Causal Bayesian machine learning to assess the heterogeneous effect of smoking or drinking with mortality: A longitudinal analysis of Chinese Longitudinal Healthy Longevity Study

TINGJIAO CUI

Causal Bayesian machine learning to assess the heterogeneous effect of smoking or drinking with mortality: A longitudinal analysis of Chinese Longitudinal Healthy Longevity Study

Tingjiao Cui

A Thesis presented for the degree of
Master of Mathematics

Department of Mathematics
Durham University
United Kingdom
October 2023

Abstract

This dissertation introduces an innovative methodology employing Bayesian machine learning to identify heterogeneous effects. This method provides a quantitative perspective on potential effect-modifying factors influencing the heterogeneity in the associations between the independent variable and the outcome.

The study cohort consisted of 43,487 individuals from the Chinese Longitudinal Healthy Longevity Survey (CLHLS), a longitudinal cohort study of elderly Chinese individuals. Numerous studies have shown that the association between smoking or drinking and mortality varies significantly with mediators such as physical activity level and diet. However, only some studies have systematically assessed the heterogeneous effects of this association. To address this gap, this research investigates the association between smoking or drinking and mortality across several subgroups identified by the Bayesian machine learning method. The results reveal significant variations in the association based on body weight and physical activity levels. Specifically, among individuals weighing 57 kilograms or more, a heightened risk of mortality is observed with low levels of physical activity. In contrast, among individuals weighing less than 57 kilograms, only a high level of physical activity is linked to an increased mortality risk.

The methodology employed in this study involves a two-step approach. First, the missForest algorithm was used for data imputation to handle missing values, ensuring a robust and accurate dataset. MissForest's non-parametric nature and effectiveness in managing complex interactions and non-linear relationships make it an ideal choice for this diverse dataset. Second, Bayesian Additive Regression Trees (BART) were applied to analyze the imputed data. BART is particularly adept at capturing non-linear relationships and interactions among predictors, enhancing the statistical power to detect true heterogeneous effects.

By handling the imputation separately, we ensured that the BART model could

focus on identifying and modeling the intricate interactions between smoking, drinking, and mortality without the additional complexity of simultaneously imputing missing values. In summary, using missForest for imputation, followed by BART for modeling, provided a robust and effective methodology for our study. This combination leveraged the strengths of both techniques, ensuring accurate and reliable imputation of missing data and powerful, flexible modeling of the relationships between variables.

This research demonstrates that the Bayesian machine learning method can effectively identify heterogeneous effects between the independent variable and the outcome. The integration of advanced statistical methods highlights the potential for precision medicine approaches in epidemiological research. Furthermore, the findings highlight the multifaceted nature of the relationships among body weight, physical activity, smoking or drinking, and the risk of mortality. This underscores the importance of considering lifestyle factors, such as smoking or drinking, along with body weight and physical activity, when examining mortality risk. These insights are valuable for precision medical interventions.

The methodology, specifically Bayesian Additive Regression Trees (BART), demonstrates transparency, reproducibility, and robustness. This research contributes to the biomedical field by providing valuable methodological insights and advancing our understanding of potential effect-modifying factors in complex associations. Further research is warranted to explore the underlying mechanisms and potential confounding factors that may influence these associations.

Declaration

The work in this thesis is based on research carried out at the Department of Mathematics, Durham University, United Kingdom. No part of this thesis has been submitted elsewhere for any other degree or qualification and it is all my own work unless referenced to the contrary in the text.

Acknowledgements

I embarked on this journey with high hopes and expectations a year ago. The school and this city have witnessed every step of my growth. Looking back, I see a broader world and have met many remarkable individuals. However, the time has come for me to bid farewell to graduate school life. As I approach the completion of my thesis, I want to express my gratitude to those who have supported, helped, and encouraged me along the way.

