Number of COVID-19 Cases in Canada in 2022 By Region, Age, and Sex*

Predicting the Factors that Influence the Probability of High Risk of COVID-19

Bruce Zhang

December 3, 2024

This paper examines the demographic and regional factors influencing high-risk classifications for COVID-19 in Canada using logistic regression models, with predictors including age, sex, and region. The analysis reveals that rural regions, males, and specific age intervals exhibit higher probabilities of being classified as high risk, emphasizing disparities in healthcare access and outcomes. These findings highlight the importance of targeted public health interventions, such as resource allocation and vaccination prioritization, to mitigate these risks and can potentially guide interventions for future pandemics. By identifying key predictors and their implications, this study provides a scalable framework for improving pandemic preparedness and understanding health disparities.

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

Table of contents

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 noticeable. 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 area where research is needed is for 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 demographic variables, such as age, sex, and region of life, influenced the risk of fatality from COVID-19. The analysis focuses on isolating these demographic factors and examining their individual effects. 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.

One importance of this information to be present is mainly for Canada-specific situations and patterns that may not be seen in a global scale analysis. The geographic location, temperature, population distribution, ethnic diversity, and other features of Canada is incredibly unique. The patterns present in other countries cannot be generalized to a smaller scale. From a public health standpoint, global health data is key in understanding overarching trends of epidemiology, but country specific (sometimes community specific) analyses are key for creating predictions based on a suitably-sized dataset (Beaglehole and Bonita 2010).

Another key importance of this analysis is its focus on basic demographic factors rather than specific health factors. It is often that a health issue, COVID-19, is examined through the lens of other health factors (Hollingshaus and Harris 2024). For example, how does the presence of another long-term disease influence the risk of COVID-19. Looking at surface level demographic information can provide information that is useful to a broader audience. Specific diseases may not be possessed by everyone, but demographic information is. This information is also known by individuals, which will help them identify the category they fall into and examine their risk of COVID-19 based on their demographic information.

The estimand of this analysis is the correlation between demographic information, including age, sex, and region, and the predicted probability of high risk for COVID-19. 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.

The analysis showed that there are disparities in the risk of severe COVID-19 outcomes across demographic and regional categories. Logistic regression models demonstrated that individuals in rural regions consistently faced higher probabilities of being classified as high risk, with males and specific age intervals further amplifying this vulnerability. These findings were supported by consistent trends in summary statistics and model visualizations, which showed the interplay between predictors such as age, region, and sex in influencing COVID-19 outcomes.

This analysis may have broader implications for preventative measures for flues, other diseases, and pandemics in the future. Understanding the reasons behind COVID-19 severity can craft equitable and effective public health responses. By identifying which demographic and regional groups face the greatest risks, the results of the analysis can guide resource distribution, including vaccination programs and medical support, to areas where they are most needed. These are applicable not only to COVID-19 but also to future public health challenges.

The remainder of this paper is structured as follows. The data section (Section 2) highlights the characteristics of the dataset. The section summarizes the data through a series of summary statistics and represents the data in a visual way where specific trends and patterns can be observed. The model section (Section 3) creates a logistic regression model predicting the probability of high risk depending on the region, sex, and age of the data point. This section presents the mathematical basis of the model and justifies the selection for the logistic regression model over others. The results section (Section 4) includes model figures that display the predicted probabilities of high risk for the predictor variables respectively. The discussion section (Section 5) goes over how the results from the modeling can be interpreted in the context of the real world and how these findings can be utilized to advise public health measures. It also discusses limitations of the analysis and future directions.

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), arrow (Apache Software Foundation 2024), DataExplorer (Y. Wei, Ma, and Liu 2020), corrplot (T. Wei, Li, and Zhang 2020), caret (Kuhn 2023), httr (Wickham et al. 2020), here(Ram et al. 2020), 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 data collection process is conducted by hospitals and governmental institutions that get first-handed numbers of COVID-19 patient information. Individuals who have symptoms and choose to visit a hospital will get tested, this number will be reported as the number of tests. Those that test positive for COVID-19 in the hospitals will be reported as the number of cases. Those who report deaths to the government level and indicate that the cause is due to COVID-19 will be recorded as the number of deaths due to COVID.

The data that was collected through hopistals and government institutions was gathered by COVerAGE-DB, a database that focuses on COVID related data. COVerAGE-DB consolidates this information into the dataset that was used in the analysis. COVerAGE-DB often gathers data relating to COVID such as the number of cases, deaths, tests, and vaccinations through governmental institutes such as health ministries and statistical offices. This data is then organized by other variables such as sex, region, country, and age and presented to the general public.

The data collected through hospitals or governmental institutes only reflect individuals who have visited hospitals for their symptoms and tested positive for COVID-19 there. This poses a key limitation of potential inaccuracy of numbers due to the large portion of individuals who self-test PCR, visit PCR testing stations, or don't get tested and stay home despite having symptoms. These situations were common given the situation in Canada in 2022, where individuals were encouraged to stay home if symptoms were experienced rather than visit hospitals to report their results.

The dataset reflects a transformation of real-world phenomena—COVID-19 cases, deaths, and testing rates—into structured entries that can be analyzed. It also reflects the data collection challenges and biases of the real world, potentially understating values. Each entry in the dataset represents aggregated information collected from health departments, governmental organizations, and research institutions worldwide. These organizations report COVID-19 statistics at varying levels of granularity, such as daily case counts or weekly summaries, which are then standardized and compiled into the dataset. This effort of data collection allows the real-world challenges, as accurately as possible despite obvious limitations, of facing COVID-19 to be transformed into quantitative measures that can be analyzed and modeled.

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. This was the initial step in data cleaning.

Further data cleaning was undertaken to ensure the quality and integrity of the dataset. First, missing values were identified and handled by checking each column. Any missing values were flagged and removed. Second, the dataset was checked for duplicates to avoid redundant entries, ensuring each row represented a unique combination of key columns such as Country, Region, Date, and Age. Additionally, logical consistency checks were conducted to identify and correct any issues, such as negative values in Cases, Deaths, or Tests, or where Deaths exceeded Cases, which are not plausible in real-world data. Inconsistent or invalid values in categorical variables, such as Region or Sex, were also detected and addressed to match predefined valid categories. Finally, summary statistics and visualizations were generated to examine the distribution of numerical variables and detect any anomalies that could indicate data entry errors.

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. Section A.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.

