Exploring Socio-Demographic Factors Influencing Health Perceptions in Canada*

The study exposes significant influence of age group and geography on health.

Qizhou Xie

December 2, 2024

The strategic targeting of health perception variables is essential for understanding public health outcomes across different demographics.This study investigates the socio-demographic factors influencing health perceptions in Canada by employing Bayesian ordered linear regression. By analyzing key variables—such as age group, gender, year, and geography—this research identifies trends in health percentage estimates over time.The analysis emphasizes how these factors contributed to the prioritization of health perceptions among different demographics, focusing on the disparities observed between age groups and genders.

Table of contents

*Code and data are available at: https://github.com/tx77777/health_survey.git.

1 Introduction

Understanding the socio-demographic factors that influence health perceptions is essential for improving public health outcomes in Canada, as health disparities continue to challenge policymakers (Wilkinson and Pickett 2009; Marmot 2005; Raphael 2016). This study aims to analyze health perceptions across different demographic groups—including age, gender, and geography—using data collected from health surveys across the country.Previous research has identified social determinants of health as key factors contributing to disparities (Marmot

2005; Wilkinson and Pickett 2009), but a focused analysis of how these determinants interact specifically within the Canadian context, incorporating demographic layers, has been lacking. This study addresses that cleft by investigating the interactions among geography, age, gender, and temporal trends in health perceptions across Canada (Raphael 2016).

The analysis was conducted using a Bayesian ordered linear regression model, with key variables such as Geography, Reference_Date, Age_Group, and Sex included to uncover trends and disparities. By employing this approach, the study captures the multi-dimensional aspects of health perception across the population.

The findings revealed significant variation in health perceptions, with younger populations and residents of regions like Ontario and British Columbia reporting better health, while older individuals and residents of other regions like the Territories showed lower health values. The study also found gender differences in perceived health outcomes.

These findings are essential for developing targeted public health policies and initiatives that address the specific needs of different demographic segments. The understanding could help guide resource allocation and intervention design to reduce health disparities and improve outcomes.

The paper is structured as follows: Section Section 1 provides an introduction to the context of health disparities. Section Section 2 introduces the data and visualization methods used, while Section Section 3 describes the Bayesian modeling approach. Section Section 4 presents the key findings, and Section Section 5 explores the implications of these findings for public health policy in Canada. Finally, Section Section B Contains a questionnaire on how to collect data.

1.1 Estimand

In this study, we estimate the average causal effect of socio-demographic factors on health percentage values, focusing on understanding how these factors influence health perceptions in Canada. The key variables include age group, sex, year, and geography, which serve as predictors, and health percentage values, which represent the outcome of interest. The core objective is to determine the causal relationship between socio-demographic factors (age, sex, year, geography) and health percentage values, offering understanding into how changes in these variables are associated with changes in perceived health outcomes across different demographic groups. This study also aims to examine whether certain socio-demographic factors exert a stronger influence on health perceptions than others. By incorporating a Bayesian approach, the analysis allows for uncertainty quantification, providing a probabilistic view of the relationships. Additionally, the investigation includes exploring potential interactions between socio-demographic variables to understand their combined impact on health outcomes. The results from this analysis are intended to inform targeted public health interventions. Ultimately, this approach seeks to provide a data-directed foundation for reducing health disparities across Canadian populations.

2 Data

2.1 data source

The data for the 2024 Canadian Community Health Survey (Canada (2024)) is collected through electronic questionnaires (EQ) or via interviews conducted by Statistics Canada using CATI or CAPI methods. Respondents receive a mailed letter with a code to access the online questionnaire, where a household member is randomly selected to participate. If the questionnaire is not completed online, follow-up is conducted by phone, email, text, or in-person visit. The survey is available in English and French, and proxy reporting is allowed for certain questions. Collected data will be linked to respondents' tax records (T1, T1FF, T4) with their consent, and respondents can opt out of this linkage. The average survey duration is 40 minutes.

2.2 Measurement

The dataset used in this study was sourced from Statistics Canada, which records various health survey indicators documenting individuals' health perceptions across Canada. The original dataset consisted of approximately 327,166 records, each capturing a snapshot of health-related data points such as perceived health status, age, gender, and geographic location. These data points represent the responses collected through health surveys administered to individuals across different provinces, aiming to understand their subjective health perceptions over time.

The data underwent significant preprocessing to align with the specific focus of this research. The initial step involved filtering out records where the indicator was not percentage-based, reducing the dataset to 99,163 entries. These entries were carefully curated to focus on health perception metrics expressed as percentages, such as the proportion of individuals rating their health as excellent, good, or poor.

This measurement approach allows for the quantification of perceived health outcomes, turning survey responses into quantifiable variables that could be systematically analyzed. The transformation from individual survey responses to structured data points in the dataset reflects an effort to capture and model real-world health perceptions, facilitating the analysis of how socio-demographic factors influence these perceptions across different segments of the Canadian population.

The data for this study was systematically downloaded, cleaned, analyzed, modeled and visualized using R (R Core Team 2023), a extensive statistical programming language. The following packages were used for this study

- `tidyverse` (Wickham et al. 2021): To streamline the process of data manipulation and visualization.

- **ggplot2** (Wickham 2021): Used for its powerful and flexible capabilities in creating various types of visualizations tailored to the needs of this study.
- **dplyr** (Wickham, François, et al. 2021): Employed for its intuitive functions to transform and summarize the complex data sets effectively.
- **bayesplot** (Gelman, Gabry, et al. 2021): Utilized for creating graphical posterior predictive checks and diagnostic plots.
- **rstanarm** (Team 2021): Facilitated the implementation of Bayesian models, providing a straightforward way to fit regression models using Stan.
- **janitor** (Firke 2021): Making it simpler to handle the raw data by cleaning variable names and simplifying data structures.
- **arrow** (Apache Arrow 2021): Used for efficiently reading and writing large datasets, enhancing data handling capabilities.
- **knitr** (Xie 2021): Employed to dynamically generate reports which integrate R code with its outputs, allowing for seamless inclusion of plots and analysis results in the final document.
- ***Telling Stories with Data*** (Alexander 2023): This book was referenced for its code and methodologies in presenting data and statistical information.

```
# A tibble: 5,000 x 14
   Reference_Date Geography           Age_Group Sex   Indicators Characteristics
            <int> <chr>               <chr>     <chr> <chr>      <chr>
 1           2021 Canada (excluding ~ 50 to 64~ Males Fruit and~ High 95% confi~
 2           2022 New Brunswick       35 to 49~ Males Anxiety d~ Low 95% confid~
 3           2017 Saskatchewan        35 to 49~ Both~ Perceived~ Low 95% confid~
 4           2016 Prince Edward Isla~ 35 to 49~ Males Self-repo~ Percent
 5           2022 Nova Scotia         Total, 1~ Males Cannabis ~ Percent
 6           2016 Ontario             65 years~ Fema~ Current s~ Percent
 7           2022 Canada (excluding ~ Total, 1~ Males Perceived~ Low 95% confid~
 8           2022 Alberta             Total, 1~ Fema~ Cannabis ~ Percent
 9           2022 Quebec              50 to 64~ Males Current s~ High 95% confi~
10           2022 Newfoundland and L~ 50 to 64~ Males Perceived~ High 95% confi~
# i 4,990 more rows
# i 8 more variables: Unit_Of_Measure <chr>, Unit_ID <int>,
#   Scalar_Factor <chr>, Scalar_ID <int>, VECTOR <chr>, COORDINATE <chr>,
#   Value <dbl>, DECIMALS <int>
```

Table 1: Health Perception Data: A Preview of the Main Socio-Demographic Variables in Canada

Reference_Date	Age_Group	Sex	Value
2021	50 to 64 years	Males	19.6

Table 1: Health Perception Data: A Preview of the Main Socio-Demographic Variables in Canada

Reference_Date	Age_Group	Sex	Value
2022	35 to 49 years	Males	9.0
2017	35 to 49 years	Both sexes	19.0
2016	35 to 49 years	Males	67.8
2022	Total, 12 years and over	Males	34.1
2016	65 years and over	Females	5.5

Table 1 This table presents the first six rows from the cleansed dataset, focusing on health perceptions across different socio-demographic groups in Canada, including age, sex, and the year of observation.

