kirill. 2020. *Here: A Simpler Way to Find Your Files*. https://here.r-lib.org/.

Müller, Kirill, and Hadley Wickham. 2023. *Tibble: Simple Data Frames*. https://tibble.tidyverse.org/.

Organization, World Health. 2024. "Air Quality, Energy and Health." https://www.who.int/teams/environment-climate-change-and-health/air-quality-energy-and-health/health-impacts#:~:text=The%20main%20pathway%20of%20exposure,and%20ultimately%20leading%20to%20disease.

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

Sachid Deshmukh, Vishal Arora, Michael Yampol. 2019. "Poisson Regression. Negative Binomial Regression. Multinomial Logistic Regression. Zero Inflated Poisson. Negative Binomial Regression." https://rpubs.com/myampol/Data621_Group4_HW5.

Slowikowski, Kamil. 2024. *Ggrepel: Automatically Position Non-Overlapping Text Labels with 'Ggplot2'*. https://ggrepel.slowkow.com/.

UCLA. 2021. "NEGATIVE BINOMIAL REGRESSION | STATA DATA ANALYSIS EXAMPLES." https://stats.oarc.ucla.edu/stata/dae/negative-binomial-regression/.

Venables, W. N., and B. D. Ripley. 2002. *Modern Applied Statistics with s*. Fourth. New York: Springer. https://www.stats.ox.ac.uk/pub/MASS4/.

Wickham, Hadley. 2007. "Reshaping Data with the reshape Package." *Journal of Statistical Software* 21 (12): 1–20. http://www.jstatsoft.org/v21/i12/.

———. 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, and Jennifer Bryan. 2023. *Readxl: Read Excel Files*. https://CRAN.R-project.org/package=readxl.

Wickham, Hadley, Romain François, Lionel Henry, Kirill Müller, and Davis Vaughan. 2023. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

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

Xie, Yihui. 2023. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.

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