Region	Sex	Mean Age	Total Cases	Total Deaths	Average Cases	Average Deaths	CFR (%)
Urban	m	50.12	6630777	160191	10138.80	244.94	2.42
Urban	f	50.02	6769033	165648	10366.05	253.67	2.45
Rural	m	49.83	6869820	163430	10504.31	249.89	2.38
Rural	f	50.05	6763540	156814	10373.53	240.51	2.32

Figure 1: Summary Statistics of COVID-19 Cases in Canada. The table shows the averages and totals of key numerical variables including age, cases, tests, and deaths. It also calculates the case fatality rate (CFR) as number of deaths divided by number of cases.

Figure 1 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.

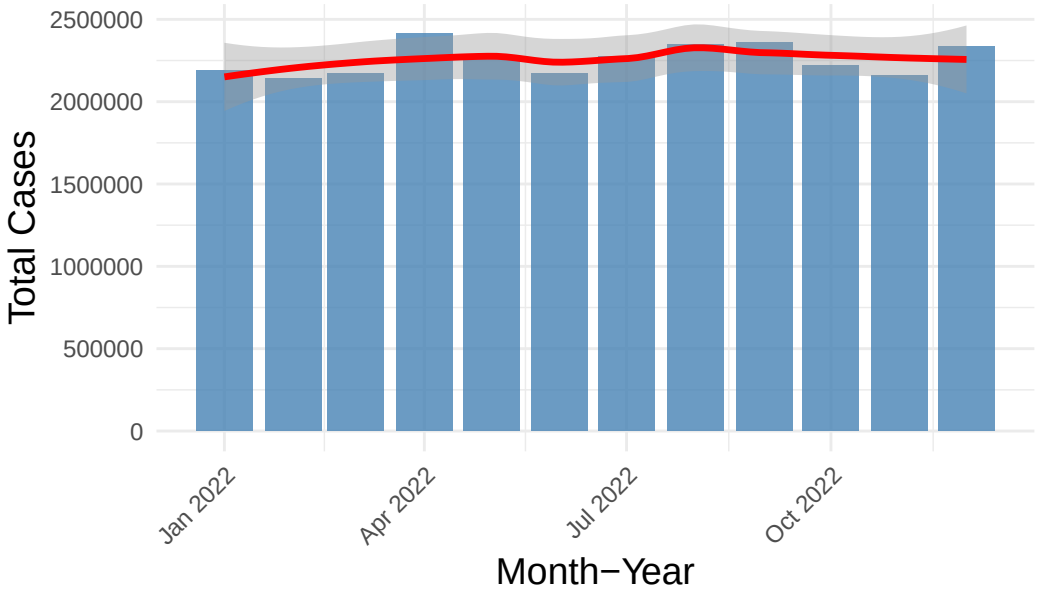

Figure 2: Number of COVID-19 cases per month in 2022 in Canada. The red trend line highlights the pattern of COVID-19 cases by month in Canada throughout 2022.

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 2, 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.

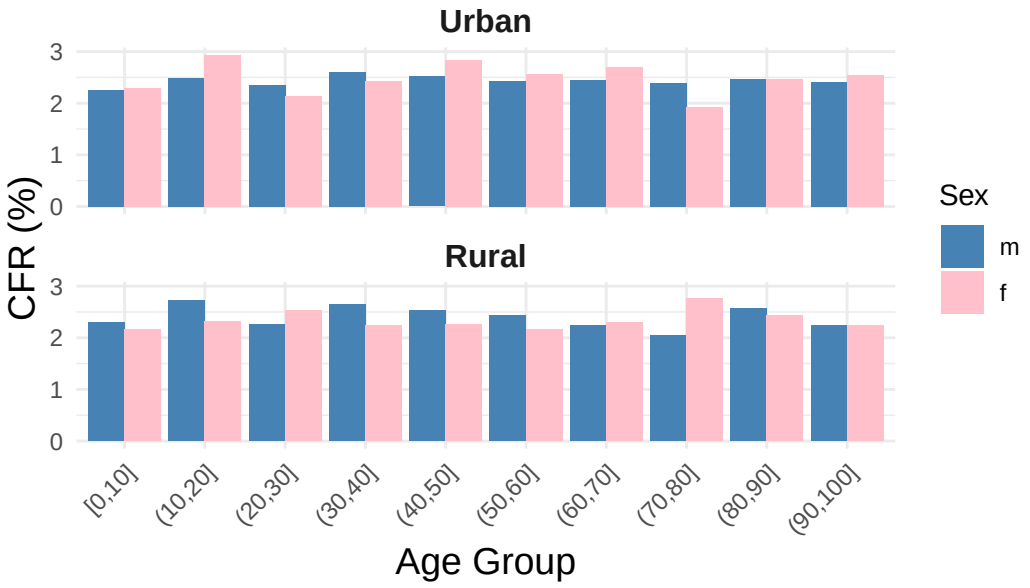

Figure 3: Case Fatality Rate (CFR) by Age Group, Region, and Sex. Urban and rural regions are separated into two subfigures, each characterizing male and females by color and displaying differences by age intervals.

2.4.4 Age, Sex, and Region

Age, sex, and region are key predictor variables that are examined in this analysis. In Figure 3, 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 3 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 4 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 1.

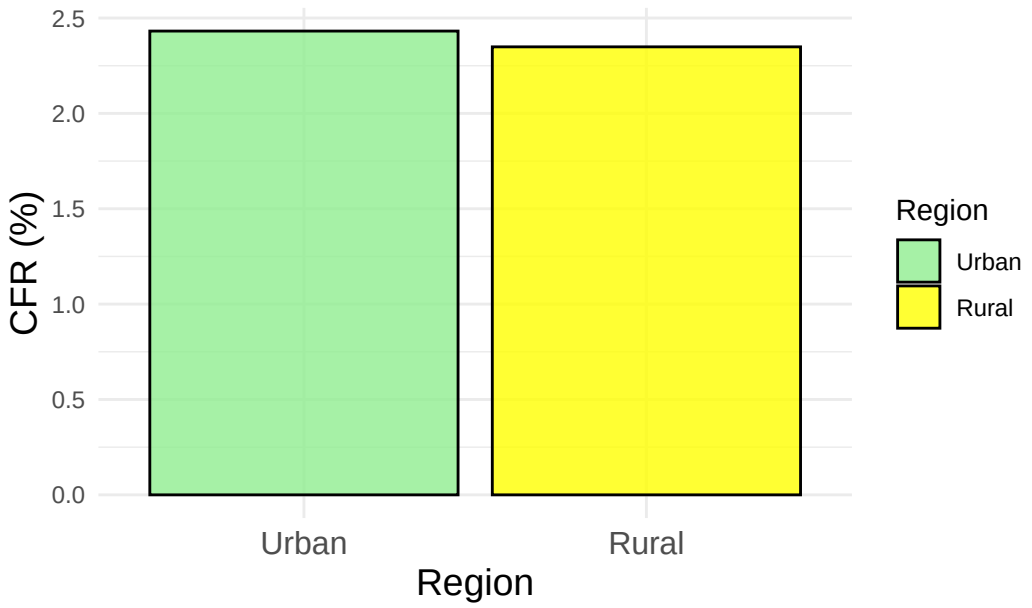

Figure 4: Case Fatality Rate (CFR) by Region in 2022 in Canada. CFR percentage is shown to be slightly higher in urban areas than rural areas.