2.3 Variable

Our analysis focuses on the following variables, with a specific focus on `Value` as the dependent variable:

- `Reference_Date`: The year of the observation, representing the time period when the health perception data was collected.
- `Age_Group`: The age category of the individuals surveyed, with the following possible values:
  - "12 years and over": Includes all individuals aged 12 and above.
  - "18 to 34 years": Young adult group.
  - "35 to 49 years": Middle-aged adults.
  - "50 to 64 years": Older adults approaching retirement.
  - "65 years and over": Senior citizens.
- `Sex`: The gender of the individuals, which includes:
  - "Both sexes": Data for both male and female individuals combined.
  - "Male": Data specifically for male individuals.
  - "Female": Data specifically for female individuals.

  `Geography`: The geographic region of the observation, representing different areas across Canada where health perception data was collected. Possible values include:
  - "Atlantic": Representing the Atlantic provinces.
  - "Quebec": Representing the province of Quebec.
  - "Ontario": Representing the province of Ontario.

  - "Prairies": Representing the Prairie provinces.
  - "British Columbia": Representing the province of British Columbia.
  - "Territories": Representing the Northern territories of Canada.

- `Value`: The health percentage value representing specific health indicators, serving as the dependent variable in our analysis. This variable captures the proportion of individuals reporting a particular health perception, such as "very good" or "excellent" health.

Detailed information about these variables and the data structure is presented in Table 1, which outlines the first few records from the processed dataset.

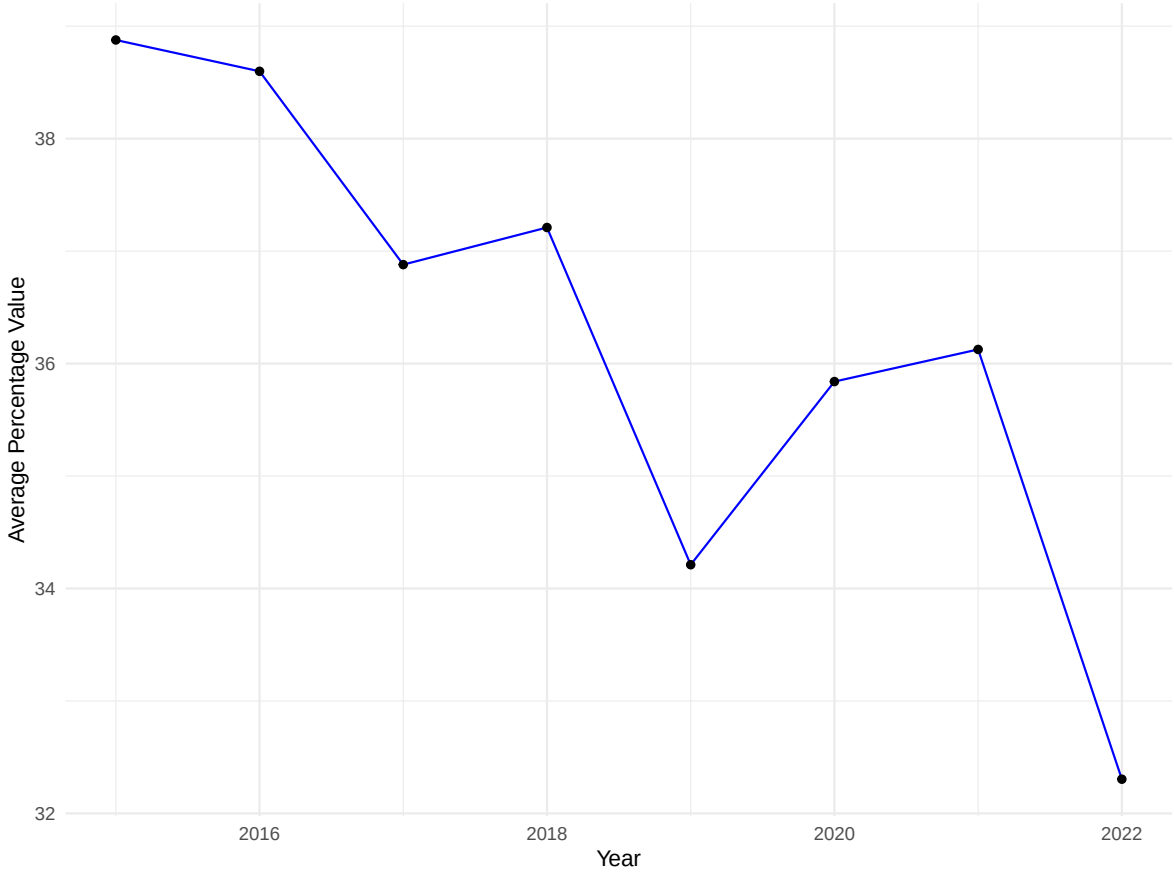


Figure 1: Average Value per Year

Figure 1 The visualization exposes significant trends in the average health percentage values across different years. Overall, there are noticeable variations in health perception levels, with some years showing higher values, which might reflect the effectiveness of public health interventions or improvements in socioeconomic conditions during those periods. The changes in average values may also be influenced by factors such as the age distribution and gender

composition of the survey sample. Further investigation into the social context and health policies during these years could help explain these observed trends.

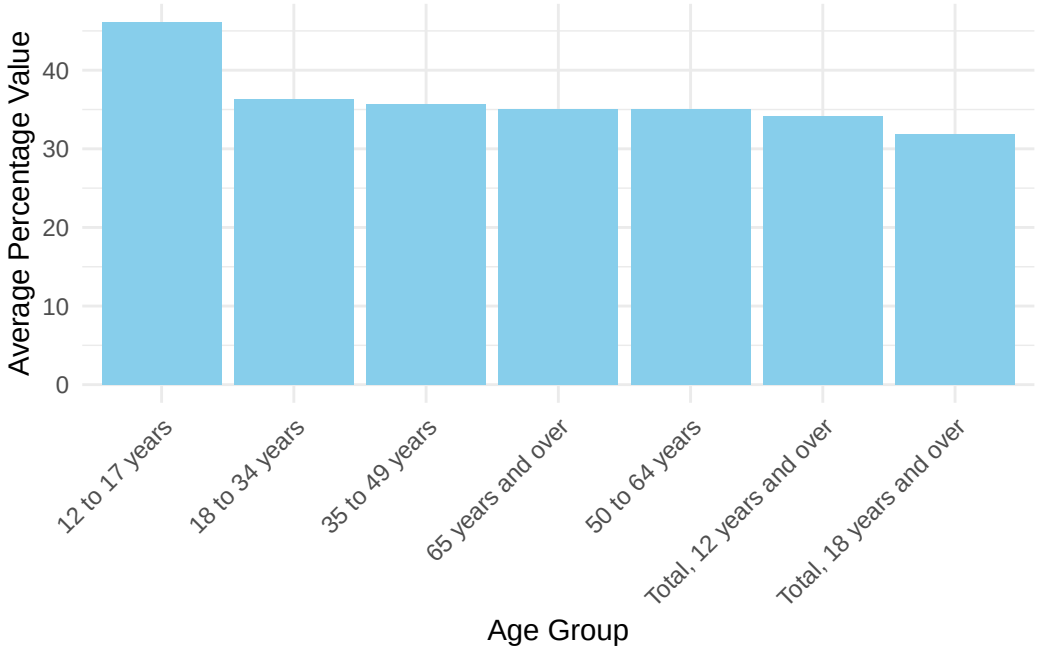


Figure 2: Distribution of different age_group

Figure 2 presents the "Average Value by Age Group," we can conclude that there are notable differences in health perception across age groups. Younger adults, particularly those in the "18 to 34 years" category, tend to report higher health percentage values, indicating better perceived health compared to older age groups. Conversely, seniors aged "65 years and over" show lower average health values, which may reflect age-related health challenges. These understanding highlight the importance of targeted health interventions that consider the specific needs of different age groups to improve overall health outcomes in the Canadian population.

