Number of COVID-19 Cases in Canada By Region, Age, and Sex*

Predicting the Factors that Influence the Number of COVID-19 Cases and Deaths

Bruce Zhang

November 24, 2024

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

1 Introduction

The COVID-19 pandemic is something that is difficult to forget. Although it is now in the past, the effects that it had on society and how we function as individuals is still profound. At the same time, the statistical information that sprouted from the pandemic has a high value for analysis. This is not only to better understand the pandemic and how it affected populations but also to form a strong idea of its patterns to suggest tactics for dealing with pandemics shall something similar happen in the future.

Studies have been done in recent years in attempts to utilize the data that was gathered during COVID times to better advise future preventions and treatment for pandemic situations. Many studies look into the relationship between demographic information of populations and their likelihood of getting diagnosed or fatality. One study found that communities on hillsides and other locations with enhanced ventilation and oxygen had reduced numbers of COVID cases (Sharma et al. 2023). Other studies have found relationships between exercise and cardiovascular function and the likelihood of getting COVID (Yang et al. 2024). Although many relationships have been examined, there has been a lack of studies focusing on a broad set of demographic traits such as age and sex. There has also not been cross comparisons of region and demographic aspects. Another gap is the lack of country-specific analysis, particularly Canada.

This analysis looks at COVID-19 data from 2022 in Canada and analyzes how the different population and individual-level variables, such as age, sex, and region of life, influenced the

[*] Code and data are available at: https://github.com/brucejczhang/covid_data.

risk of being diagnosed with COVID and the risk of fatality. The paper uses the demographic data gathered from 2022 to predict the characteristics of populations and individuals that may be at higher risk of fatality once they have been diagnosed with COVID. This analysis may have broader implications for preventative measures for flues, other diseases, and pandemics in the future.

The analysis focuses on the outcome variable of predicted probability of high risk, which is a function of number of cases and number of deaths per datapoint and is further defined in Section 3. The model predicts the probability of high risk as a result of a series of predictor variables including region, sex, and age of the reported data.

Results paragraph (what was found?)

Why it matters paragraph (why it matters?)

The remainder of this paper is structured as follows. Section 2 highlights the characteristics of the dataset. The section summarizes the data through a series of summary statistics (Figure 2) and represents the data in a visual way where specific trends and patterns can be observed (Figure 3, Figure 4). Section 3 creates a logistic regression model predicting the probability of high risk depending on the region, sex, and age of the data point. The section includes model figures that display the predicted probabilities of high risk for the predictor variables respectively.

2 Data

2.1 Overview

I use the statistical programming language R (R Core Team 2023) to analyze the data and to create graphs and models. Th packages that were used include tidyverse (Wickham, Averick, et al. 2023), tidyr (Wickham, Henry, and Vaughan 2023), dplyr (Wickham, François, et al. 2023), caret (Kuhn 2023), and pROC (Robin et al. 2011). My data (Open Science Framework (OSF) 2022) was obtained from COVerAGE-DB, which was housed in Open Science Framework. The data anlysis was conducted based on the guidance of Alexander (2023).

2.2 Measurement

The dataset used in this analysis

Some paragraphs about how we go from a phenomena in the world to an entry in the dataset.

2.3 Data Cleaning

The original dataset contained data from five different countries including. For the purpose to focus on data in Canada for this analysis, the datapoints for other countries were removed. Further cleaning was done to remove any problematic datapoints such as when the number of deaths exceeded the number of cases.

2.4 Outcome Variables and Predictor Variables

The cleaned dataset contains variables including the country, region, date, sex, and age where the number of cases, deaths, and tests were reported. Figure 1 shows a sample of the dataset, presenting the first six rows of the analysis data. The outcome variables highlighted include the total number of cases and the total number of deaths.