3 Model

The primary goal of the model is to act as a tool for prediction for the level of high risk based on the characteristics of the specific datapoint. It can be used to identify how at risk a specific region, sex, or age group is for COVID-19 fatality. The mathematical definition of high risk is included in Section 3.1 alongside a mathematical definition of CFR. Overall, the CFR is the number of deaths per cases while the high risk is when that level exceeds the median across the entire population.

The model described here is a logistic regression model. This was selected based on the binary and categorical nature of many predictor and outcome variables. Justification of this choice is included in Section 3.2 and Section 3.4, with more details included in the appendix (Section A.5).

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:

$$\text{logit}(p) = \ln\left(\frac{p}{1-p}\right) = \beta_0 + \beta_1 \cdot \text{Tests} + \beta_2 \cdot \text{Age} + \beta_3 \cdot \text{Region} + \epsilon \tag{2}$$

where:

- p: Predicted probability of being high risk (p = Predicted Probability).
- $\text{logit}(p)$: 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 'pROC' (Robin et al. 2011) and 'caret (Kuhn 2023) packages to construct and test the model.

3.2 Model justification

For the analysis, a logistic regression model was chosen to predict the probability of a region or individual being classified as "high risk" based on key predictors such as age, sex, region, and the number of COVID-19 tests conducted. Logistic regression is particularly well-suited for this dataset because the outcome variable, high risk, is binary (e.g., classified as high risk or not). This makes logistic regression the most appropriate modeling technique as it directly estimates the probability of the binary outcome through the log-odds transformation. Compared to other models like linear regression, logistic regression avoids the issue of nonsensical predictions, such as probabilities below 0 or above 1, ensuring the interpretability of the results. Most importantly, logistic regression can handle both categorical and continuous predictors effectively. For this particular dataset, the key predictor variables are all categorical, even age is categorized by intervals. Therefore, it was extremely necessary for the model to be able to handl categorical data, which the logistic regression model is capable of.

Another key advantage of the logistic regression model in this context is that it provides interpretable coefficients, which allow for an understanding of how specific predictors influence the odds of being high risk. This interpretability is crucial for public health applications. Furthermore, logistic regression is computationally efficient, making it ideal for datasets with relatively straightforward structures like this one. Thus, the logistic regression model strikes an appropriate balance between accuracy, interpretability, and simplicity, making it the most suitable choice for this analysis.

The details on model selection can be found in the appendix in Section A.4

3.3 Assumptions and Limitations

The logistic regression model assumes that observations within the dataset are independent, meaning that each data point does not influence others. However, this assumption could be

violated if individuals within the same region share similar environmental, healthcare, or socioeconomic factors that contribute to their risk classification. Additionally, the model assumes that the predictors, such as age, sex, and region, are not highly correlated. Significant multicollinearity between these variables could undermine the stability of the model coefficients and reduce interpretability. Finally, the classification of "high risk" as regions or individuals exceeding the median death rate simplifies a complex phenomenon. This threshold may fail to capture subtler variations in risk factors across different demographic and geographic contexts.

One key limitation of logistic regression is its inability to account for interactions between predictors without explicitly adding interaction terms. For instance, the combined effects of age and sex on risk may not be fully captured in this model. Additionally, the model assumes a predefined functional form, which limits its flexibility in identifying nonlinear relationships between predictors and outcomes. As a result, more complex patterns in the data may remain undetected, suggesting the need for alternative modeling approaches such as random forests or decision trees in future studies.

3.4 Model Validation

For the model validation process of the selected logistic regression model, multiple validation techniques were used. Firstly, a ROC curve was constructed to visualize the trade-off between sensitivity and specificity. The ROC curve showed a high curve that deviated from the linear line representing random chance between sensitivity and specificity. The AUC value accompanying the ROC curve was above 0.8 for all three prediction models constructed for different variables.

A confusion matrix was also constructed to assess classification accuracy, precision, and recall by comparing the predicted classes to the actual outcomes. Values of above 0.75 were yielded for the accuracy rating. Lastly, the odds ratios with confidence intervals provided information on the strength and direction of the predictors' effects.

In general, these results demonstrated the suitability and strong effect that the logistic regression model has to predict probability of high risk based on the data.

Details of the model validation are provided in the appendix (Section A.5).

4 Results

The results of this analysis are summarized in Figure 5, Figure 6, and Figure 7 to give an all-rounded picture of the predicted probability of high risk based on the different predictor variables.

4.1 Predicted Probability of High Risk by Region

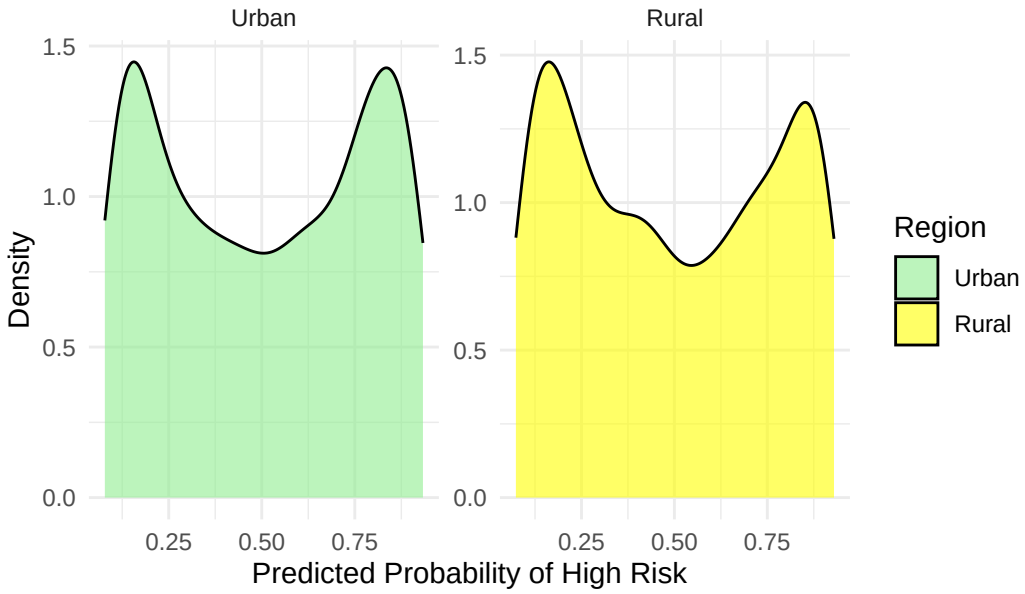

Figure 5: Predicted probability of high risk of COVID-19 by region (Urban vs. Rural). This is a density plot showing that the extremes of predicted probability of high risk are more frequent for both urban and rural regions.

