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 15, 2024

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

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 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

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

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 key datasets were utilized to extract relevant variables for analysis, aiming to uncover the relationship 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 comprehensive 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, 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. 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. This dataset contains 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 Alberta.ca (Government of Alberta) 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 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 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) and Wickham et al. (2019a) for data cleaning and processing, utilizing packages like tidyverse (Wickham et al. 2019b) 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. 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 comprehensive analysis. Similar to the previous datasets, the raw mortality 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. The raw data on AQHI were filtered to include observations from the years 2011 to 2021 for consistency with other datasets. Additionally, the data were aggregated at the municipal level for further analysis.

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. A snapshot of this data is referenced in Table X. 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. These datasets were instrumental in examining the impact of PM2.5 on each type of illness, as previously discussed. Overall, our methodology ensured the creation of comprehensive datasets that allowed for a detailed investigation into the relationships between PM2.5 levels and various health outcomes in Alberta.

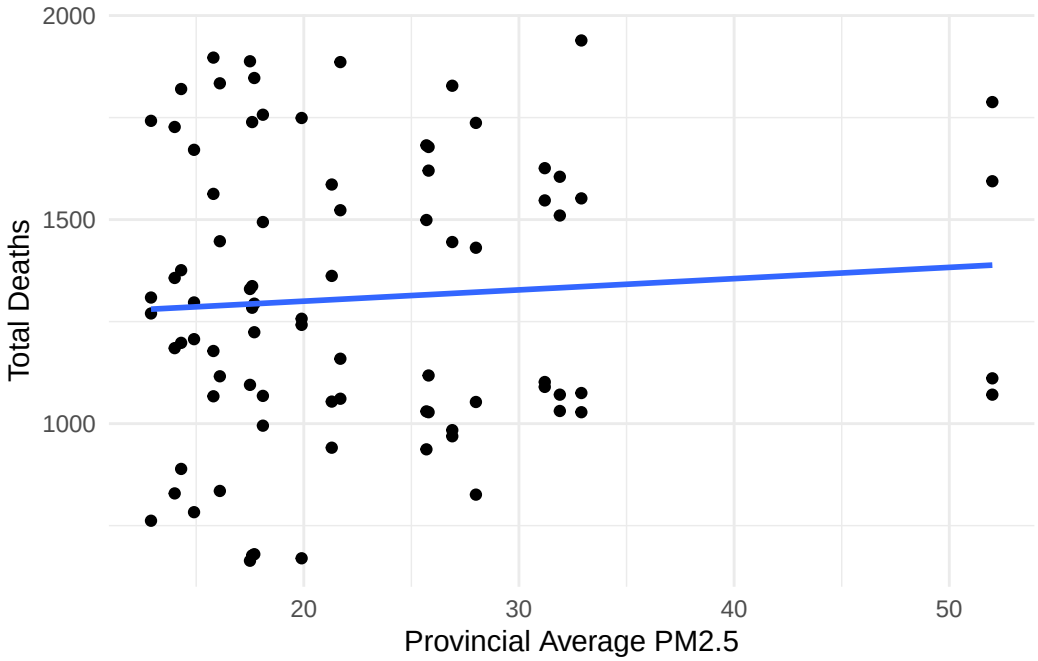


Figure 1: Provincial Average PM2.5 Quantity Vs Total Deaths

Talk more about it.

And also planes (**?@fig-planes**). (You can change the height and width, but don't worry about doing that until you have finished every other aspect of the paper - Quarto will try to make it look nice and the defaults usually work well once you have enough text.)

Talk way more about it.