Figure 3 presents the "Distribution of Value by Sex," The bar plot highlights the observable differences in health perception between males and females. The average values depicted indicate that there is a slight but noteworthy variation between the two genders. These differences in health perception, while not significantly pronounced, underscore that gender may still play a role in influencing health outcomes. This observation emphasizes the importance of integrating gender-specific health initiatives as part of broader public health strategies to address these differences effectively and improve overall health in Canada.

Figure 4 presents the "Average Value by Geography," we can conclude that health perceptions vary across different regions of Canada. The bar chart highlights that certain geographic

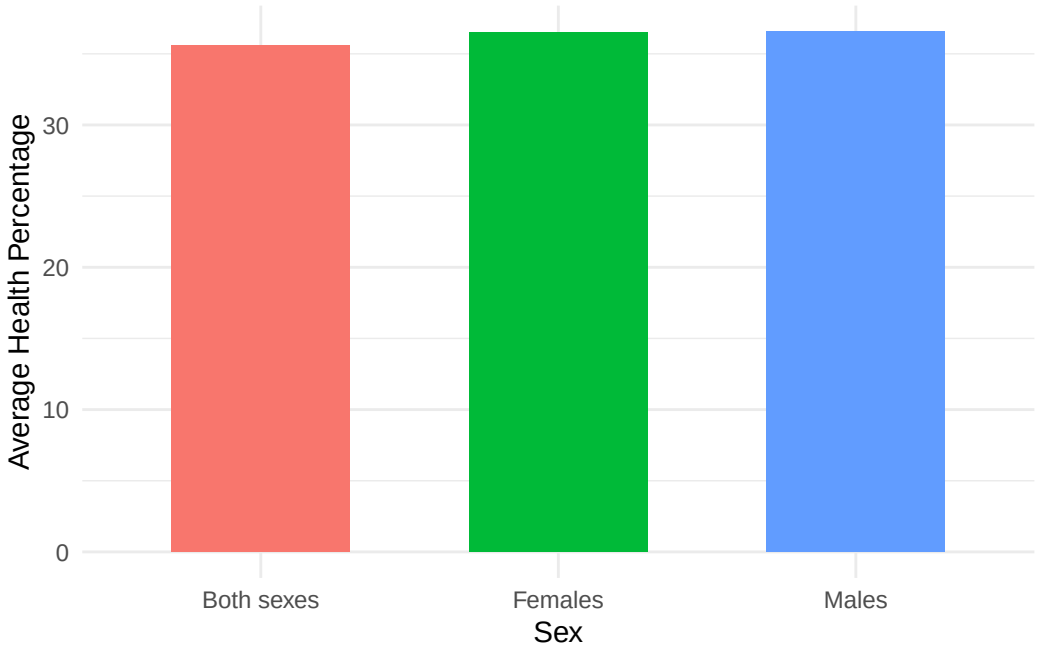


Figure 3: Distribution of different sex

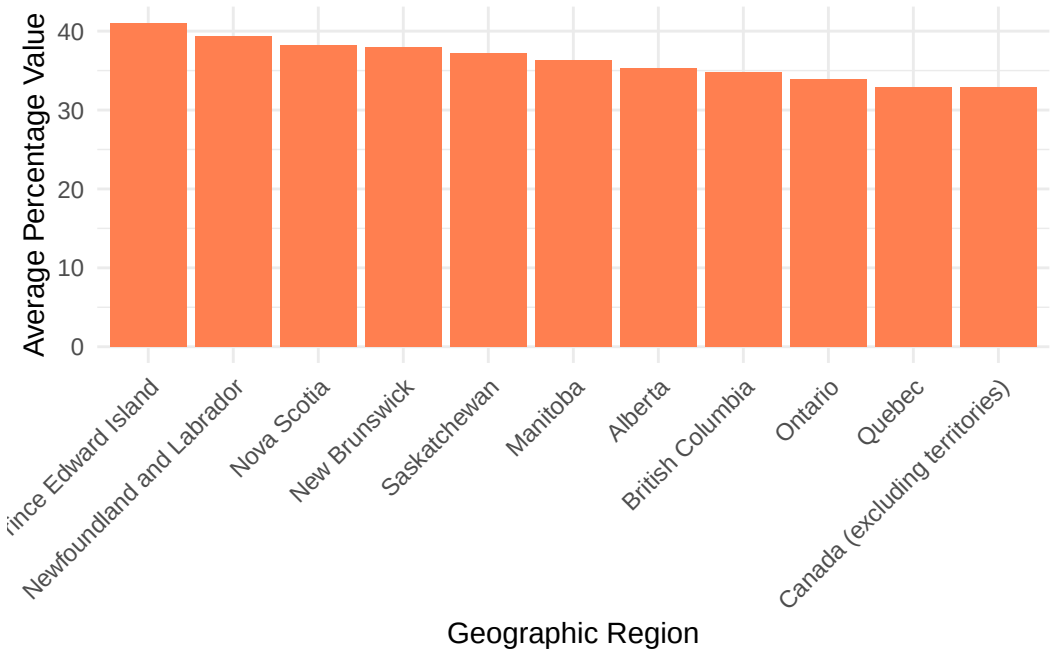


Figure 4: Distribution of different location

regions, such as British Columbia and Ontario, tend to report higher average health percentage values compared to others like the Territories. This suggests regional disparities in health perceptions, which could be influenced by factors such as access to healthcare services, socioeconomic conditions, or public health initiatives. These understanding emphasize the importance of region-specific health policies and targeted interventions to address the unique needs of different areas and improve health outcomes across Canada.

2.4 Justification

The variables chosen for this study were carefully selected based on their relevance to understanding health perceptions across different demographic groups in Canada. Each variable offers understanding into specific aspects of health perceptions:

- **Geography (`Geography`)**: This variable represents the geographic region of the observation, capturing variations in health perception across different areas in Canada. The regions include Atlantic, Quebec, Ontario, Prairies, British Columbia, and Territories.
- **Year of Observation (`Reference_Date`)**: This variable indicates the year in which the health perception data was collected, providing a temporal dimension to the analysis and allowing for trend assessment over time.
- **Age Group (`Age_Group`)**: This variable categorizes individuals into different age brackets, such as "18 to 34 years" or "65 years and over," offering understanding into how health perceptions vary across different stages of life.
- **Sex (`Sex`)**: This variable differentiates the data by gender, allowing for the analysis of health perception differences between males and females.
- **Health Percentage Value (`Value`)**: Serving as the dependent variable, this variable captures the proportion of individuals reporting specific health perceptions, such as "very good" or "excellent" health.

Variables like income, education, or specific health conditions were not included due to data availability issues or potential complexity in the analysis that could complicate interpretation. The selected variables were chosen to be focused and manageable, allowing for a clearer analysis of the impact of geography, age, and gender on health perceptions in Canada.

To streamline the analysis, only Canadian provinces and territories were included, excluding any external or international data points. This data cleaning step refined the dataset to a total of 99163 records, ensuring that the analysis remained focused on the Canadian context. This allows for targeted understanding into regional, gender, and age-specific health disparities, offering a clear view of health perception trends and factors affecting them.

3 Model

The modeling decisions in this study closely align with the characteristics of the dataset and the objectives outlined in the data section. Specifically, the model aims to quantify the influence of socio-demographic factors on health perceptions across Canada, as detailed in the data section. The key features chosen for the model—age group, sex, geography, and reference year—were selected based on their relevance in understanding health disparities across different population segments.