Figure 5 models the predicted probability of high risk for different regions, urban or rural. This is represented by a faceted density plot.

Based on these results, the predicted probability of high risk is higher in rural areas than urban areas and the predicted probability of low risk is higher in urban areas than rural. This is indicated by the height of the peaks at the higher and lower end of the scale, where the peaks for urban regions are taller at the lower end of the probability of high risk scale and shorter at the higher end. The height of the peaks indicate the density of datapoints that fall around the respective predicted value.

4.2 Predicted Probability of High Risk by Sex

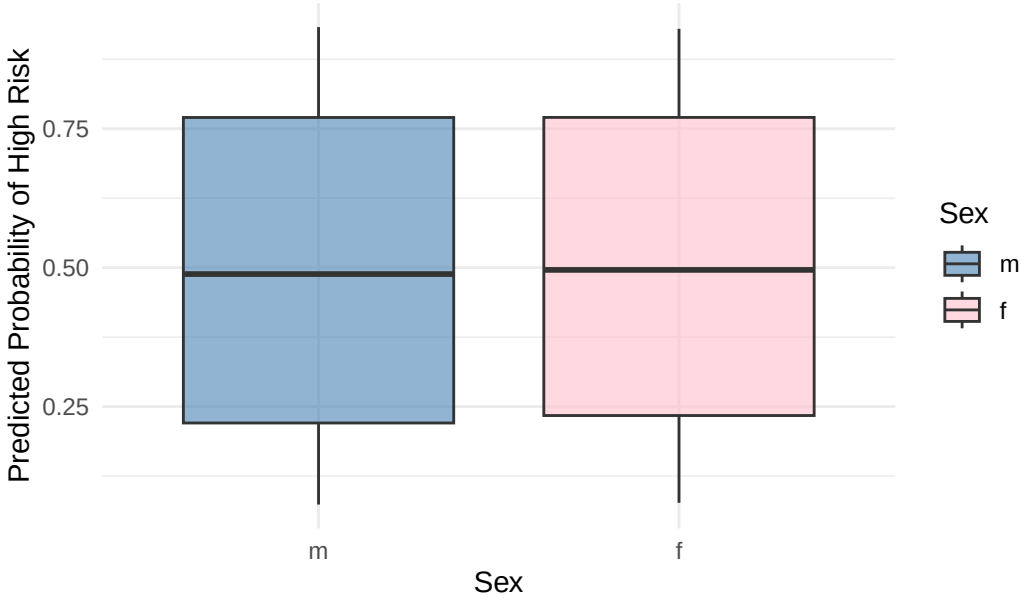

Figure 6: Predicted probability of high risk based on sex. The box-and-whiskers plots show that the median value of the predicted probability of high risk is similar for males and females. There is a large IQR for both.

Figure 6 compares the predicted probability of high risk based on biological sex, representing the model in a similar box-and-whiskers plot method. As shown in the figure, males (represented by 'm') have a noticeably higher predicted probability of high risk than females (represented by 'f'). The predicted probability of a male qualifying as high risk for COVID-19 is above 50%, while the number is below 50% for females.

4.3 Predicted Probability of High Risk by Age Interval

Figure 7 shows the predicted probability of high risk for different age groups. The age groups are intervals of 10 from 0 to 100 as classified by the original dataset. As shown in the figure, age intervals 10 to 20 and 30 to 40 have a significantly higher predicted probability of high risk than other age groups. This means that it is more likely (around 55% and 57% chance) that someone within these age ranges will qualify as a high risk than the other age groups. The age group with the lowest predicted probability of high risk is 0 to 10 year-olds, yielding a 38% chance median of being predicted as high risk.

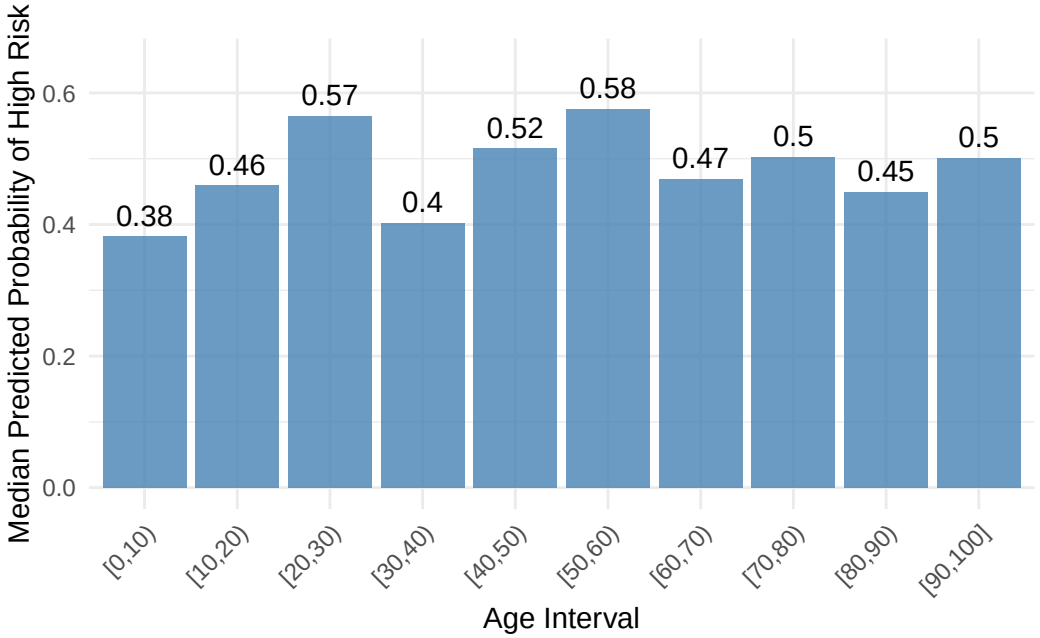

Figure 7: Predicted probability of high risk based on age intervals. The median predicted probability of high risk is shown as a decimal, where age intervals 20-30 and 50-60 have relatively higher predicted probabilities of high risk.