Country	Region	Date	Sex	Age	Cases	Deaths	Tests	CFR
Canada	Urban	2022-01-01	m	0	1000	239	26704	23.900000
Canada	Rural	2022-01-01	m	0	5850	346	40541	5.914530
Canada	Urban	2022-01-01	m	0	1513	24	23200	1.586253
Canada	Rural	2022-01-01	m	0	14629	447	10587	3.055575
Canada	Urban	2022-01-01	m	0	5523	29	27726	0.525077
Canada	Rural	2022-01-01	m	0	783	134	8949	17.113665

Figure 1: First six rows of the dataset after cleaning

		Mean	Total	Total	Average	Average	CFR
Region	Sex	Age	Cases	Deaths	Cases	Deaths	(%)
Rural	f	50.28	6626386	163171	10178.78	250.65	2.46
Rural	m	49.58	6631565	169585	10202.41	260.90	2.56
Urban	f	49.98	6633795	159529	10112.49	243.18	2.40
Urban	m	49.88	6592721	162133	10111.54	248.67	2.46

Figure 2: Summary Statistics of COVID-19 Cases in Canada

Figure 2 further breaks down the data summarizing the mean number of cases and deaths by region and sex. This creates an all-rounded summary of the dataset and gives a preliminary idea on the differences in cases and deaths depending on the categories of sex and region.

2.4.1 Number of Cases and Number of Deaths

The total and average numbers of cases and deaths are key outcome variables that can be correlated with the categorical predictors. These number can demonstrate the differences in COVID-19 susceptibility and threat level based on the characteristics of the region and the individual.

2.4.2 Case Fatality Rate

The case fatality rate (CFR) is an additional aspect of the summary that was calculated to standardize the deaths and cases by category. This can be used as a processed outcome variable that can better represent the dataset and aid modeling later on. The total number of deaths and cases may lead to inaccurate representations of a region, sex, or age group due to the absolute number of individuals that fall within the category. The CFR value allows better comparison across categories.

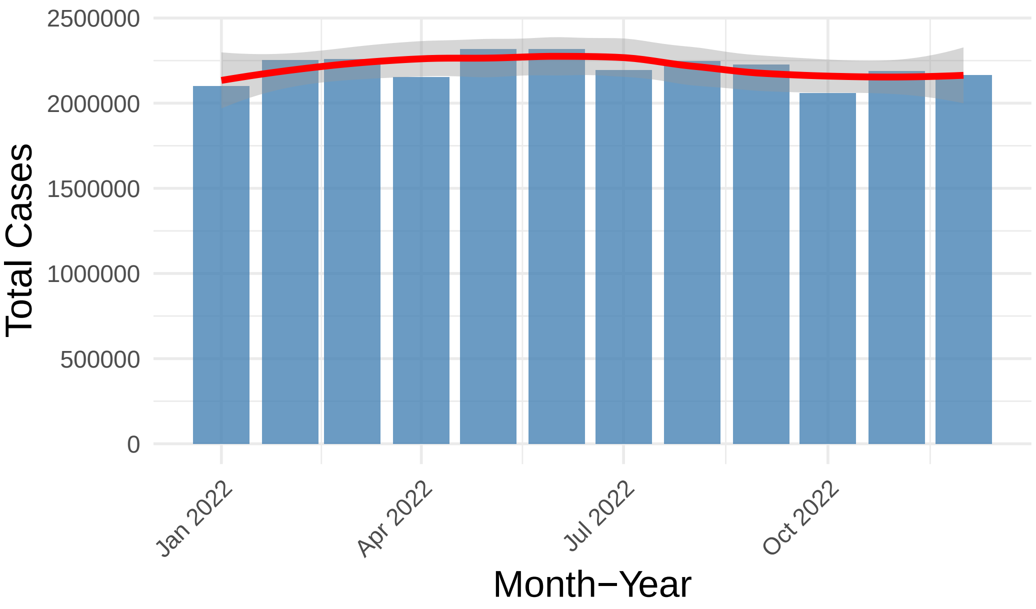

Figure 3: Number of cases per month in 2022 in Canada

2.4.3 Time

The time of year can also be correlated with the likelihood of getting COVID-19 independent of other predictor variables. As shown in Figure 3, there is a weak pattern of increased total number of cases in the early summer months of May and June. The number of cases in winter and near winter months, specifically January and October, seem to be lower.