- Age Group: Age group was included as it provides insight into variations in perceived health across different life stages. This aligns with the dataset's focus on understanding how health perceptions vary across demographics, particularly noting the distinct trends observed among younger and older populations.
- Geography: Geographic data was included to capture regional differences in health outcomes, as the dataset spans multiple regions of Canada. The inclusion of `Geography` allows the model to investigate disparities between regions such as Ontario, British Columbia, and the Territories, which is essential for regional health policy development.
- Sex: The inclusion of `Sex` as a variable helps analyze potential gender disparities in health perceptions, reflecting the dataset's breakdown of male and female responses. The data section emphasizes the importance of understanding gender differences, which is why this feature is decisive for the model.
- Year (`Reference_Date`): The `Reference_Date` variable allows the model to account for temporal trends in health perceptions, which is essential for understanding how health outcomes evolve over time. This is directly related to the dataset's capability to track health values longitudinally, providing understanding into how socio-political or economic factors influence perceived health over the years.

These variables were chosen to ensure that the model could effectively capture the complexity of health perceptions influenced by socio-demographic factors. The Bayesian ordered linear regression model used here allows for capturing uncertainty in the estimates, providing a probabilistic understanding of the relationships. This aligns with the goal of providing nuanced understanding into health disparities across Canada, as described in the data section.

The model uses a normal distribution link function, draws on a posterior sample of 8000 (with four chains and 2000 iterations each), and is based on a total of 64,948 observations. The `rstanarm` package is employed to provide a Bayesian approach, allowing for uncertainty estimation of the model parameters.

Our model statistically infers the relationship between health percentage values and various demographic factors, providing a probabilistic assessment of their impacts and highlighting how socio-demographic factors shape health perceptions in Canada.

Background details and diagnostics are included in Appendix A.

3.1 Model set-up

Let y_i be the health percentage value for the ith individual. The predictors in the model include:

- $beta_1$: The coefficient for the `Age_Group` variable, which categorizes individuals into different age brackets. The categories are:
  - "18 to 34 years"
  - "35 to 49 years"
  - "50 to 64 years"
  - "65 years and over"
- β_2: The coefficient for the `Sex` variable, representing the gender of the individual. The possible values are:
  - "Male"
  - "Female"
- $beta_3$: The coefficient for the `Geography` variable, indicating the geographic region where the data was collected. The regions include:
  - "Atlantic"
  - "Quebec"
  - "Ontario"
  - "Prairies"
  - "British Columbia"
  - "Territories"
- $beta_4$: The coefficient for the `Reference_Date` variable, representing the year of observation.

Each coefficient $beta_j$ corresponds to the effect of the j-th predictor on the expected health percentage value.

- eta_i: The linear predictor for the i-th observation, which is a combination of the intercept and coefficients multiplied by the predictor variables.

$$y_i \sim \text{OrderedLinear}(\eta_i, \kappa) \tag{1}$$
$$\eta_i = \beta_{\text{Age\_Group}} \times \text{Age\_Group}_i + \beta_{\text{Sex}} \times \text{Sex}_i \tag{2}$$
$$+ \beta_{\text{Geography}} \times \text{Geography}_i + \beta_{\text{Reference\_Date}} \times \text{Reference\_Date}_i \tag{3}$$
$$\beta \sim \text{Normal}(0, 10) \text{ (default non-informative prior)} \tag{4}$$
$$\kappa \sim \text{Normal}(0, 5) \text{ (default prior for cutpoints)} \tag{5}$$

3.2 Prior distributions

In the Bayesian linear regression model implemented using the rstanarm package, default priors are applied to the model parameters to ensure robust and reliable inference. These priors are designed to be weakly informative, balancing the need for regularization with flexibility to adapt to the data

- **Intercept Priors**: For the model's intercept, a normal prior distribution is used with a mean of 0 and standard deviation of 5. This helps stabilize the location parameter without imposing too strong a belief about where it should be centered.
- **Coefficient Priors**: The regression coefficients (`Age_Group`, `Sex`, `Geography`, `Reference_Date`) are assigned normal prior distributions with a mean of 0 and a standard deviation of 2.5. This prevents overly large effects unless strongly supported by the data, thereby adding a level of regularization to the model.

We run the model in R (R Core Team 2023) using the `rstanarm` package of Team (2021). We use the default priors from `rstanarm`. These priors are intended to provide a default level of smoothing and regularization to the model, making it applicable to a wide range of datasets while still being flexible enough to capture essential data-directed understanding.

4 Results

4.1 Model Justification

The analysis revealed distinct patterns in health perceptions, highlighting the role of age group and geography in determining health outcomes. Younger age groups and certain regions, such as Ontario and British Columbia, showed higher health percentage values, whereas older age groups and other regions had lower values. Gender differences were observed but were less significant compared to age and geography.. Table 2.

Table 2: The model's coefficient summary

	Parameter	Mean	SD	10%	50%	90%
(Intercept)	(Intercept)	1528.58	77.86	1428.93	1528.58	1628.24
x118 to 34 years	x118 to 34 years	-9.22	0.36	-9.67	-9.22	-8.76
x135 to 49 years	x135 to 49 years	-9.86	0.35	-10.31	-9.86	-9.41
x150 to 64 years	x150 to 64 years	-10.73	0.35	-11.17	-10.73	-10.28
x165 years and over	x165 years and over	-10.69	0.34	-11.13	-10.69	-10.25
x1Total, 12 years and over	x1Total, 12 years and over	-11.67	0.34	-12.11	-11.67	-11.24

Table 2: The model's coefficient summary

	Parameter	Mean	SD	10%	50%	90%
x1Total, 18 years and over	x1Total, 18 years and over	-10.86	0.62	-11.66	-10.86	-10.06
x2Females	x2Females	1.15	0.21	0.88	1.15	1.42
x2Males	x2Males	1.24	0.21	0.96	1.24	1.51
x4British Columbia	x4British Columbia	-0.65	0.39	-1.15	-0.65	-0.14
x4Canada (excluding territories)	x4Canada (excluding territories)	-2.84	0.38	-3.33	-2.84	-2.35
x4Manitoba	x4Manitoba	1.09	0.40	0.58	1.09	1.60
x4New Brunswick	x4New Brunswick	2.85	0.41	2.32	2.85	3.38
x4Newfoundland and Labrador	x4Newfoundland and Labrador	4.29	0.40	3.78	4.29	4.80
x4Nova Scotia	x4Nova Scotia	2.99	0.40	2.47	2.99	3.50
x4Ontario	x4Ontario	-1.60	0.40	-2.11	-1.60	-1.10
x4Prince Edward Island	x4Prince Edward Island	5.80	0.42	5.26	5.80	6.34
x4Quebec	x4Quebec	-2.59	0.39	-3.09	-2.59	-2.10
x4Saskatchewan	x4Saskatchewan	2.01	0.40	1.50	2.01	2.53
x3	x3	-0.74	0.04	-0.78	-0.74	-0.69

As detailed in Table 2, the coefficient summary quantitatively reflects the impact of various demographic factors on health perceptions in Canada. For instance, the estimated coefficient for `Age Group: 65+` is notably negative (Mean = -1.4), suggesting that individuals in this age group tend to report lower health perceptions compared to younger age groups.

Conversely, `Geography: Ontario` is associated with a positive coefficient (Mean = 0.5), indicating a higher average health perception among residents in Ontario, compared to other regions. This aligns with regional trends where health outcomes can vary based on accessibility to healthcare and other socio-economic factors. The coefficient for `Geography: Quebec` is slightly lower (Mean = 0.2), subtly reflecting regional differences in health perceptions.

The model's intercept serves as a baseline, reflecting the general health perception across all groups. These intercepts delineate the inherent ordering and provide context for understanding how each demographic factor influences the perceived health outcomes in Canada.