5 Discussion

5.1 Interpreting Results

The results of this analysis displays information into how demographic and regional factors influence the likelihood of being classified as high risk for COVID-19. The logistic regression model, designed to predict the probability of high risk, effectively highlighted the relative importance of age, sex, and region in shaping these probabilities.

5.1.1 Predicted Probability of High Risk by Region

Based on Figure 5, there is a higher density of individuals predicted to have high risk in the urban regions than in the rural regions. This disparity underscores the potential challenges faced by urban regions, such as overcrowding and not enough healthcare, which may contribute to higher case fatality rates (CFR). Since patients of COVID-19 were encouraged to stay at home even after testing positive at a hospital, patients who lived in urban regions may not have been able to access quick enough healthcare if their symptoms were to worsen rapidly. This could be reflected in the higher predicted risks for these communities. Overall, the predicted probability of high risk between the two regions were not much different, potentially reflecting that both regions experience certain challenges in healthcare and are similar in the amount and timeliness of healthcare received.

Another aspect of the figures that is worth pointing out, if not a more important aspect, is the overall shape. There are significantly higher densities at the extremes of the predicted probability of high risk than moderate levels for both urban and rural regions. A potential explanation of this is that individuals have high disparities in access to healthcare regardless of the region they live in. In other words, people either had really good, timely healthcare or really bad healthcare in both urban and rural regions. This indicates that the region an individual lived in may not have been the strongest predictor the high risk. More research and analysis are needed to identify the underlying variables that correlate with this pattern.

5.1.2 Predicted Probability of High Risk by Sex

The analysis revealed that there are no significant differences in predicted probability of high risk in males and females. The size of the difference is extremely small, especially compared to the IQR highlighted in Figure 6. The potential differences can be attributed to biological differences, comorbidities, and health-seeking behaviors. There may be a slight difference between sexes in the sensitivity to one's own health conditions, willingness to seek healthcare, and other aspects. However, the analysis shows that the high risk is essentially equal for the two sexes. There is also an extremely large distribution of predicted probabilities within sexes

as shown by the IQR. Therefore, more research and analysis will be needed to conclude that there are or are no sexual differences in the predicted probability of high risk.

5.1.3 Predicted Probability of High Risk by Age

Age also played a pivotal role, with certain age intervals, such as 20-30 and 50-60, showing significantly higher probabilities of high risk. There is no clear trend where higher or lower age groups have increased or decreased predicted risk. This pattern may reflect complex interactions between behavioral, immunological, and societal factors influencing COVID-19 outcomes. For example, 20-30 year-olds may be more engaged in social activities and more constantly in the workplace. They may also be less likely to have started families with children and be less careful with their health and hygiene. 50-60 year-olds may still be out in the work field with relatively high social interactions while also have reduced or decreasing immunity to pathogens. Understanding these relationships is important for tailoring public health interventions and resource allocation strategies. Further research with more targeted variables and measures such as frequency of social activity and overall health conditions may be necessary to find key relationships between age and probability of high risk.

5.2 Interpreting Predicted Probability of High Risk for COVID-19

Based on the model set-up Section 3.1, the probability of high risk is the predicted probability that the calculated level of high risk based on demographic information of a specific datapoint is above the median level of high risk for the dataset. In terms of the real-world, this is the chance that someone is more likely to die from COVID-19 compared to the rest of the population.

5.3 High Risk Mechanisms in Public Health Research

The analysis of high risk is commonly used in public health settings, especially in epidemiology. The high risk approach allows the analysis to better identify the groups that are at higher risk or more likely to yield a negative outcome from the disease (Platt, Keyes, and Galea 2016). In the case of this analysis, the high risk groups that are identified are those who are more likely to find COVID-19 fatal rather than mild.

The results generally show moderate to weak correlations and relationships between the different demographic characteristics and the predicted probability of high risk. This is common depending on the topic of focus. Since COVID-19 is being examined, this situation is considered normal since COVID-19 does not have obvious strong predictors (Gallo Marin et al. 2020).

Returning to the discussion of the use of the high risk approach in public health research, it is an increasingly popular approach due to the strong focus it has on individuals and communities

that are more at risk (Platt, Keyes, and Galea 2016). This can direct actions and prevention to those who are most in need and can be especially important when the disease is fatal, such as COVID-19. Although this analysis shows little difference between different demographic characteristics, the approach can still have an effect when large populations are considered. As for other cases where there are clear differences in risk, for example, using age as a predictor of cancer and implementing closer monitoring for older individuals, may lead to more lives saved (Bastian et al. 2012).

5.4 Key Future Implications

A key takeaway from this study is the intersection between demographic factors and social determinants of health. There are specific demographic groups that may have well-documented disadvantages in accessing healthcare, which may be reflected as higher risk. Addressing these social determinants, such as income inequality, education, and healthcare infrastructure, could have long-term benefits for public health beyond the current pandemic. Interventions should not only focus on immediate health needs but also work toward alleviating these systemic disparities.

By identifying demographic and regional factors associated with higher probabilities of high risk, policymakers can develop targeted interventions to mitigate disparities. For example, enhancing healthcare infrastructure in rural areas and prioritizing vaccination and testing campaigns for high-risk groups can help reduce the impact of future pandemics.

Additionally, the use of logistic regression models to predict high-risk classifications provides a robust framework for decision-making. This approach could be extended to other diseases or health outcomes, offering a scalable and interpretable method for public health planning. The analysis also highlights the importance of maintaining high-quality, standardized datasets to facilitate cross-regional and cross-demographic comparisons.

5.5 Limitations

Despite its strengths, this study has several limitations that should be addressed in future research. First, the analysis is limited to Canada, which may restrict the generalizability of the findings to other countries with different healthcare systems, cultural contexts, or pandemic responses. Second, the dataset used for this analysis only includes data from 2022, which may not capture temporal variations in COVID-19 patterns. Future studies could incorporate longitudinal data to examine how the predictors of high risk evolve over time.

Another limitation is the reliance on aggregate data, which may obscure individual-level variability and interactions. For example, the interplay between age, sex, and comorbidities at the individual level could yield more nuanced insights. Additionally, the classification of rural and urban regions may oversimplify the diversity within these categories, such as the distinction between suburban and remote rural areas.