First and foremost, I extend my heartfelt thanks to my supervisor, Associate Professor Georges, who has been a source of great inspiration and assistance in my studies and personal life over the past year. Under his guidance, inspiration, and encouragement, I have made significant progress and had the privilege of attending various academic lectures, where I learned extensively about Bayesian statistics. Whether in my daily studies or during the thesis-writing process, Associate Professor Georges provided detailed suggestions and enthusiastic support whenever I needed it. I am also deeply grateful to the teachers at the School of Mathematics. Their classes have enriched my understanding of mathematics and statistics, and I have gained invaluable expertise.

Home has always been my warmest harbor. I would like to express my sincere gratitude to my dear family. Their selfless love and unconditional support have given me the courage to take on challenges and pursue my aspirations. They are my strongest support and source of strength. I wish to care for myself better and be there for my dear family.

I am also deeply thankful for the friendships I have formed. My friends have enriched my graduate life and provided warmth and comfort. During times of anxiety, they offered the utmost patience and understanding. Every moment with them has brought me ease and renewed my positive energy. I would not have discovered so much beauty without their companionship in an unfamiliar city. We will all have

bright futures and shine in various ways in our respective fields.

Contents

List of Figures

List of Tables

CHAPTER 1

Introduction

In this chapter, I provide the background for this research, which includes three key components: 1) the urgent need to identify heterogeneous effects, 2) the epidemiological background of smoking or drinking in relation to mortality, and 3) the research objective of identifying the heterogeneous effects of smoking or drinking on mortality.

In recent years, the identification of heterogeneous effects in clinical studies has gained prominence [1]. These effects reveal varying responses to therapies among individuals, aligning with the goals of precision medicine, which aims to tailor prevention and treatment strategies for optimal allocation of health resources [2]. A key aspect of this field is identifying who benefits most from specific interventions and who may not or might be harmed. Traditional subgroup analyses, while commonly used, have limitations such as underpowered results, potential overlaps, and reliance on pre-specified analyses that may underestimate the heterogeneous effects [1]. Additionally, single-variable subpopulation definitions, like age or sex, often fail to capture the complex mechanisms in treatment responses. More sophisticated, data-driven methods that consider multiple characteristics are needed to address these limitations.

Recently, novel data-driven methods have been introduced to effectively identify combinations of heterogeneous effects in epidemiological studies [3]. These innovative approaches leverage advanced statistical techniques and artificial intelligence algorithms to overcome the challenges posed by traditional subgroup analyses. One significant advantage of data-driven methods is their ability to analyze large datasets containing diverse patient characteristics and treatment responses, allowing researchers to discern patterns and associations that might remain undetected through conventional methods. Moreover, these approaches can identify interactions among variables that play a crucial role in determining the efficacy or risk of a particular intervention without human bias. The increasing complexity of data from epidemiological or clinical studies, such as genomics or proteomics, underscores the need for such advanced methods [4].

Smoking and drinking are critical public health issues, often occurring together. Numerous scientific investigations have unequivocally demonstrated the severe and detrimental consequences of lifelong tobacco smoking. Simultaneously, a multitude of studies have highlighted increased mortality rates associated with chronic heavy drinking. Interestingly, evidence suggests that regular light drinking may have minimal impact on overall mortality and could potentially offer some protection against coronary heart disease. The intertwining reality is that many individuals engage in both smoking and drinking behaviors.

Furthermore, the neurochemical mechanisms of action of nicotine and alcohol exhibit mutual reinforcement [5]. In response to the growing public health challenge `https://publichealth.nyu.edu/events-news/news/2023/01/23/global-public-health-goes-global-climate-summit` posed by the coexistence of smoking and drinking, numerous smoking cessation and alcohol reduction programs have been implemented globally. These programs aim to promote healthier behaviors and mitigate associated risks. However, despite the widespread adoption of these interventions, a critical gap remains in our understanding. Specifically, no comprehensive studies have systematically assessed the heterogeneous effects on mortality in individuals who both smoke and drink. Such investigations are essential for identifying subpopulations that might be at heightened risk and for tailoring interventions

to address the specific needs of these individuals [6]. By identifying and targeting high-risk subgroups, healthcare resources can be allocated more effectively, thereby improving overall population health outcomes.