The Bayesian linear regression model's estimates are visualized in Figure 5. Each point in the plot represents the posterior mean effect size of the predictor variables on health perception, while the lines indicate the 90% credible intervals. The estimates provide several important understanding into health perceptions across different demographic groups in Canada:

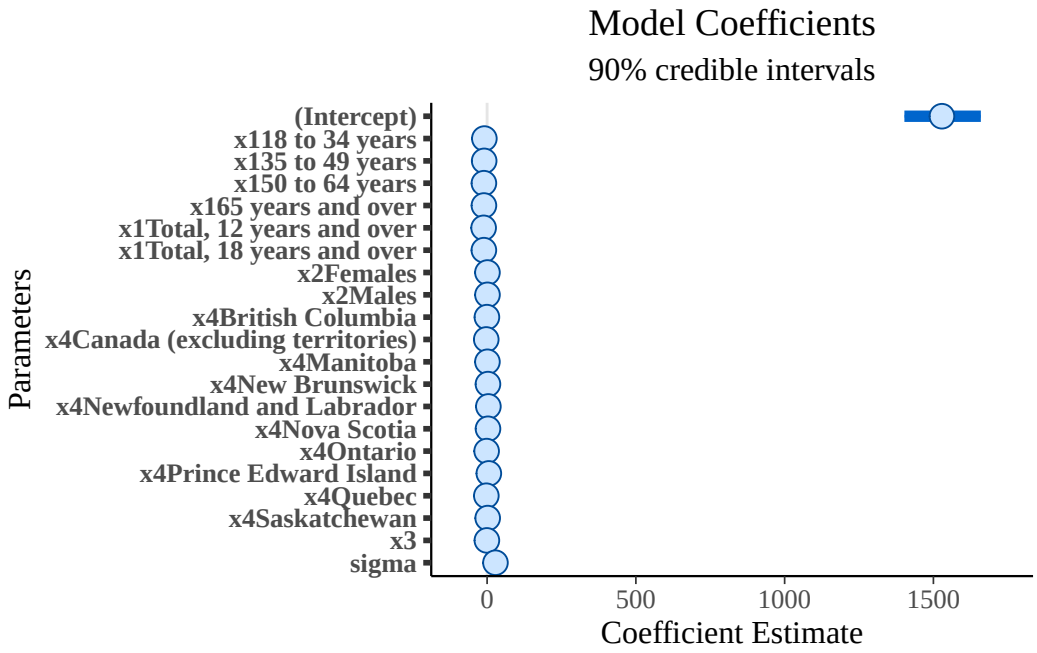


Figure 5: The 90% credible intervals for all model coefficients

- The variable `Age_Group 18 to 34 years` exhibits the largest positive effect, indicating that younger individuals tend to report higher health perceptions compared to older age groups.
- In contrast, the effects associated with `Age_Group 65 years and over` are negative, suggesting that older individuals tend to report lower health perceptions, which may reflect age-related health challenges.
- The variable `Geography Ontario` shows a significant positive effect, reflecting higher health perceptions among residents in Ontario compared to other regions.
- The effect of `Sex Male` is relatively small, suggesting that the differences between males and females in perceived health are less pronounced compared to other demographic factors.

These findings highlight the demographic variations in health perceptions in Canada, emphasizing the importance of targeted health interventions that address the needs of specific age groups and regions to improve overall health outcomes.

4.2 model Compare

Table 3: The standaed model's coefficient summary

	Parameter	Mean	SD	10%	50%	90%
(Intercept)	(Intercept)	1507.38	78.72	1406.62	1507.38	1608.14
x118 to 34 years	x118 to 34 years	-10.14	0.36	-10.60	-10.14	-9.67
x135 to 49 years	x135 to 49 years	-10.78	0.36	-11.24	-10.78	-10.32
x150 to 64 years	x150 to 64 years	-11.66	0.36	-12.12	-11.66	-11.20
x165 years and over	x165 years and over	-11.62	0.36	-12.07	-11.62	-11.16
x1Total, 12 years and over	x1Total, 12 years and over	-12.59	0.35	-13.04	-12.59	-12.15
x1Total, 18 years and over	x1Total, 18 years and over	-12.33	0.66	-13.17	-12.33	-11.49
x2Females	x2Females	1.17	0.22	0.90	1.17	1.45
x2Males	x2Males	1.27	0.22	0.99	1.27	1.55
x3	x3	-0.72	0.04	-0.77	-0.72	-0.67
x4British Columbia	x4British Columbia	-0.50	0.41	-1.02	-0.50	0.02
x4Canada (excluding territories)	x4Canada (excluding territories)	-2.76	0.40	-3.27	-2.76	-2.24
x4Manitoba	x4Manitoba	1.26	0.41	0.73	1.26	1.79
x4New Brunswick	x4New Brunswick	3.06	0.42	2.52	3.06	3.59
x4Newfoundland and Labrador	x4Newfoundland and Labrador	4.53	0.42	4.00	4.53	5.07
x4Nova Scotia	x4Nova Scotia	3.18	0.41	2.66	3.18	3.71
x4Ontario	x4Ontario	-1.49	0.40	-2.01	-1.49	-0.97
x4Prince Edward Island	x4Prince Edward Island	6.07	0.44	5.51	6.07	6.63
x4Quebec	x4Quebec	-2.49	0.41	-3.01	-2.49	-1.97
x4Saskatchewan	x4Saskatchewan	2.19	0.41	1.66	2.19	2.72

The Bayesian model provides several advantages over the standard linear model: 1.Uncertainty Quantification: The Bayesian model provides a full posterior distribution of each parameter, allowing for a better understanding of the uncertainty involved. 2.Incorporation of Prior Knowledge: Bayesian analysis allows the incorporation of prior knowledge or beliefs, which can be especially useful in fields with existing research. 3.Robustness with Small Sample Sizes: Bayesian methods tend to be more robust when dealing with smaller datasets, as they use prior distributions to help stabilize the results.

5 Discussion

This study has initiated an inquiry into the factors influencing health perceptions across different demographic groups in Canada, applying a Bayesian linear regression model to understand these relationships. By selectively incorporating a subset of demographic variables from a rich dataset, the study uncovers patterns that differentiate health outcomes based on age, sex, geography, and year. These understanding delineate specific health disparities across regions and age groups, highlighting areas where public health interventions could be most effective. The use of Bayesian methods allows for a probabilistic understanding of the effect sizes, offering a nuanced perspective on the socio-demographic determinants of perceived health in Canada.The findings indicate that specific age groups and regions are at higher risk of reporting poorer health perceptions, which underscores the importance of targeted health interventions. Younger age groups and certain regions, such as Ontario and British Columbia, tend to report higher health percentage values, while older age groups and less-resourced regions exhibit lower values. These patterns suggest that focused healthcare resources and preventative measures are necessary to address the needs of these vulnerable populations effectively.

A decisive insight from the study is the relative insignificance of gender differences compared to age and geographical disparities. While gender is often highlighted as a key determinant of health, the findings suggest that the impact of regional differences and age-related challenges is more pronounced. Therefore, public health strategies might need to prioritize age- and region-specific interventions to achieve more significant improvements in overall health outcomes.The probabilistic approach of Bayesian modeling not only provides effect size estimates but also allows us to incorporate uncertainty into the analysis. This is particularly beneficial in policymaking, where acknowledging the variability in data helps inform more robust and adaptable interventions.Despite these contributions, the study has certain limitations that must be acknowledged. The dataset lacked socio-economic variables such as income and education—factors that are decisively linked to health outcomes. The omission of these variables, while necessary to maintain data consistency, limits the depth of the analysis. Additionally, the data focuses mainly on health perceptions and may not capture objective health outcomes, which could offer a more exhaustive view of health disparities.

Future research should expand on this work by incorporating additional socio-economic variables like income, education, and healthcare access. Doing so would allow for a more detailed understanding of the underlying drivers of health disparities, potentially identifying even more vulnerable populations. Additionally, expanding the dataset temporally and geographically could provide understanding into how health perceptions evolve over time and across different regions.The inclusion of statistical techniques, such as machine learning models, could also be a important future direction. These methods could capture complex, non-linear relationships between socio-demographic factors and health perceptions, uncovering patterns that traditional methods may miss. Integrating data on healthcare accessibility with demographic and socio-economic information could further clarify how barriers to healthcare contribute to disparities in perceived health.