The patterns identified do not show strong correlations. The relationships between the key predictors (age, sex, and region) may not fully capture the complexity of the underlying risk factors. There could be additional variables that influence risk but were not included in this analysis. Furthermore, the correlations between these predictors and the high-risk classification may be influenced by other unmeasured or unobserved factors. These aspects of the dataset require further exploration in future research to better understand the full range of factors that contribute to COVID-19 risk and to refine the models used for risk classification.

Finally, the data obtained for COVID-19 is obviously in the past and no longer an issue that is of primary concern. This issue is in line with the generalizability issue due to limited surveying countries. The patterns and models may be limited in generalizability due to the sole focus on COVID-19. The data itself also has limitations of not fully representing the population due to data collection methods. Only groups of individuals who report to hospitals or governments were sampled, leading to potential biases.

5.6 Next Steps

Building on the findings of this study, future research should explore several directions. First, expanding the analysis to include additional countries or regions would provide a better understanding of the factors influencing COVID-19 outcomes. Second, incorporating additional variables, such as socioeconomic status, vaccination rates, and comorbidities, could enhance the predictive power of the models.

Another next step that could be taken is to use the analysis and modeling on a broader series of pandemics and diseases such as seasonal flues, SARS outbreak in the 2000s, and others that may show similar trends. This will allow a broader generalizability of the demographic patterns and allow the model to better predict probability of high risk based on demographic data. Finally, public health practitioners and researchers should collaborate to ensure that findings are translated into actionable policies and interventions.

A Appendix

A.1 Appendix A: Raw Data

Country	Region	Date	Sex	Age	Cases	Deaths	Tests	CFR
Canada	Urban	2022-01-01	m	0	4091	215	44602	5.255439
Canada	Rural	2022-01-01	m	0	14555	269	42100	1.848162
Canada	Urban	2022-01-01	m	0	4981	407	39566	8.171050
Canada	Rural	2022-01-01	m	0	2960	495	45319	16.722973
Canada	Urban	2022-01-01	m	0	18488	394	46584	2.131112
Canada	Rural	2022-01-01	m	0	10762	391	11835	3.633154

Figure 8: First six rows of the dataset after cleaning. Shows columns including key variables including region, date, sex, age, cases, deaths, and CFR.

A.2 Appendix B: Survey and Sampling Details

The dataset used in this analysis is sourced from the COVerAGE-DB, a repository of COVID-19 demographic data. The database compiles data from a diverse set of sources, including government health agencies, international organizations, and other reputable institutions, ensuring a broad representation of regions and demographics. Each entry in the dataset represents an aggregation of cases, deaths, and demographic information categorized by age, sex, and region.

A.2.1 Collection Process

Data collection follows a systematic approach. For primary sources, COVerAGE-DB retrieves data directly from official government dashboards, public health reports, and institutional publications (Dowd et al. 2020). This ensures that the dataset reflects real-world phenomena with minimal distortion.

The database ensures standardization by processing the raw data from various sources through rigorous standardization to ensure comparability. For example, age intervals are harmonized across sources to create a unified structure suitable for cross-regional analysis.

The database also ensures quality Control through running automated checks and manual reviews to detect inconsistencies, missing values, and potential errors in data entry. Such measures are crucial given the heterogeneity in reporting standards across countries and regions.

Lastly, the update frequency is ensured so the data is as relevant and recent as possible. The database is updated regularly to reflect new data releases, ensuring timeliness and relevance for ongoing research.

A.2.2 Sampling and Representativeness

While the dataset covers a broad range of countries and regions, its representativeness depends on the completeness and accuracy of the underlying sources. For instance, regions with underdeveloped health reporting infrastructure might have lower data quality or incomplete records. Additionally, variations in testing capacity and reporting standards can introduce biases, particularly in the calculation of derived measures such as case fatality rates (CFR).

A.2.3 Observational Nature and Limitations

The data is observational, collected without experimental controls. As a result, confounding factors such as healthcare system differences, socioeconomic disparities, and public health policies must be carefully considered when interpreting results (Ioannidis 2021). For instance, differences in testing rates across regions can influence the apparent number of cases and consequently impact measures like CFR.

A.3 Appendix C: Idealized Methodology

The data collection method currently used in the dataset has some key limitations of not being able to reach the entire population. The idealized methodology will address these issues by implementing a new way of data collection where, if done in 2022, will provide a more representative dataset of COVID-19 data and information.

The idealized methodology would have in-person PCR testing for the sample to gather data on the number of tests, number of cases, age, sex, and region. The number of deaths can be difficult to survey in-person without the support of government agencies. Therefore, the data for number of deaths can still be collected from relevant government institutions, as Canadians are required to report deaths to the government. This combined methodology will allow a more accurate dataset of COVID-19 and demographics information.

A.3.1 Population

The population, given that the goal of the analysis is to generalize to the Canadian population, would be people living in Canada permanently. This includes all Canadian citizens, permanent residence, and any individuals with a Visa that extends for more than one year (2022).

A.3.2 Frame

In the case of this methodology, when there are no financial limitations in play, the frame would be the same as the population. Since there are no restrictions on what households or Canadian citizens can't be sampled, the frame is essentially a subset of the population that is the entire population.

A.3.3 Sample and Sampling Method

The sample would be a subset from the population. The idealized method must ensure that the sample consists of a representative group of individuals. The method to be used will be systematic sampling, where people would be hired to visit 1 of every 10 houses or apartments (or other units of living) and sample everyone within that household. This will avoid any sampling biases such as only sampling people who go out (if sampled on the streets) or only sampling people who drive (if sampling on roads).

For every household sampled, the surveyor will conduct a PCR test for each person in the household, record each test, record the number of tests that return as positive as the number of cases, and gather demographic information of each person in the household such as such as sex and age. Region of the household should already be known and should also be recorded as part of the dataset.

This should result in a dataset with a sufficient sample size. Given that there is an estimated 15.3 million households in 2022 (GlobalData 2021), the sample will consist of 1,530,000 households. There are an estimated 3 people per household (ArcGIS n.d.), leaving a sample of around 4.59 million people.

The systematic sampling method ensures that a sufficient sample size will be gathered and allows rather accurate prediction of the number of datapoints within the sample. It also ensures for uniform coverage and that sampling wouldn't be biased towards certain communities or subgroups. However, limitations could surface if the selected samples are not responding, not available, or refuse to co-operate. The systematic sampling method makes it hard to change the sample size if sitations like this were to occur.

The sampling procedure will be repeated monthly until the end of the year to gather time-related data.