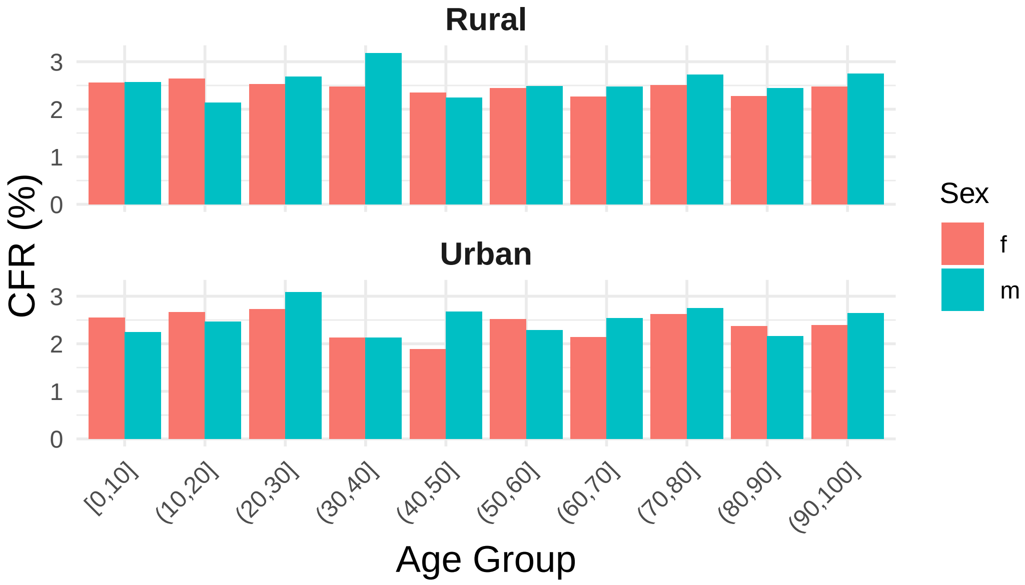

Figure 4: Case Fatality Rate (CFR) by Age Group, Region, and Sex

2.4.4 Age, Sex, and Region

Age, sex, and region are key predictor variables that are examined in this analysis. In Figure 4, these variables are organized relative to the CFR, which calculates a rate based on the number of cases and number of deaths and allows standardized comparisons across different predictor variables. Figure 4 shows that the urban and rural regions have similar patterns in CFR across different age groups. The CFR for males tend to be higher more often than not compared to the CFR of females, indicating that males may have a higher likelihood of death after diagnosis of COVID in Canada. The pattern of CFR in relation to age seem to vary quite significantly between urban and rural regions and between male and females. For females, the highest CFR seems to be for individuals aged under 30 and for those around 70 to 80 regardless of region. Male CFR values peak at age interval 20 to 30 for urban regions and 30 to 40 for rural regions.

Figure 5 shows a more direct comparison of the CFR between urban and rural regions of Canada, combining different age groups and sexes. This figure shows that the CFR in rural regions is noticeably higher than that of urban. Although the difference is small, it is still a considerable size difference when considering the number of people that a small percentage can be responsible for, based on the summary statistics from Figure 2.

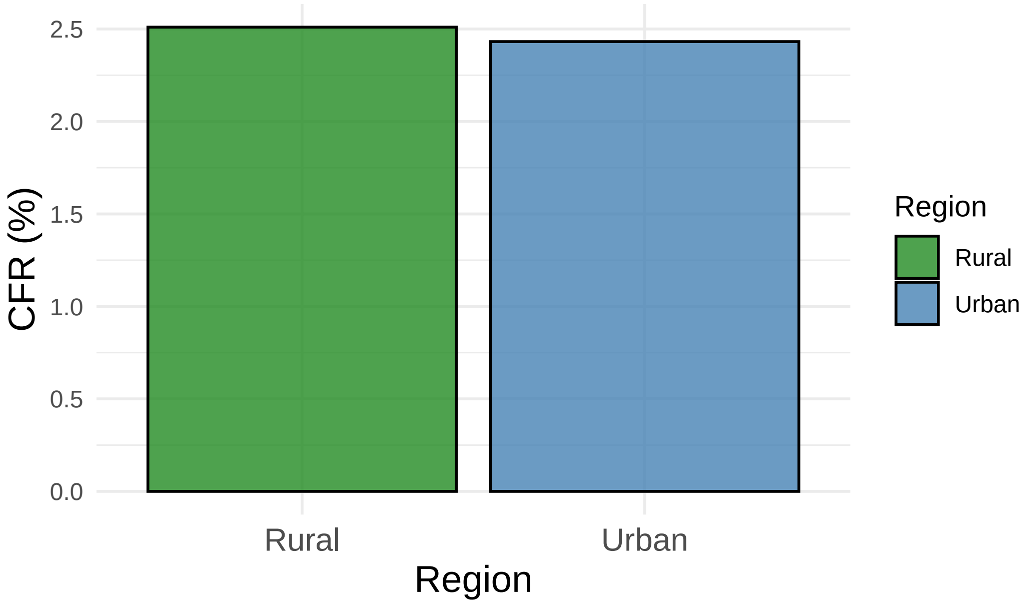

Figure 5: Case Fatality Rate (CFR) by Region in 2022 in Canada

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 B.