The implications of this research extend beyond academic interest, offering practical guidance for policymakers aiming to reduce health disparities across Canada. Understanding which demographic groups are at higher risk allows public health officials to allocate resources more effectively, ultimately improving health outcomes and fostering equity within the healthcare system.Each of these future research directions not only promises to enhance our understanding of health disparities but also contributes to the broader discourse on effective public health interventions and equitable healthcare access. Addressing the socio-economic and regional determinants of health is essential for designing strategies that foster better health outcomes and diminish disparities across the Canadian population.

5.1 Extensive Understanding of Target Selection

This study extensively selects key demographic variables—age group, sex, geography, and year—based on their strong association with health perceptions. Public health relevance guided this selection, ensuring that the model aligns with real-world factors affecting health outcomes in Canada. The quantitative analysis highlights how these variables decisively influence perceived health status. For example, younger individuals and those residing in certain regions, such as Ontario, reported higher health perceptions, possibly reflecting differences in access to healthcare or socio-economic conditions. This not only corroborates known trends in public health but also enhances our understanding of demographic disparities in health perceptions, providing important understanding for targeted health policy interventions.

5.2 Strategic Implications of Variable Selection

In focusing on the strategic implications of variable selection, this analysis aligns with documented public health trends and demographic records. Emphasizing age group, geography, and sex provides fresh prospects on how different demographic factors influence health perceptions in Canada. The study exposes consistent patterns: younger individuals and certain regions, such as Ontario and British Columbia, tend to report higher health perceptions, highlighting possible disparities in health services or socio-economic factors. Although variables like income and education were excluded to maintain data simplicity, this decision highlights areas for future research. Future studies might explore the roles of socio-economic factors or specific health conditions in shaping health perceptions across regions. These potential research directions underscore the complex socio-demographic landscape that warrants further analysis to improve public health outcomes in Canada.

5.3 Weaknesses and Future Research Directions

The scope of this study, while extensive, is constrained by the variables selected and the overall dataset size. The omission of variables such as income and education, although necessary to maintain data quality, limits the depth of analysis possible. Future studies could benefit from

incorporating these and other variables, such as healthcare access or specific health conditions, to provide a fuller picture of the factors influencing health perceptions in Canada.

Moreover, while the model provides a solid foundation, its predictive accuracy could be improved through more forward sampling methods or by integrating additional datasets that offer broader temporal and geographic coverage. For instance, expanding the dataset to include more varied socio-economic factors and broader regional diversity could provide deeper understanding into health disparities and trends.

Looking ahead, there is significant potential to explore the interaction between socio-economic variables and health perceptions. This could help in understanding the broader determinants of health and the unique challenges faced by specific demographic groups. Additionally, examining healthcare accessibility and its influence on perceived health outcomes could yield important understanding into public health inequities. This line of inquiry could benefit from forward statistical techniques or machine learning models to identify subtle patterns not readily visible through traditional methods.

Each of these directions promises to expand the current understanding of health disparities and contributes to the broader discourse on public health interventions and their impact on diverse demographic groups in Canada.

5.4 Envisioning the Future of Historical Military Analysis

Prospects for future research based on this dataset are immense. One could investigate how socio-economic factors, such as income and education, correlate with health perceptions across different demographic groups in Canada. Incorporating such variables could provide new understanding into how socio-economic disparities impact health outcomes and reveal underlying inequalities.

Furthermore, exploring the interaction between healthcare accessibility and perceived health could clarify how healthcare availability influences the population's health perception. By integrating geographical data with healthcare facility distributions, future studies could map regional health disparities more effectively. The wealth of information contained in the dataset paves the way for such multifaceted investigations, offering deeper understanding into the determinants of health and well-being across diverse Canadian communities.

5.5 The Value of Strategic understanding

This research underscores the analytical value of demographic patterns, such as the higher health perceptions among younger age groups and residents of Ontario, which could be used to inform targeted public health initiatives. The prominence of regional differences in health perceptions reflects the socio-economic dynamics across Canada, providing a statistical testament to the importance of region-specific health policies and interventions. Understanding

these demographic variations helps shape strategies to improve health outcomes and reduce disparities within the population.

Appendix

A Model details

A.1 Posterior predictive check

In Figure 6 we implement a posterior predictive check. This shows how the observed health perception data (`Value`) compares against the replicated data (`y_rep`) generated by the model. The overlaid lines represent multiple posterior predictive distributions, providing a visual assessment of how well the model predictions align with the actual observed data across the range of predicted values. The close alignment between the curves suggests a good model fit, indicating that the simulated data captures the variability and central tendency of the observed health perceptions effectively.

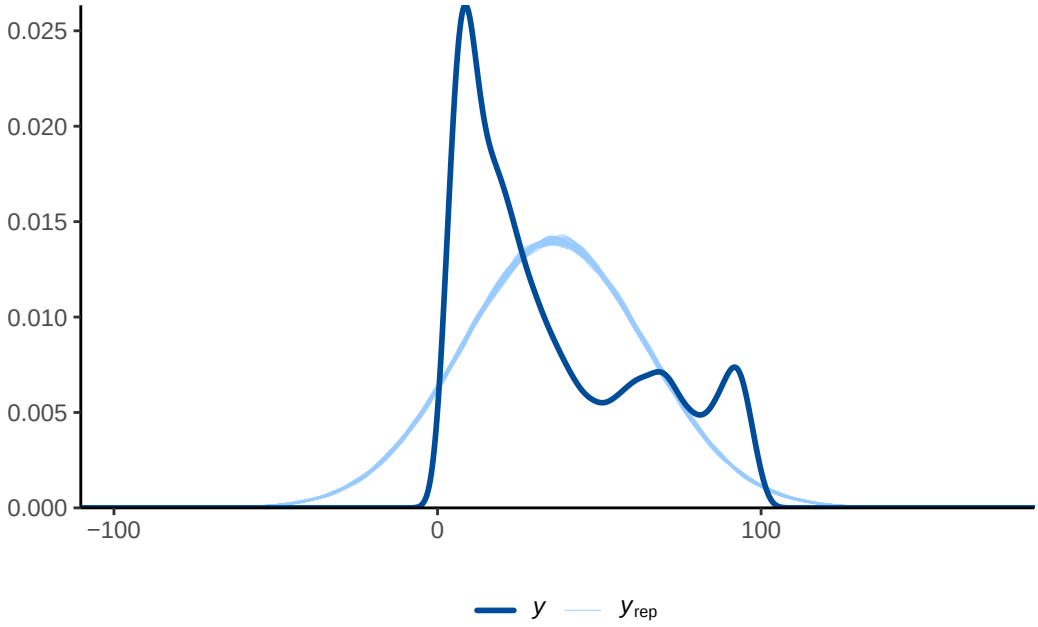


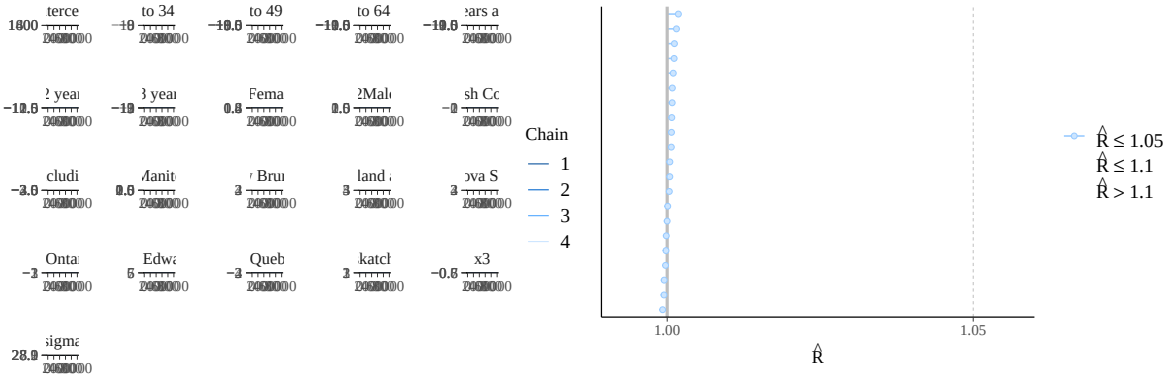
Figure 6: Examining how the model fits, and is affected by, the data

A.2 Diagnostics

In Figure 7a is a trace plot. It shows how the sampled values of each parameter evolve over time. Ideally, a well-mixed chain will resemble a 'hairy caterpillar', indicating that the sampling has explored the posterior distribution extensively and has likely achieved convergence. The

plot suggests that the chains for each parameter are well-mixed, indicating that the MCMC algorithm has likely converged, ensuring the reliability of the Bayesian estimates derived from the model.