This study leverages the Chinese Longitudinal Healthy Longevity Survey (CLHLS) [7] to examine the effects of smoking and drinking on mortality. It aims to understand their interplay and influence on health. Our study will thoroughly assess the demographic, lifestyle, and health-related factors associated with smoking and drinking behaviors. By examining these factors in tandem, we can elucidate potential interactions and patterns that might contribute to differential mortality outcomes in various subgroups. Moreover, we will consider other contextual factors, such as socioeconomic status, access to healthcare, and environmental `https://vorstcanada.com/blogs/news/artemisinin-for-malaria` influences, which may further contribute to the heterogeneous effects of smoking and drinking on mortality.

The methodology employed in this study involves a two-step approach. We first address the missing data using the missForest algorithm, ensuring that the imputed values are consistent with the underlying data structure. We then apply BART for modelling; BART, with its ability to model complex interactions and improve statistical power, was well-suited for analyzing the heterogeneous effects of smoking and drinking on mortality. By combining missForest and BART, we achieve a robust and efficient analytical framework that maximizes the strengths of both methods. This integrated approach not only simplifies the modelling process but also ensures that the imputed data are of high quality, thereby enhancing the overall reliability and validity of our findings.

In conclusion, the convergence of smoking and drinking behaviours represents a significant public health challenge. While the adverse effects of lifelong tobacco [5] smoking and chronic heavy drinking are well-established, the interplay between these behaviours and their combined impact on mortality remains less explored. By leveraging data from the CLHLS and By combining missForest and BART, our research seeks to bridge this knowledge gap and provide valuable insights into [8] the heterogeneous effects of smoking and drinking on mortality [9]. The integration of advanced statistical techniques and data modelling highlights the potential for pre-

cision medicine approaches in epidemiological research; we aim to identify high-risk subpopulations and tailor interventions to enhance healthcare resource allocation and improve overall population health; we hope to contribute to the development of evidence-based and personalized public health strategies that pave the way for a healthier and more resilient population.

1.1 Background

Heterogeneous exposure associations (HEAs) in public health reflect the variability of exposure-outcome relationships across different subgroups. This variability is influenced by numerous factors, including genetics, environment, lifestyle, and social determinants of health, making it a complex yet vital area of study.

For example, research on smoking and lung cancer has shown that not only genetic factors but also environmental and lifestyle factors significantly impact the risk of lung cancer among smokers [10]. Individuals exposed to secondhand smoke or those living in areas with high air pollution may face a higher risk, even if their smoking habits are less severe [11] [12].

In the context of medication efficacy, consider the case of anti-hypertensive drugs. The effectiveness of these drugs can vary significantly among different ethnic groups. Certain blood pressure medications, for instance, are more effective in African American populations compared to Caucasian populations, underscoring the importance of considering ethnic and racial backgrounds in treatment plans [13].

Similarly, in the field of mental health, the response to antidepressants is another area where HEAs are evident. Studies have demonstrated that factors like age, gender, genetic makeup, and the severity of depression can influence how individuals respond to these medications. This variability necessitates a more tailored approach to treatment `https://biocertica.com/blogs/what-medicine-should-i-take/revolutionizing-gout-management-a-personalized-approach-through-pharmacogenetics`, moving away from the 'one size fits all' model [14] [15].

In nutrition and dietetics, the concept of HEAs is increasingly recognized. The impact of various diets on health outcomes, such as obesity, diabetes, and car-

diovascular diseases https://primewomen.com/wellness/health/male-menopause/, can vary significantly based on an individual's genetic background, metabolic rate, and lifestyle. For instance, the effectiveness of a low-carbohydrate diet in weight loss [4]and metabolic improvement might be more pronounced in some individuals, depending on their genetic predisposition to metabolize different types of nutrients [16].