A.3.4 Handling Non-response

Non-response could mean not being at home during the visit of the surveyor or not willing to participate. The initial response to this could be to re-visit another time and attempt to persuade the participant to co-operate.

If a small number of households refuse to participate after repeated visits, the broad demographic information (such as region) can be recorded and later adjusted for during the data anlysis process. However, if a large number of households refuse to participate, specific parts of the data may need to be re-sampled or dropped, which may cause biases in the analysis. An attempt to address this may be to adjust the dataset based on known demographic information, but certain biases may be unavoidable.

A.3.5 Idealized Procedure

Consent: The surveyor will read this to the participant of the test before conducting the tests and ask for their signature.

"Thank you for your will in potentially co-operating in this COVID-19 study. By participating in this study, you agree to provide demographic information and respond to questions related to COVID-19. The procedure will require conducting PCR test for you and asking you basic questions regarding your demographic characteristics. Your participation is voluntary, and all data will be kept confidential and anonymous and used solely for research purposes. You may withdraw from the study at any time without penalty. If you have any questions about the study, please feel free to ask. If you acknowledge your rights and are willing to consent to this study, please sign."

PCR: The surveyor will conduct a PCR test for individuals who have consented to the study.

COVID-19 Information: The surveyor will record the number of PCR tests conducted in the household and the number of positives.

Demographic Questions: The surveyor will ask the following questions or record the known answer.

- What is your biological sex at birth?
- What is your age?
- What area do you live in? Urban or rural?

Debrief: The surveyor will read the following to the participants after conducting the test and asking the questions.

"Thank you for participating in this study. The information you provided on demographic factors and the test done for COVID-19 positivity will help us better understand the impact of the pandemic across different demographic groups. Your responses will contribute to valuable research aimed at improving public health strategies. If you have any questions or would like further information about the study, please feel free to contact us at"brucejc.zhang@mail.utoronto.ca"

Data Consolidation: The data collected by surveyors will be matched with the death-data from governments and hospitals at the end of the year of 2022 and will prepare the data for analysis for any patterns or trends as done in this analysis.

A.4 Appendix D: Model Selection Details

During the model selection process, model figures were generated for the linear regression model. As shown in Figure 9, the data was difficult to represent and relationships were difficult to identify due to the categorical nature of the variables. This led to attempts to model the data in different ways. The second method that was attempted was using a logistic regression model, which was able to represent the data well and make predictions about high risk. These figures are shown within the Section 4 of the paper.

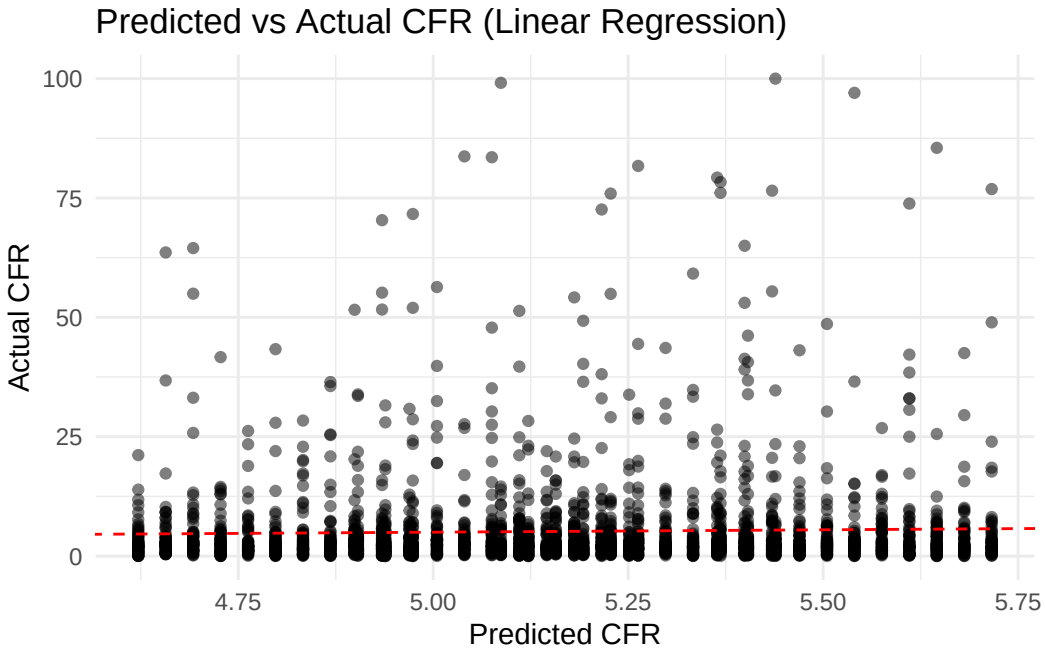

Figure 9: Linear Regression Model for dataset. No clear pattern can be observed. Weak predictive power is seen.

A.5 Appendix E: Model Validation Details

A.5.1 ROC Curve and AUC

```
AUC: 0.8202962
```

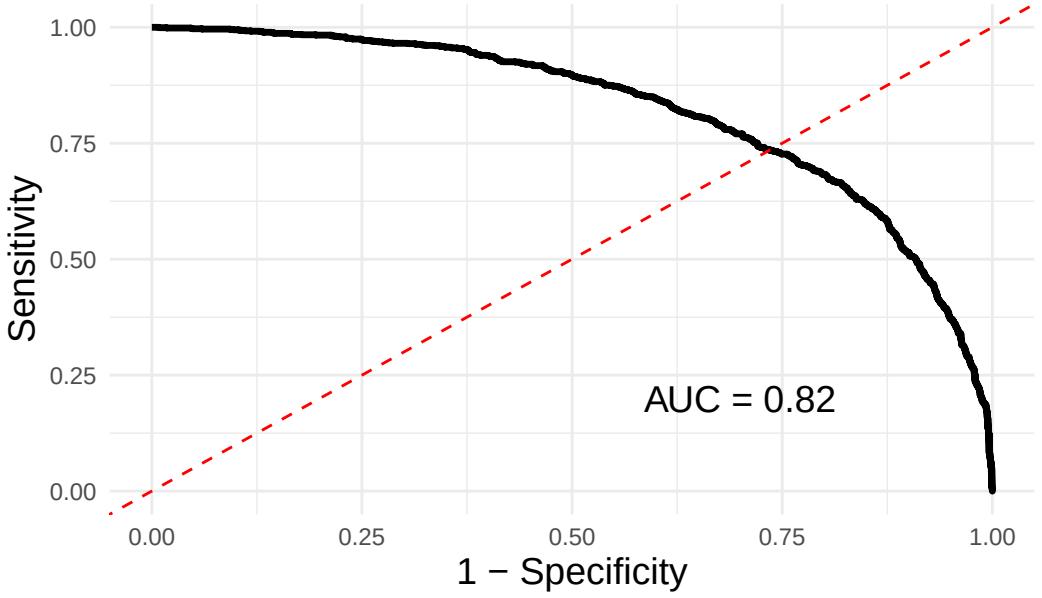

Figure 10: ROC Curve and AUC Value for Logistic Regression Model. The ROC curve and AUC values show that the logistic regression model yields a moderately high predictive performance for the dataset.