3.1 Model set-up

A logistic model has been selected as part of this analysis to examine the relationship between region, sex, age, and the number of cases and deaths of COVID-19.

The logistic regression model predicts the probability of a region or individual being classified as "high risk" based on key predictors. In this case, "high risk" is defined as regions where the death rate (CFR) exceeds the median death rate across the dataset.

The CFR is calculated as:

$$\text{CFR} = \frac{\text{Deaths}}{\text{Cases} + 10^{-6}} \times 100 \tag{1}$$

The **box-and-whisker plot** visualizes the distribution of predicted probabilities ((p)) across different **age intervals**. Each box represents the interquartile range (IQR) of predicted probabilities, with the line inside the box indicating the median predicted probability for that age interval. Whiskers extend to 1.5 times the IQR, and any points outside this range are considered outliers. This visualization highlights the variability of predicted probabilities within each age group.

The logistic regression model equation is:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Tests} + \beta_2 \cdot \text{Age Interval} + \epsilon \tag{2}$$

where:

- p: Predicted probability of being high risk (p = Predicted Probability).
- logit(p): Log-odds of the high-risk classification.
- β_0: Intercept of the model.
- β_1, β_2: Coefficients for the predictors (e.g., number of tests, age interval).
- ϵ: Error term.

This model explains how testing rates and age intervals affect the likelihood of a region or demographic being classified as high risk. The box-and-whisker plot shows the variability of predicted probabilities across age groups, helping to identify specific age intervals with consistently higher or lower risks.

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**.

```

Call:
glm(formula = High_Risk ~ Tests + Region + Age, family = binomial(), 
    data = filtered_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)  
(Intercept)  1.928e-01  1.060e-01   1.818   0.0690 .
Tests       -5.233e-06  2.696e-06  -1.941   0.0522 .
RegionUrban -4.139e-02  7.844e-02  -0.528   0.5978  
Age         -8.753e-04  1.243e-03  -0.704   0.4812  
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3616.8  on 2608  degrees of freedom
Residual deviance: 3612.2  on 2605  degrees of freedom
AIC: 3620.2

Number of Fisher Scoring iterations: 3
```