Understanding HEAs is crucial for developing more personalized and effective healthcare and public health interventions. It allows researchers and practitioners to tailor their approaches based on the unique characteristics of different subgroups, thereby enhancing the effectiveness of treatments and preventive measures. As research in this field advances, it becomes increasingly clear that a one-size-fits-all approach is often inadequate in addressing the complex and varied nature of human health and disease. Therefore, exploring heterogeneous exposure in societies is not only a scientific necessity but also a practical imperative to ensure optimal health outcomes for diverse populations [2] [17].

1.2 Research Objectives

The primary objective of this research is to investigate the relationship between smoking and drinking habits and mortality within a study population, characterized by a diverse range of baseline variables. Specifically, this study aims to explore whether additional factors among these baseline variables significantly influence the relationship between smoking or drinking and mortality. The goal is to identify factors beyond smoking and drinking that may modulate or alter the impact of these behaviours on the risk of death. To address this objective, the study will collect and analyze multiple variables, including age, gender, race, socioeconomic status, lifestyle factors, genetic background, medical history, and environmental exposures. The analysis will assess whether these variables interact with smoking or drinking habits in predicting mortality risk. This approach will help determine if certain groups are more susceptible to the adverse effects of smoking or drinking and identify factors that may mitigate or exacerbate the health risks associated with

these behaviours.

This study uses the missForest algorithm for imputation, followed by BART for modelling. By combining missForest and BART, we achieve a robust and efficient analytical framework that maximizes the strengths of both methods. Separating the imputation process from the analysis process ensures methodological transparency, which allows us to maintain methodological rigour while optimizing computational efficiency. We leverage advanced statistical techniques, specifically Bayesian Additive Regression Trees (BART), to uncover complex interactions and heterogeneous effects that traditional methods might overlook. The innovative use of BART allows for a more nuanced data analysis, providing insights into the potential synergistic effects of multiple variables on mortality. This methodological approach not only enhances the robustness and accuracy of the findings but also highlights the potential for precision medicine approaches in epidemiological research by demonstrating the value of data-driven techniques in uncovering intricate relationships in public health data.

The overarching aim of this research is to achieve a comprehensive understanding of the multifactorial influences on mortality risk, particularly in the context of smoking and drinking habits. This understanding is crucial for developing more effective public health strategies and individualized preventive measures `https://newswebsite.com/5-signs-of-depression-in-women/`. By identifying key baseline variables that interact with smoking and drinking behaviours, this study seeks to formulate precise health interventions that can reduce mortality risks associated with unhealthy lifestyle choices. This will enable the development of targeted public health interventions and improve our understanding of how various baseline characteristics interact to affect health outcomes `https://www.fruitandveggie.com/berry-compounds-can-reduce-high-blood-pressure-3494/`.

CHAPTER 2

Longitudinal cohort data for analysis

This section describes the data used in this study, including 1) study population, which describes the cohort design and population size of CLHLS; 2) data on mortality, which describes the collection of the mortality outcome in this study; 3) exposure, which describe the data collection procedure of smoking and drinking in the study and its encoding method; 4) effect modifiers, which describe the variables used in this study as mediator; 5) imputation technique, which describe the methodology used for missing data imputation.

2.1 Study population

Our study draws on data from the Chinese Longitudinal Healthy Longevity Study (CLHLS), a large-scale study of health status and quality of life conducted in 23 provinces out of 31 provinces in China since 1998 [18]. Follow-ups are conducted every 2-3 years [18], and eight waves have been completed so far. The study covers about 85% of the Chinese population and aims to identify the determinants of human health and longevity. CLHLS attempted to interview all consenting centenarians in the surveyed counties and cities [19]. The CLHLS researched 16,557

centenarians and 23,081 people aged 90 and older by gender and place of residence (i.e., living on the same street, city, town, or county). We conducted 96,805 face-to-face interviews with 25,842 octogenarians and 19,650 younger adults aged 65 to 79, collecting detailed longitudinal data on older adults' physical and mental health, social participation, and cognitive functioning [20].