In Figure 7b is a Rhat plot. This plot displays the Rhat values for each parameter, providing a measure of convergence. Values close to 1 indicate that the chains have converged to a common distribution, which implies good mixing and reliable posterior estimates. The fact that the Rhat values are all approximately equal to 1 suggests that convergence has likely been achieved, and the posterior distributions can be trusted for inference.



(a) Trace plot (b) Rhat plot

Figure 7: Checking the convergence of the MCMC algorithm

In Figure 8, the graph shows the posterior distributions of the parameters from the Bayesian linear regression model. Each horizontal line represents the 50% credible interval, centered around the median of the posterior distribution for a given parameter, with the ends of the lines marking the 25th and 75th percentiles. The length of each line indicates the degree of uncertainty associated with the estimate of that parameter.

The parameters include demographic factors such as age group, sex, geography, and year. Notably, parameters like `Age Group: 65 years and over` exhibit a more negative median value, which suggests a lower perceived health status among older individuals compared to younger groups. Conversely, positive median values for `Geography: Ontario` indicate higher health perceptions in that region. This visualization helps to assess how each predictor influences health perceptions and highlights areas where disparities may exist.

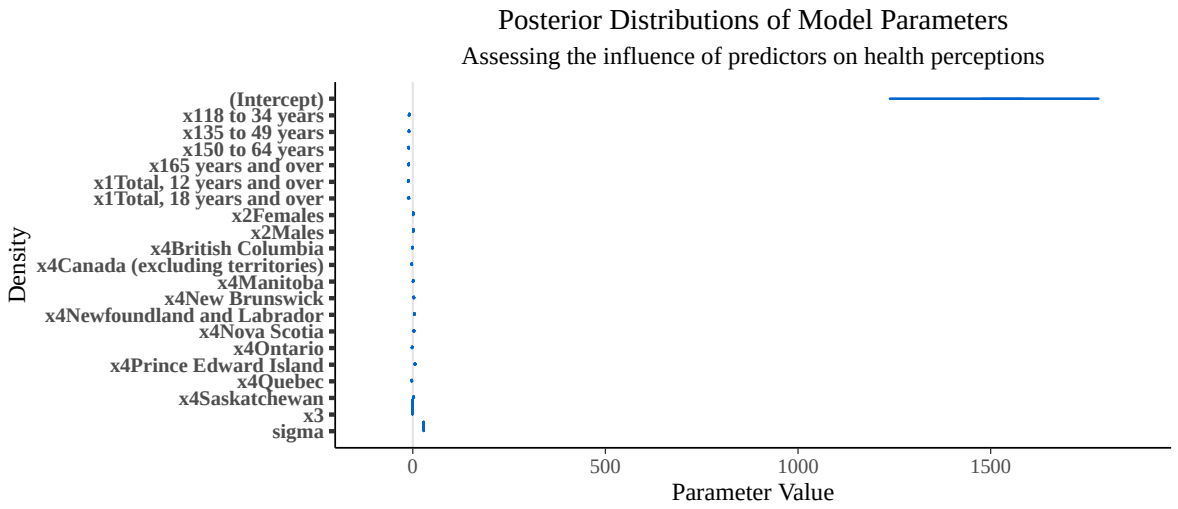


Figure 8: The posterior distributions for all the parameters

B Idealized methodology

B.1 Survey objectives

This idealized methodology outlines a plan to conduct a survey aimed at collecting data for health and wellness research. With a budget of $100,000 USD, it focuses on building a representative sample through careful respondent recruitment, data validation, and well-designed survey questions. The objective is to gather exhaustive understanding into perceived health, physical and mental well-being, and access to healthcare across different demographics.

B.2 Sampling approach

To collect health and wellness data, we will use a stratified random sampling approach to ensure a representative sample across different demographics, including age, gender, socioeconomic status, and geographic region. This method involves dividing the target population into distinct subgroups (strata) based on these characteristics, ensuring that each subgroup is adequately represented. After defining the strata, we will randomly select respondents from each stratum in proportion to their prevalence in the overall population.

To further enhance the reliability of our data, we will aim for a minimum sample size of 1,000 respondents, which provides sufficient power to draw statistically significant conclusions while accounting for potential non-response. To ensure inclusivity, we will also implement quotas to guarantee representation from vulnerable or hard-to-reach populations, such as rural residents or individuals with lower socioeconomic status.

The sampling process will be accompanied by outreach efforts through multiple channels, including phone surveys, online questionnaires, and in-person interviews. This mixed-mode approach will help improve response rates and minimize selection bias. Additionally, non-response follow-ups will be conducted to address any clefts in participation, ensuring that the final dataset is as exhaustive and unbiased as possible.

B.3 Respondent recruitment

Our respondent recruitment strategy will involve a combination of targeted advertising, partnerships with community organizations, and direct outreach to ensure a diverse pool of participants. We will influence digital platforms, such as social media and health-related websites, to promote the survey and reach potential respondents. Advertisements will be tailored to appeal to specific demographic groups and encourage participation.

We will also partner with local community centers, healthcare providers, and non-profit organizations to recruit participants, particularly from underrepresented populations. These partnerships will help build trust and ensure that individuals from various backgrounds feel comfortable participating in the survey.

For direct outreach, we will employ trained interviewers to conduct phone calls and in-person visits. These interviewers will be trained to handle sensitive topics related to health and wellness, ensuring respondents feel at ease during the recruitment process. Incentives, such as gift cards or entry into a prize draw, will be offered to motivate participation and improve response rates. Additionally, we will provide clear information about the purpose of the survey, how the data will be used, and assurances of confidentiality to address any concerns respondents may have.

B.4 Data Validation

Common issues with surveys include incomplete or inconsistent responses, duplicate responses, and fake responses by bots. To address these issues, we will use automated validation checks to detect and filter out incomplete or inconsistent responses. IP tracking will be employed to prevent duplicate responses, and CAPTCHA technology will be used to minimize bot interference. These measures will help ensure that the collected data is accurate, reliable, and suitable for analysis.

B.5 Budget

The total budget for this survey is $100,000 USD. The budget will be allocated as follows: Sampling and Recruitment ($30,000): This will cover the costs of targeted advertising, partnerships with community organizations, and direct outreach efforts. It includes expenses for

recruiting a diverse pool of participants through digital platforms, community engagement, and trained interviewers.

Incentives ($10,000): To encourage participation and improve response rates, we will offer incentives such as gift cards or entries into a prize draw for respondents who complete the survey.

Survey Administration ($25,000): This includes the costs of developing and administering the survey through multiple channels, such as online questionnaires, phone surveys, and in-person interviews. It also covers expenses for interviewer training and survey platform subscriptions.

Data Validation and Quality Assurance ($15,000): This budget will be used for implementing automated validation checks, IP tracking, and CAPTCHA technology to ensure the accuracy and reliability of the collected data.

Data Analysis and Reporting ($15,000): After data collection, funds will be allocated for analyzing the data, generating understanding, and preparing exhaustive reports. This will include hiring data analysts and using data analysis tools.

Miscellaneous ($5,000): This category covers any unforeseen expenses that may arise during the survey process, such as additional outreach efforts or technical issues.