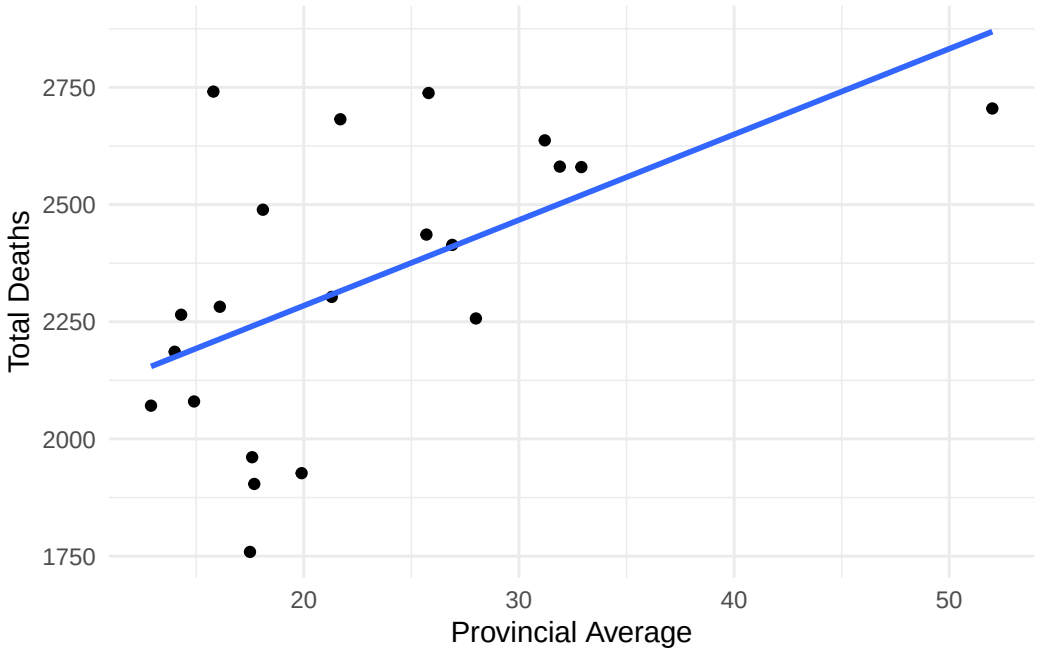


Figure 2: Lung Related Causes Mortality Rates vs Average PM2.5 Quantity

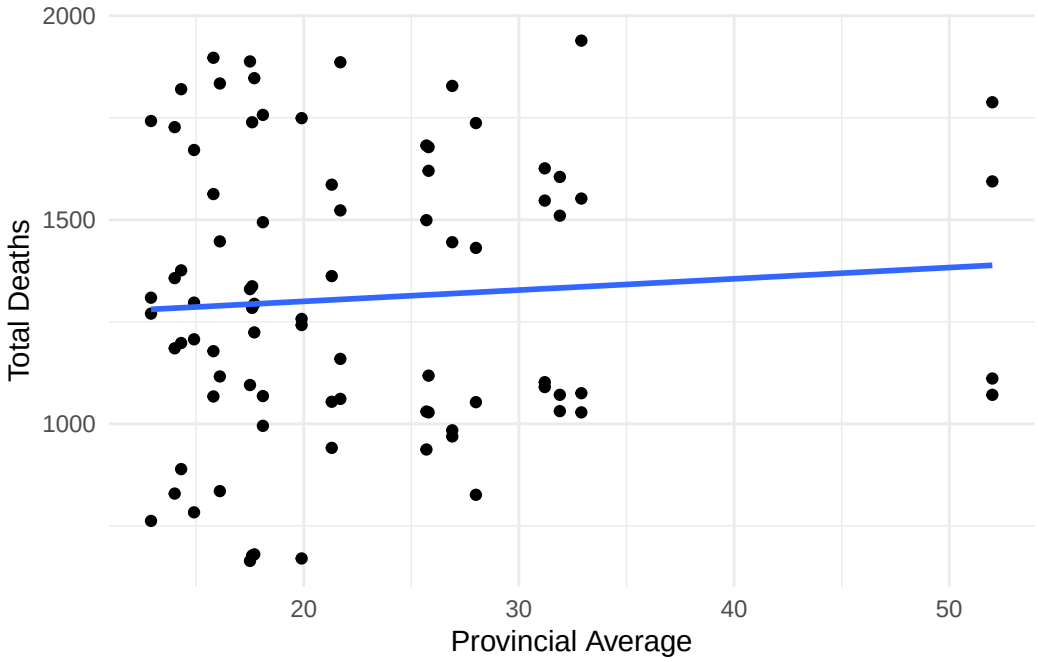


Figure 3: Heart Related Causes Mortality Rates vs Average PM2.5 Quantity

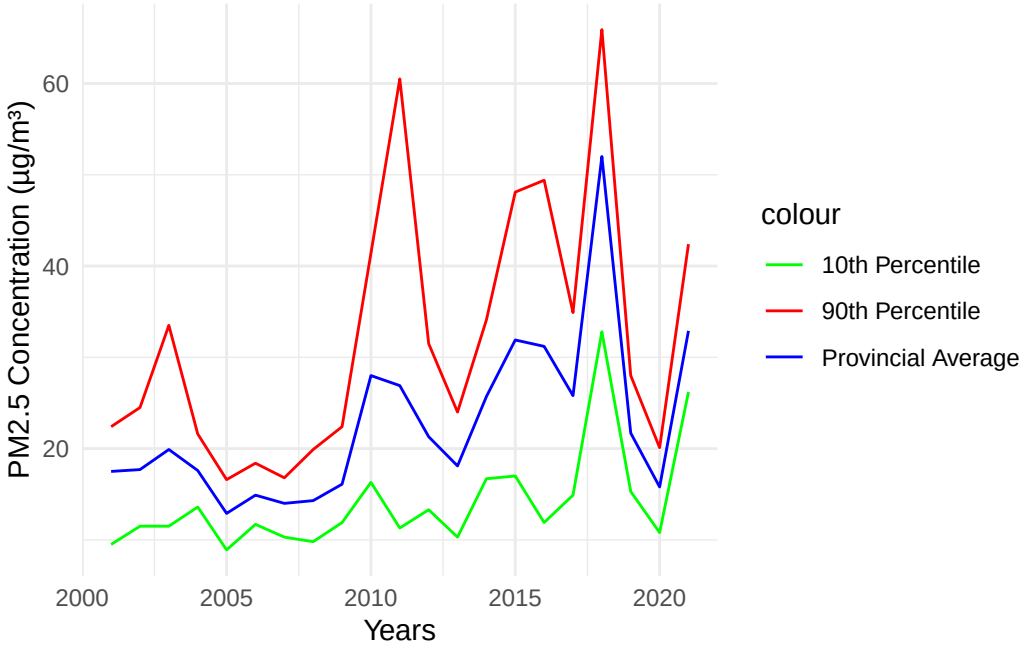


Figure 4: Annual Trends of PM2.5 Concentrations in Alberta

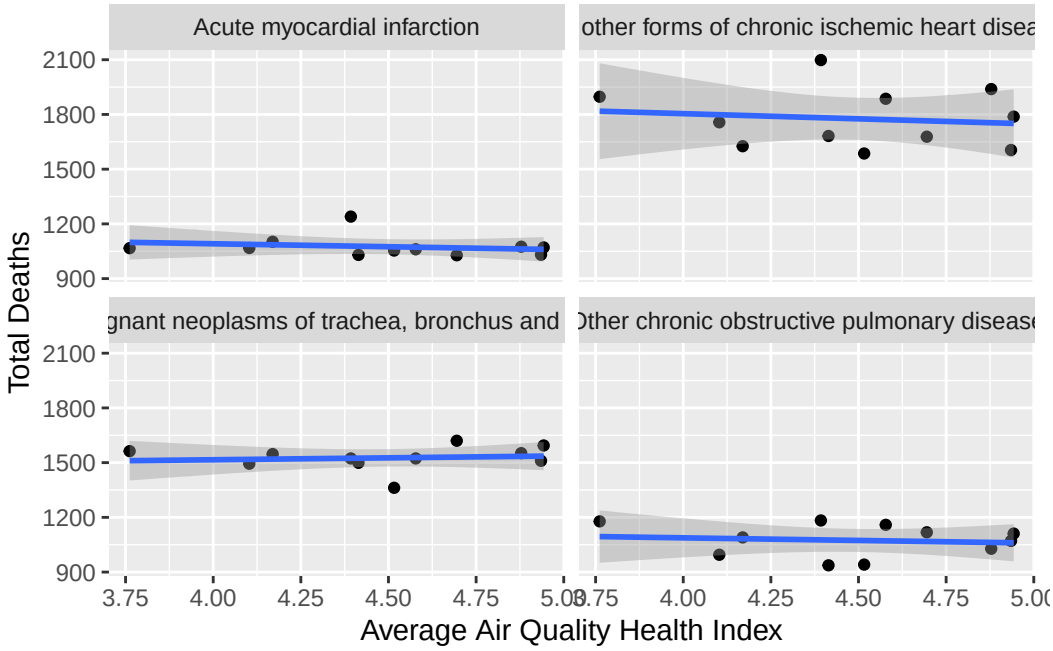


Figure 5: Mortality Rates vs. Air Quality Health Index

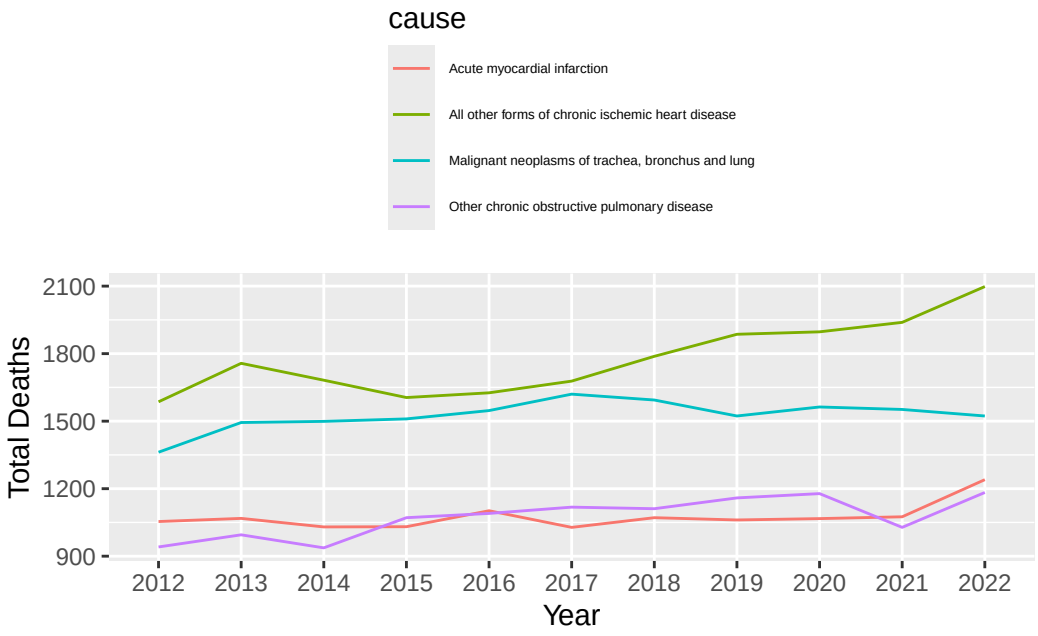


Figure 6: Total Deaths by Year for Each Cause

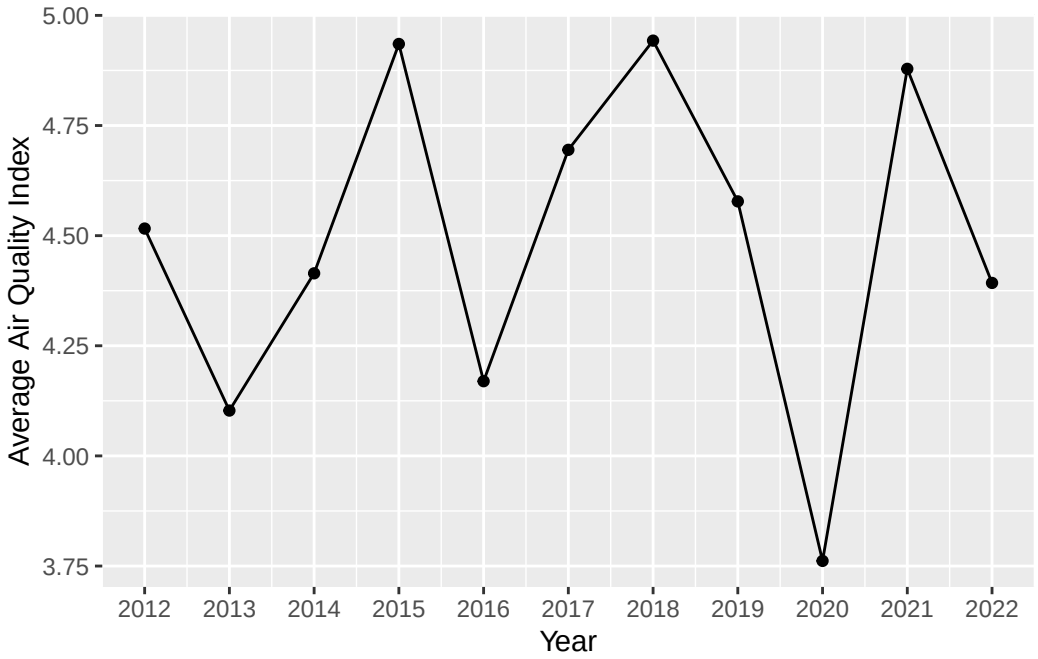


Figure 7: Average Air Quality Health Index by Year