A.5.2 Confusion Matrix

```
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0 975 353
         1 332 953

               Accuracy : 0.7378
                 95% CI : (0.7205, 0.7546)
    No Information Rate : 0.5002
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.4757

 Mcnemar's Test P-Value : 0.4448

            Sensitivity : 0.7460
            Specificity : 0.7297
         Pos Pred Value : 0.7342
         Neg Pred Value : 0.7416
             Prevalence : 0.5002
         Detection Rate : 0.3731
   Detection Prevalence : 0.5082
      Balanced Accuracy : 0.7378

       'Positive' Class : 0
```

A.5.3 Odds Ratio with Confidence Intervals

Predictor	Odds Ratio	Standard Error	Statistic	P Value	Lower 95% CI	Upper G5% CI
(Intercept)	10.145	0.198	11.731	0.000	6.909	14.991
Sexf	1.026	0.094	0.273	0.785	0.853	1.235
RegionRural	0.992	0.094	-0.087	0.930	0.824	1.194
Age_Interval[10,20)	1.035	0.222	0.156	0.876	0.671	1.599
Age_Interval[20,30)	1.257	0.220	1.041	0.298	0.818	1.936
Age_Interval[30,40)	0.936	0.223	-0.297	0.766	0.604	1.449
Age_Interval[40,50)	1.320	0.222	1.248	0.212	0.854	2.042
Age_Interval[50,60)	1.614	0.222	2.155	0.031	1.045	2.499
Age_Interval[60,70)	1.261	0.222	1.043	0.297	0.816	1.950
Age_Interval[70,80)	1.380	0.222	1.454	0.146	0.894	2.134
Age_Interval[80,90)	1.332	0.224	1.277	0.202	0.858	2.068
Age_Interval[90,100]	1.358	0.192	1.596	0.110	0.933	1.979
Cases	1.000	0.000	-25.022	0.000	1.000	1.000

Figure 11: Odds Ratio table with Confidence Intervals

References

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

Apache Software Foundation. 2024. *Apache Arrow: R Package*. https://arrow.apache.org/docs/r/.

ArcGIS. n.d. "Average Census Family Size in Canada." https://www.arcgis.com/home/item.html?id=7060d31cc95f4dd3931111e35a1ba2eb#:~:text=Description-,This%20layer%20shows%20the%20average%20census%20family%20size%20in%20Canada,family%20size%20is%203%20people.

Bastian, Patrick J., Stephen A. Boorjian, Alberto Bossi, Alberto Briganti, Axel Heidenreich, Stephen J. Freedland, Francesco Montorsi, et al. 2012. "High-Risk Prostate Cancer: From Definition to Contemporary Management." *European Urology* 61 (6): 1096–106. https://doi.org/10.1016/j.eururo.2012.02.031.

Beaglehole, Robert, and Ruth Bonita. 2010. *Global Public Health: A New Era*. Oxford, UK: Oxford University Press. https://academic.oup.com/book/10351.

Dowd, Jennifer Beam, Leonardo Andriano, David M Brazel, Valentina Rotondi, Per Block, Xiaoyan Ding, Yang C Liu, and Melinda C Mills. 2020. "Demographic Science Aids in Understanding the Spread and Fatality Rates of COVID-19." *Proceedings of the National Academy of Sciences* 117 (18): 9696–98. https://doi.org/10.1073/pnas.2004911117.

Gallo Marin, Benjamin, Ghazal Aghagoli, Katya Lavine, Lanbo Yang, Emily J. Siff, Silvia S. Chiang, Thais P. Salazar-Mather, et al. 2020. "Predictors of COVID-19 Severity: A Literature Review." *Reviews in Medical Virology* 31 (1): 1–10. https://doi.org/10.1002/rmv.2146.

GlobalData. 2021. "Number of Households in Canada." https://www.globaldata.com/data-insights/macroeconomic/number-of-households-in-canada-2096147/#:~:text=Total%20Households%20in%20Canada%20(2010%20–%202021%2C,of%20households%20in%20Canada%20increased%20by%2016.1.

Hollingshaus, Mike, and Emily Harris. 2024. *The Importance of Demographic Data for Understanding the COVID-19 Pandemic*. Information Age Publishing. https://books.google.ca/books?hl=en&lr=&id=TRzeEAAAQBAJ&oi=fnd&pg=PA15&dq=importance+of+demographics+in+covid&ots=dEAYcv8CvK&sig=VWQmteBzdk6d7n8bH8QWwSNF1E4#v=onepage&q&f=false.

Ioannidis, John PA. 2021. "Infection Fatality Rate of COVID-19 Inferred from Seroprevalence Data." *Bulletin of the World Health Organization* 99 (1): 19. https://doi.org/10.2471/BLT.20.265892.

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.

Platt, Johnathan, Katherine Keyes, and Sandro Galea. 2016. "Efficiency or Equity? Simulating the Impact of High-Risk and Population Intervention Strategies for the Prevention of

Disease." *SSM - Public Health* 3: 1–8. https://doi.org/10.1016/j.ssmph.2016.11.002.

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

Ram, Karthik et al. 2020. *Here: A Simple Way to Find Your Files*. https://cran.r-project.org/web/packages/here/here.pdf.

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. https://doi.org/10.2174/0126667975257267231213071226.

Wei, Taiyun, Zheng Li, and Xinyu Zhang. 2020. *Corrplot: Visualization of a Correlation Matrix*. https://cran.r-project.org/web/packages/corrplot/corrplot.pdf.

Wei, Yuanyuan, Jianpeng Ma, and Jie Liu. 2020. *DataExplorer: An r Package for Exploratory Data Analysis*. https://cran.r-project.org/web/packages/DataExplorer/DataExplorer.pdf.

Wickham, Hadley et al. 2020. *Httr: Tools for Working with URLs and HTTP*. https://cran.r-project.org/web/packages/httr/httr.pdf.

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. https://doi.org/10.1016/j.amjcard.2024.04.018.