This matched recruitment procedure resulted in an oversampling of the oldest and older men. In CLHLS, to reflect the unique sampling design, the age and gender weights of urban and rural residents in the sample are consistent with the total population distribution of the 22 provinces in the sample. In this study, we excluded some participants from the CLHLS data to increase the accuracy and validity of the analysis. We excluded participants younger than 65 years old (n=381) and participants with missing values for smoking and drinking (n=308). Therefore, the final dataset used for analysis in our study included 43,487 older adults aged 65 years or older from the CLHLS [21].

A significant advantage of using the CLHLS dataset is the information available about participants' sociodemographic characteristics, lifestyle behaviors, health status, and other relevant factors. This allows researchers to control for potential confounding variables and improve the accuracy of study results. Given the multifaceted nature of healthy aging and longevity [22], we need to consider a wide range of variables and their potential impact, for example, lifestyle factors such as diet and physical activity, as well as underlying health conditions and access to health care, which may affect play an essential role in influencing the relationship between smoking, drinking, and longevity. Therefore, we will carefully consider these variables in our analysis to ensure the robustness of the results [23].

Many previous studies have published in-depth analyses of CLHLS [24], revealing complex interactions among various determinants of healthy longevity, and these results laid the foundation for our research. Our study aimed to focus on the association between smoking, drinking, and longevity in older adults, as well as the heterogeneous effects of this association. To study these associations, we will use sophisticated statistical methods to analyze data sets and draw meaningful conclusions. We will use descriptive statistics, regression models [25], and other appro-

priate techniques to examine the relationships between smoking, alcohol use, and longevity outcomes. Additionally, we will identify potential interactions and changes in these associations between different demographic groups by conducting subgroup analyses. The study aimed to shed light on the link between smoking, drinking, and longevity in older adults. Findings from this study have the potential to inform targeted public health interventions and help advance knowledge in the areas of healthy aging and longevity research.

2.2 Data on mortality

In this study, we ensured the reliability and validity of mortality data through the following three aspects:

First, we obtained comprehensive information on vital status and date of death through accurate and efficient data collection methods [26]. We used officially issued death certificates whenever available to obtain the correct details of the deceased. When death certificates were not available, we obtained information from the next of kin or a local residential committee with in-depth knowledge of the deceased [26]. This approach ensured the reliability and completeness of the mortality data.

Secondly, to calculate the duration of follow-up for each participant [27], we recorded the time interval between the date of the first interview and the date of death [27], allowing us to estimate the length of time each individual was monitored during the study.

Finally, for participants who were still alive when the last wave of data was collected in 2018, their data were censored at the time of the final survey [28] . This censorship technique is essential to ensure the integrity of the study. Data on individuals who have not experienced the event of interest (in this case, death) at the end of the study is critical. By examining data from the last survey, the analysis accounted for the possibility that these individuals may still experience the event in the future.

The accuracy and completeness of mortality data were ensured through the use of official death certificates [29] and information from next of kin and local neighbor-

hood committees [29]. This is crucial in epidemiological studies, as reliable mortality data are essential for drawing accurate conclusions and making informed decisions about public health interventions https://freescience.info/epidemiology-a-comprehensive-guide/.

Furthermore, the calculation of follow-up duration provides essential insights into the study's temporal aspects. By accurately determining the period during which each participant was monitored, researchers can appropriately weigh the data and account for different exposure times in their analyses. This leads to a more precise assessment of associations between variables and outcomes of interest [30].

Another crucial aspect of the study design is the application of review techniques. By appropriately processing data from participants who have not experienced the event of interest at the end of the study [30], the results remain unbiased and reflect valid event rates in the population. Review techniques also enable researchers to extend the duration of the research and capture long-term results, providing a more complete understanding of the phenomenon being investigated.