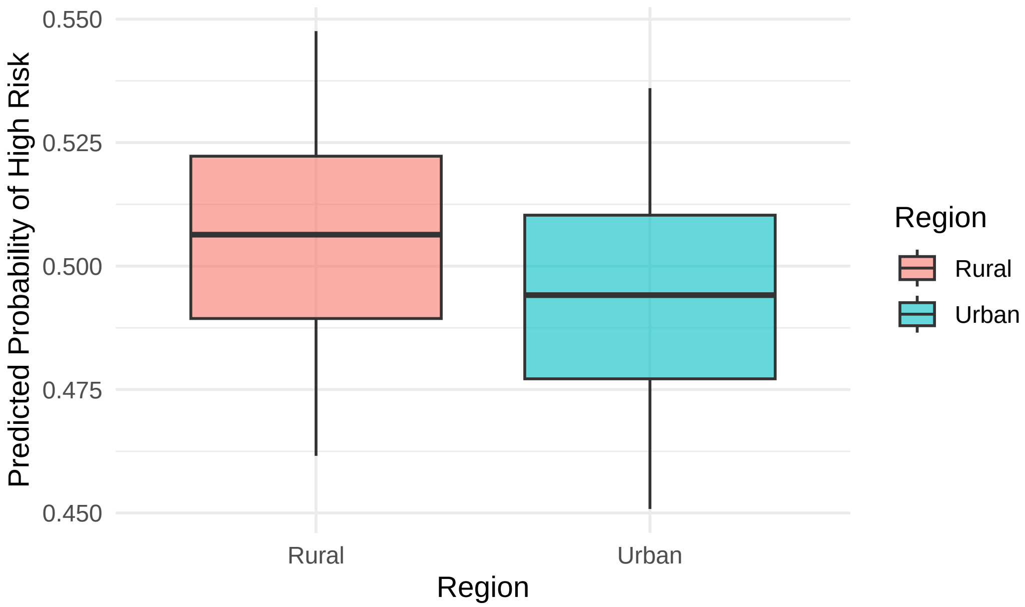

Figure 6: Predicted probability of high risk of COVID-19 by region (Urban vs. Rural)

```
Call:
glm(formula = High_Risk ~ Tests + Sex + Age, family = binomial(),
    data = filtered_data)

Coefficients:
              Estimate Std. Error z value Pr(>|z|)
(Intercept)  1.453e-01  1.072e-01   1.355   0.1754
Tests       -5.276e-06  2.694e-06  -1.959   0.0502 .
Sexm         5.524e-02  7.839e-02   0.705   0.4810
Age         -8.701e-04  1.243e-03  -0.700   0.4839
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3616.8  on 2608  degrees of freedom
Residual deviance: 3612.0  on 2605  degrees of freedom
AIC: 3620

Number of Fisher Scoring iterations: 3
```

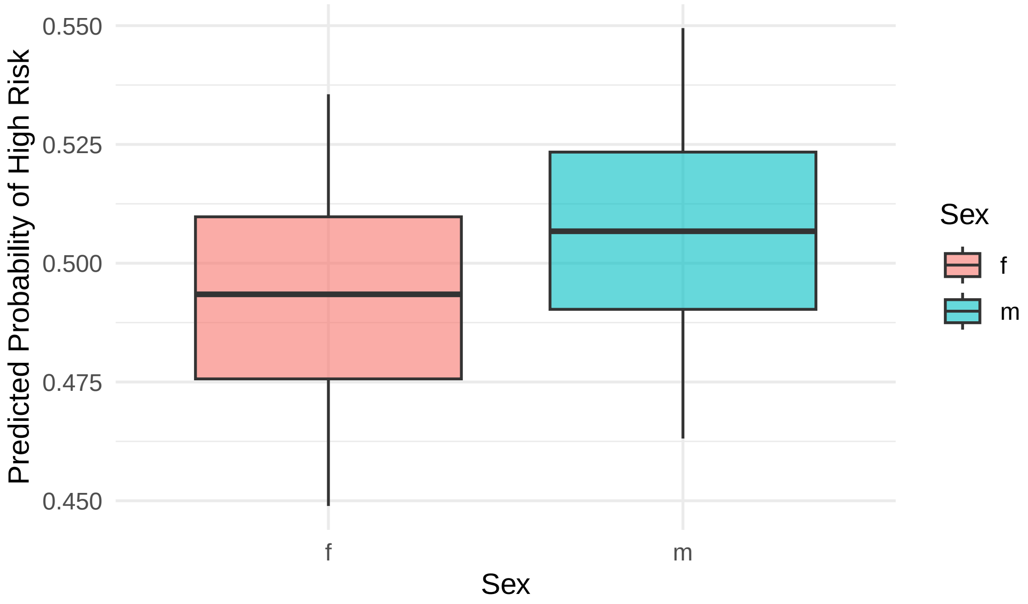

Figure 7: Predicted probability of high risk based on sex

```
Call:
glm(formula = High_Risk ~ Tests + Age_Interval, family = binomial(),
    data = filtered_data)

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)           5.286e-02  1.454e-01   0.363   0.7163
Tests                -5.473e-06  2.700e-06  -2.027   0.0427 *
Age_Interval[10,20)   2.752e-01  1.846e-01   1.490   0.1361
Age_Interval[20,30)   6.756e-02  1.842e-01   0.367   0.7137
Age_Interval[30,40)   2.888e-01  1.847e-01   1.564   0.1179
Age_Interval[40,50)   9.658e-02  1.842e-01   0.524   0.6000
Age_Interval[50,60)  -4.542e-02  1.841e-01  -0.247   0.8052
Age_Interval[60,70)   5.578e-03  1.840e-01   0.030   0.9758
Age_Interval[70,80)  -5.016e-02  1.848e-01  -0.271   0.7860
Age_Interval[80,90)   8.543e-02  1.840e-01   0.464   0.6424
Age_Interval[90,100]  8.571e-02  1.601e-01   0.535   0.5923
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 3616.8  on 2608  degrees of freedom
Residual deviance: 3605.5  on 2598  degrees of freedom
AIC: 3627.5

Number of Fisher Scoring iterations: 3
```