3 Model

The goal of our modelling strategy is twofold. Firstly,…

Here we briefly describe the Bayesian analysis model used to investigate… Background details and diagnostics are included in Appendix **??**.

3.1 Model set-up

Define y_i as the number of seconds that the plane remained aloft. Then β_i is the wing width and γ_i is the wing length, both measured in millimeters.

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

We run the model in R (R Core Team 2023) using the `rstanarm` package of (**rstanarm?**). We use the default priors from `rstanarm`.

3.1.1 Model justification

We expect a positive relationship between the size of the wings and time spent aloft. In particular…

We can use maths by including latex between dollar signs, for instance θ.

4 Results

Our results are summarized in **?@tbl-modelresults**.

5 Discussion

5.1 First discussion point

If my paper were 10 pages, then should be be at least 2.5 pages. The discussion is a chance to show off what you know and what you learnt from all this.

Table 1: 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

Table 2: 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

Table 3: 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

5.2 Second discussion point

5.3 Third discussion point

5.4 Weaknesses and next steps

Weaknesses and next steps should also be included.

Appendix

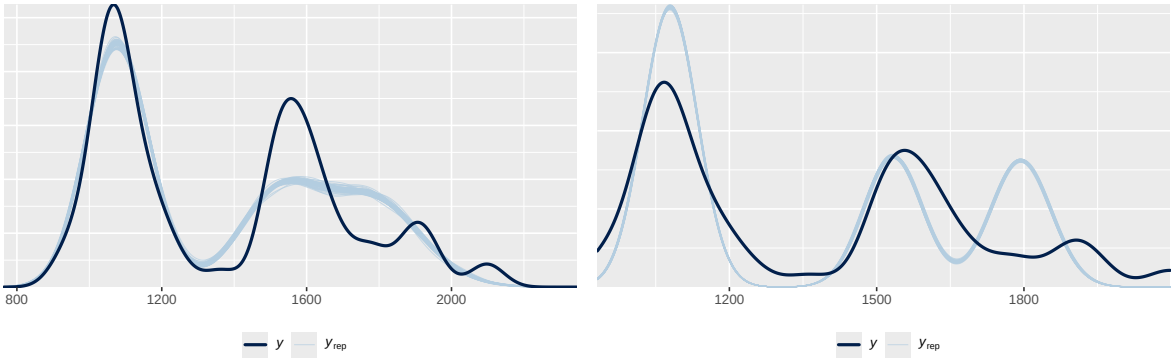
A Additional data details

B Model details

B.1 Posterior predictive check

In **?@fig-ppcheckandposteriorvsprior-1** we implement a posterior predictive check. This shows...

In **?@fig-ppcheckandposteriorvsprior-2** we compare the posterior with the prior. This shows...



(a) Posterior prediction check

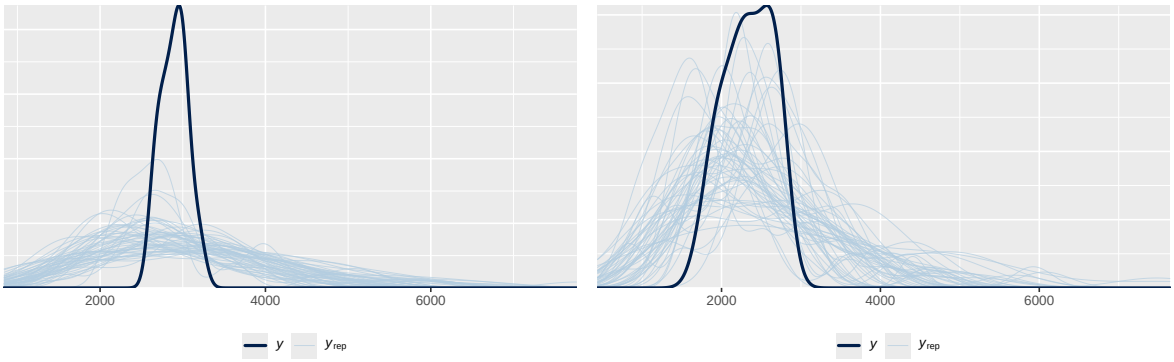
(b) Comparing the posterior with the prior