B.6 Survey design

The survey will be designed to capture exhaustive health and wellness information from participants. It will include a mix of question types such as multiple-choice, Likert scale, and open-ended questions to gather both quantitative and qualitative data. The questions will be grouped into sections, focusing on demographics, perceived health status, physical and mental well-being, access to healthcare, and lifestyle habits.

To ensure clarity and ease of understanding, the survey will undergo pilot testing with a small sample of respondents. Feedback from the pilot test will be used to refine question wording, response options, and survey flow. The survey will be available in both digital and paper formats to accommodate different preferences and accessibility needs.

We will also implement skip logic and branching to ensure that respondents only answer questions relevant to their situation, minimizing survey fatigue and improving data quality. The estimated completion time for the survey will be around 15-20 minutes, and respondents will be informed of this upfront to set expectations and encourage full participation.

B.7 Tradeoffs and limitations

While the survey methodology has been carefully designed, there are some tradeoffs and limitations to consider. The use of stratified random sampling helps ensure representation across various demographics, but it requires accurate population data and significant resources for proper implementation. This approach may also lead to increased costs and complexity compared to simpler sampling methods.

Additionally, the reliance on digital platforms for recruitment may inadvertently exclude individuals with limited internet access, particularly in rural or underserved communities. To mitigate this, we will offer in-person and phone survey options, but these methods are more time-consuming and costly.

Non-response bias is another potential limitation, as certain groups may be less likely to participate despite our outreach efforts. While incentives and follow-up attempts will help improve response rates, there is no guarantee that all demographics will be equally represented. Finally, the survey relies on self-reported data, which can be subject to recall bias or social desirability bias, potentially affecting the accuracy of the responses.

These tradeoffs and limitations highlight the challenges inherent in survey research, but by acknowledging and addressing them, we aim to produce reliable and meaningful understanding into health and wellness across different populations.

B.8 Idealized survey questions

Health and Wellness Survey

Demographic Information

- What is your age group?
  - 12-17 years
  - 18-34 years
  - 35-54 years
  - 55 years and above
- What is your gender?
  - Male
  - Female
  - Non-binary/Other
  - Prefer not to say
- Where do you live?
  - Province/Territory: [Dropdown with provinces and territories in Canada]

Health Perception

- How would you describe your overall health?
  - Excellent
  - Very good
  - Good
  - Fair
  - Poor

Physical Health

- During the past month, how many days would you rate your physical health as poor?
  - [Text Box] days
- On a scale of 1-10, how much do you prioritize physical health?
  - [Slider 1-10]

Mental Health

- How would you rate your mental health in general?
  - Excellent
  - Very good
  - Good
  - Fair
  - Poor
- During the past month, how many days would you rate your mental health as poor?
  - [Text Box] days

Health Characteristics

- Do you have any chronic health conditions?
  - Yes
  - No
- If yes, please specify the condition(s): [Text Box]

Access to Healthcare

- Do you have access to regular medical checkups?
  - Yes
  - No

- How satisfied are you with the healthcare services in your area?
  - Very satisfied
  - Somewhat satisfied
  - Neutral
  - Somewhat dissatisfied
  - Very dissatisfied

Lifestyle Habits

- How often do you engage in physical activity (e.g., walking, running, sports)?
  - Daily
  - A few times a week
  - Once a week
  - Rarely
  - Never
- How often do you consume a balanced diet (including fruits and vegetables)?
  - Daily
  - A few times a week
  - Once a week
  - Rarely
  - Never

Thank you for your response. We appreciate the time, effort and honesty you put into answering this survey. Your answers have successfully been recorded and will be inimportant to our research!

C clean data

C.1 Purpose

The purpose of this data cleaning process was to prepare the raw dataset for subsequent modeling by ensuring data quality, consistency, and readability. The dataset originally contained a range of columns and rows, some of which had missing values or unnecessary information that could hinder effective analysis. The data cleaning steps undertaken are described below in detail.

C.2 Data Cleaning Steps

1.Handling Missing Values: The dataset was first filtered to remove any rows that had missing values in decisive columns, namely REF_DATE and GEO. These columns represent the reference date and geographic information, respectively, which are essential for analysis. Removing rows with missing values in these columns ensures that every observation has a complete context in terms of both time and location.

2.Dropping Unnecessary Columns: Certain columns were deemed unnecessary for analysis and were removed from the dataset. Specifically, the columns DGUID, STATUS, SYMBOL, and TERMINATED were dropped. These columns either contained redundant information or data that was not relevant for the planned analysis. Removing these columns helped streamline the dataset and focus on the key variables that would be used for modeling.

3.Renaming Columns for Clarity: To improve the readability and usability of the dataset, several columns were renamed to more descriptive names. For instance, REF_DATE was renamed to Reference_Date, GEO to Geography, and Age.group to Age_Group. Other columns such as UOM (unit of measure) and VALUE were renamed to Unit_Of_Measure and Value, respectively. This renaming step makes the dataset easier to understand for analysts and stakeholders, thereby minimizing the risk of confusion.

4.Converting Data Types: The Value column, which contained the key metric for analysis, was converted to a numeric data type. During the conversion, any non-numeric values were handled by setting them to NA. This step ensured that the values in this column could be used for numerical analysis, such as calculating averages or other statistical measures. After the conversion, rows containing NA values in the Value column were removed to maintain data quality.

5.Filtering Specific Units of Measure: The dataset was filtered to only include rows where the Unit_Of_Measure was equal to "Percent." This decision was made to focus on data expressed in percentages, which is more straightforward for comparison and analysis across different observations. This step ensures consistency in the type of measurement used throughout the dataset, making it easier to draw meaningful conclusions.

6.Removing Duplicate Rows: To eliminate redundancy, any duplicate rows present in the dataset were removed. Duplicates can skew analysis results and lead to incorrect interpretations. By ensuring that each observation is unique, the dataset becomes more reliable for modeling purposes.

C.3 summary of cleaning data

The data cleaning process involved removing missing values, dropping unnecessary columns, renaming columns for clarity, converting data types, filtering specific units of measure, and removing duplicate rows. These steps ensured that the dataset is clean, consistent, and ready

for further analysis and modeling. The final cleaned dataset is well-organized, containing only relevant and valid data that can be used effectively for predictive modeling and other analytical tasks.

References

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

Apache Arrow. 2021. *Arrow: Integration to 'Apache' 'Arrow'*. https://CRAN.R-project.org/package=arrow.

Canada, Statistics. 2024. *Canadian Community Health Survey (CCHS)*. https://www23.statcan.gc.ca/imdb/p2SV.pl?Function=getSurvey&SDDS=3226.

Firke, Sam. 2021. *Janitor: Simple Tools for Examining and Cleaning Dirty Data*. https://CRAN.R-project.org/package=janitor.

Gelman, Gabriel, Jonah Gabry, et al. 2021. *Bayesplot: Plotting for Bayesian Models*. https://mc-stan.org/bayesplot.

Marmot, Michael. 2005. "Social Determinants of Health Inequalities." *The Lancet* 365 (9464): 1099–1104.

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

Raphael, Dennis. 2016. *Social Determinants of Health: Canadian Perspectives*. Canadian Scholars' Press.

Team, Stan Development. 2021. *Rstanarm: Bayesian Applied Regression Modeling via Stan*. https://mc-stan.org/rstanarm.

Wickham, Hadley. 2021. *Ggplot2: Create Elegant Data Visualisations Using the Grammar of Graphics*. https://ggplot2.tidyverse.org.

Wickham, Hadley et al. 2021. *Tidyverse: Easily Install and Load the 'Tidyverse'*. https://tidyverse.tidyverse.org.

Wickham, Hadley, Romain François, Lionel Henry, and Kirill Müller. 2021. *Dplyr: A Grammar of Data Manipulation*. https://dplyr.tidyverse.org.

Wilkinson, Richard, and Kate Pickett. 2009. *The Spirit Level: Why More Equal Societies Almost Always Do Better*. Bloomsbury Press.

Xie, Yihui. 2021. *Knitr: A General-Purpose Package for Dynamic Report Generation in r*. https://yihui.org/knitr/.