In summary, this study adopted meticulous data collection methods and rigorous research design to ensure the reliability and validity of mortality data. The study accurately and comprehensively captured vital status and date of death by using information from officially issued death certificates [31] [32], next of kin, and local neighborhood committees [32] [27]. The calculation of the follow-up duration allowed for a precise estimate of exposure time, thereby increasing the accuracy of the analysis. Additionally, review techniques appropriately considered individuals who had yet to experience the event of interest, providing unbiased and meaningful results. Together, these methodological factors contribute to the robustness and completeness of the study, ultimately yielding valuable insights into factors that influence mortality outcomes.

2.3 Exposure

The data collection process consisted of face-to-face interviews facilitated by trained interviewers who were key local staff members in the Chinese county-level

network [33] system operated by the National Bureau of Statistics [22]. To ensure the credibility and expertise of the interviewers, they all had at least 12 years of formal education, with a significant number of them successfully obtaining university degrees, which ensured a high level of proficiency and reliability during the data collection process.

Researchers adopted a strategic approach to increase the comprehensiveness of the health-related aspects examined during the interviews. Each interviewer was accompanied by a local doctor, nurse, or medical student [34] who was fully trained in conducting health examinations. This approach can incorporate additional health-related measures and assessments, enriching the dataset with a broader range of health-related information.

During the physical examination phase of the interview [35], skilled medical staff took specific biometric measurements following a standardized protocol, in which weight and height measurements (two critical indicators of health and well-being) were carefully recorded. Medical staff strictly adhere to standardized procedures to ensure the accuracy and consistency of collected physical data and enhance the reliability of research results. The questionnaire and data of CLHLS can be obtained through `https://opendata.pku.edu.cn/dataset.XHTML?persistentId=doi:10.18170/DVN/WBO7LK&version=2.0` [36].

We use an unambiguous classification system when analyzing health-related behaviors such as smoking and drinking [37]. Based on two specific questions, participants' smoking status was classified as a current or non-current smoker, and alcohol consumption was classified as a current or non-current drinker. This classification method reflects participants' smoking and drinking habits and facilitates a thorough investigation of their potential health effects.

The primary focus of our study was to examine smoking or drinking status as the primary exposure variable. Firstly, we identified smoking and drinking status as current or non-current smoking or drinking via a questionnaire. Secondly, by isolating and examining these behavioral aspects, we aimed to establish Significant correlations between them and various health-related outcomes. Finally, the comprehensive dataset obtained through a rigorous data collection process allows us to

explore potential associations between smoking or alcohol consumption and other health factors in an accurate and evidence-based way that leads to the conclusion.

2.4 Effect modifiers

In this study, we selected candidate subgroup variables as variables that might be evaluated in a population-based survey. My analysis included a total of 48 different variables, carefully chosen to provide a comprehensive understanding of the characteristics of the study participants, as detailed in Table 2.1.

In Table 2.1, lots of mediators, which were measured in the study, were listed, including a variety of demographics, lifestyle factors, health indicators, and disease history [38]. Those mediators are divided into two catalogs: 1) basic demographics and lifestyles, the variables adopted in the CLHLS including age, gender, body weight, etc. 2) Disease history and health indicators: the variables were measured by the medical records, healthcare information system, and self-reported questionnaires. All those variables were included in the analysis as the mediator in the model.

Table 2.1: Mediators of this study

Basic demographics and lifestyle
Age, gender, ethnicity, co-residence, years of schooling, marital status, residence, occupation before retirement, physical activity, consumption of fruit, vegetables, meat, fish, egg, bean, salt-preserved vegetables, sugar, tea, garlic, height.
Disease history and health indicators
Times of suffering from serious illness in the past two years, including hypertension, diabetes, heart disease, stroke, bronchitis, emphysema, pneumonia, asthma, tuberculosis, cataracts, glaucoma, cancer, ulcer, Parkinson's disease, bedsore, arthritis, systolic and diastolic blood pressure, BMI, self-reported health, self-reported life satisfaction, activity of daily living, social and leisure activities, MMSE score, psychological well-being.