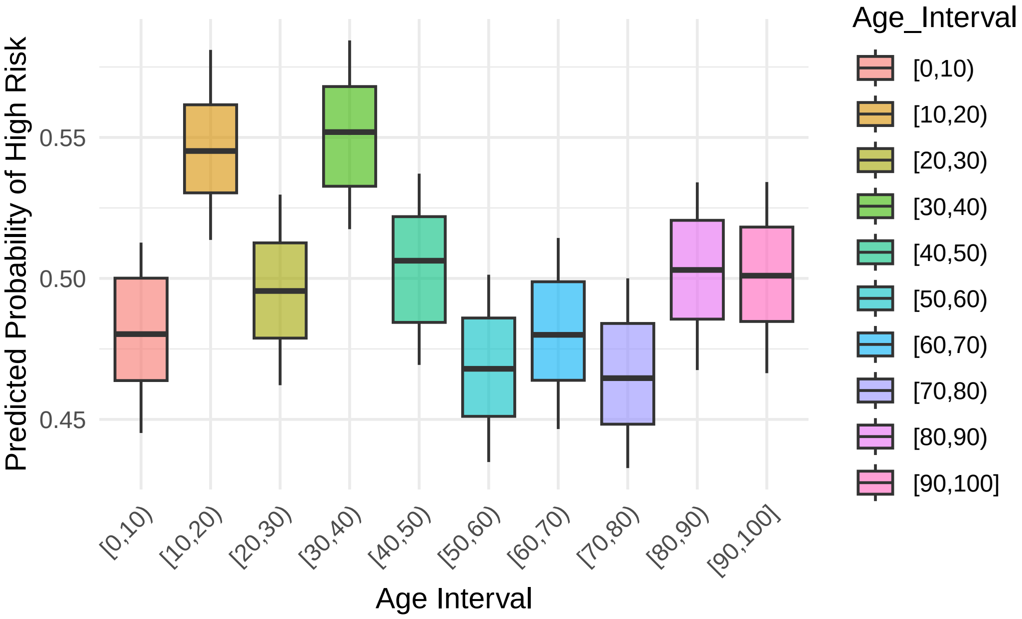

Figure 8: Predicted probability of high risk based on age intervals

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.

5.2 Second discussion point

Please don't use these as sub-heading labels - change them to be what your point actually is.

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

B.2 Diagnostics

References

Alexander, Rohan. 2023. *Telling Stories with Data*. Chapman; Hall/CRC. https://tellingstorieswithdata.com/.

Kuhn, Max. 2023. *Caret: Classification and Regression Training*. https://CRAN.R-project.org/package=caret.

Open Science Framework (OSF). 2022. "COVID-19 Cases, Deaths, and Tests by Age, Gender, and Region." Open Science Framework (OSF). https://osf.io/43ucn.

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

Robin, Xavier, Natacha Turck, Alexandre Hainard, Natalia Tiberti, Frédérique Lisacek, Jean-Charles Sanchez, and Markus Müller. 2011. "pROC: An Open-Source Package for r and s+ to Analyze and Compare ROC Curves." *BMC Bioinformatics* 12: 77. https://doi.org/10.1186/1471-2105-12-77.

Sharma, Rajesh, Anjali Singh, Poonam Verma, and Suresh Kumar. 2023. "Analyzing Uttarakhand's COVID-19 Outbreak: Demographic Insights and Strategies for Future Pandemic Prevention." *Current Medical Research and Opinion* 39 (1): 45–52.

Wickham, Hadley, Mara Averick, Jennifer Bryan, Winston Chang, Lucy D'Agostino McGowan, Romain François, Garrett Grolemund, et al. 2023. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://CRAN.R-project.org/package=tidyverse.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2023. *Dplyr: A Grammar of Data Manipulation*. https://CRAN.R-project.org/package=dplyr.

Wickham, Hadley, Lionel Henry, and Davis Vaughan. 2023. *Tidyr: Tidy Messy Data*. https://CRAN.R-project.org/package=tidyr.

Yang, Ziang, Xieraili Tiemuerniyazi, Fei Xu, Yang Wang, Yang Sun, Peng Yan, Liangxin Tian, et al. 2024. "Partial Cardiac Denervation to Prevent Postoperative Atrial Fibrillation After Coronary Artery Bypass Grafting: The pCAD-POAF Randomized Clinical Trial." *Progress in Cardiovascular Diseases* 80: 1–9.