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

B.2 Diagnostics

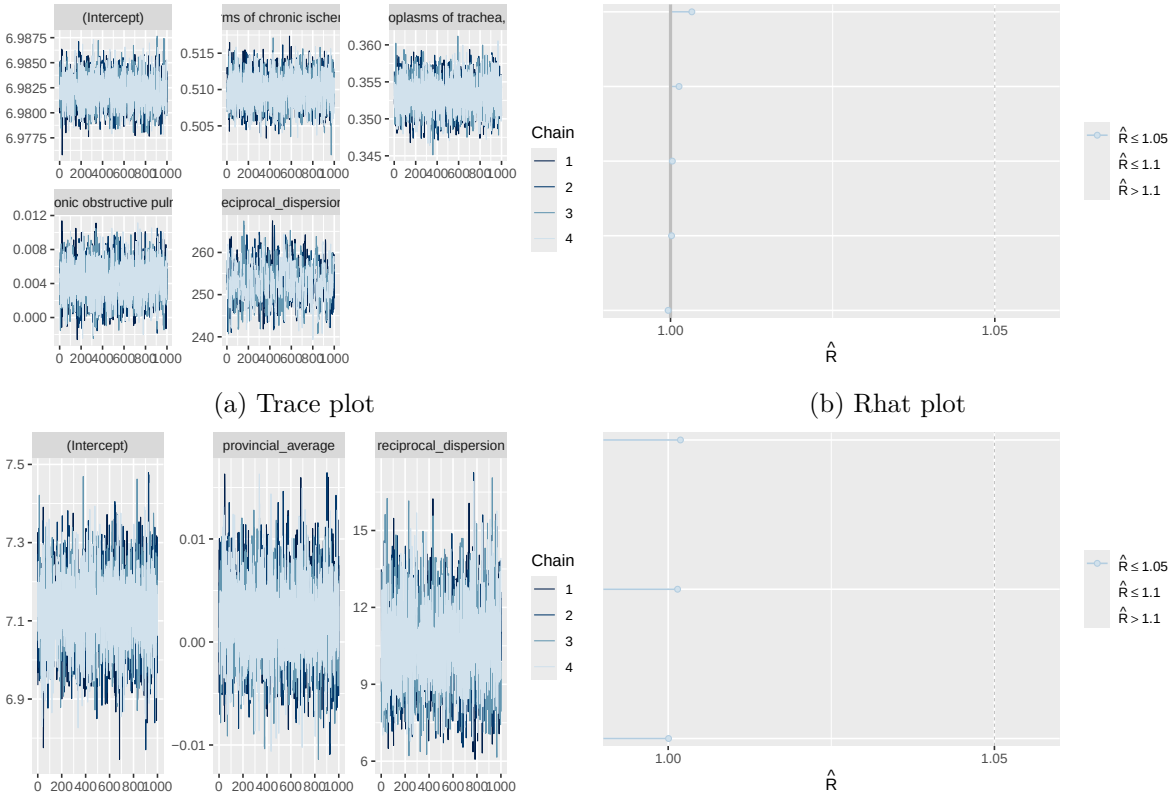
Figure **??** is a trace plot. It shows... This suggests...

Figure **??** is a Rhat plot. It shows... This suggests...



(a) Posterior prediction check (b) Comparing the posterior with the prior

Figure 9: Examining how the model fits, and is affected by, the data



(a) Trace plot (b) Rhat plot

(c) Trace plot (d) Rhat plot

Figure 10: 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.

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.

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/.

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/.

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

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. 2019b. "Welcome to the tidyverse." *Journal of Open Source Software* 4 (43): 1686. https://doi.org/10.21105/joss.01686.

———, et al. 2019a. "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.