The first set of variables captures essential demographic information, including variables such as age, gender, and race. These essential characteristics play a crucial role in https://www.bizmanualz.com/library/what-does-control-group-mea n shaping an individual's health status and are often examined in epidemiological studies to discern potential associations with health outcomes. By including these variables, I aimed to account for demographic effects that may confound or alter the relationships under investigation.

The second set of variables measured statistical information about lifestyle factors, including variables such as diet and physical activity. The data can provide insight into participant behaviors that directly impact health, so exploring these lifestyle aspects can provide valuable information about participants' overall health and potential risk factors for specific health conditions.

The third set of variables incorporates statistical information on health indicators, including variables such as selected activities of daily living (ADL), leisure activities, blood pressure, and cognitive function [39]. To comprehensively assess participants' physical and cognitive health, these metrics provide a more detailed understanding of participants' functional abilities and physiological status, thereby revealing underlying mechanisms behind the associations of interest.

Finally, a set of variables focused on disease history was included in the analysis, including measures of disease variables such as hypertension, diabetes, and stroke to account for preexisting health conditions that may influence the findings of this study. Illness history provides insight into participants' past health experiences and highlights potential vulnerability or protective factors that influence the relationships under study.

In summary, selecting candidate subgroup variables is carefully designed to capture the full range of factors [40] relevant to the study objectives. By incorporating demographics, lifestyle, health indicators, and disease history, our analysis comprehensively explains the complex interactions between these variables and the health outcomes examined. These carefully selected variables are a vital component of our study, allowing us to draw reliable and meaningful conclusions and helping us explore the complex factors that influence population health.

2.5 Imputation technique

The overall percentage of missing data stands at 4.7%. To address this issue, we have chosen to utilize the missForest algorithm to imputation missing data within the CLHLS dataset, as introduced by Stekhoven and Buhlmann in 2012.

The missForest algorithm, known for its nonparametric imputation technique, involves the construction of a dedicated random forest model for each variable, iteratively refining the imputed values. Extensive empirical evidence has indicated the superior performance of this method compared to numerous other imputation techniques, particularly in data scenarios characterized by intricate interactions and non-linear relationships.

While Bayesian methods, including BART, Bayesian trees, and, in general Bayesian statistics, are very adept at dealing with missing values and inherently possess mechanisms to handle missing data, this technique has been chosen over other imputation methods, was motivated by several practical considerations that enhance the robustness and efficiency of our analytical approach.

Firstly, missForest is well-regarded for its robustness and accuracy in imputing missing values in datasets with complex interactions and non-linear relationships, which is consistent with the characteristics of our dataset. Random forests, the foundation of missForest, are adept at capturing non-linear relationships and intricate interactions among variables. This ability ensures that the imputed values are reflective of the underlying data patterns, thereby preserving the integrity and inherent structure of the dataset.

Secondly, Unlike parametric methods that assume a specific distribution for the data, missForest is nonparametric, meaning it does not make strong assumptions about the underlying data distribution. This flexibility is crucial for the diverse and potentially non-normally distributed variables in the CLHLS dataset, such as sociodemographic characteristics, lifestyle behaviours, and health indicators.

Thirdly, The computational efficiency of missForest makes it practical for large datasets, allowing for quick and effective handling of missing values without extensive computational resources.This efficiency ensures timely imputation without sacrificing accuracy. Given the large size of the CLHLS dataset, the algorithm's abil-

ity to quickly and effectively handle missing values without extensive computational resources was a practical consideration.

Lastly, separating the imputation process from the subsequent analysis phase allows for clearer methodological transparency and reproducibility. By using missForest for imputation, we ensure that the missing data is handled consistently and independently of the BART modelling process. This separation allows for an independent evaluation of the imputation process, ensuring it does not introduce bias into the subsequent analysis.

In addition, implementing missForest as a preprocessing step simplifies the overall modelling process. BART is known for its powerful capability to model complex, non-linear interactions and its effectiveness in estimating heterogeneous treatment effects (HTE). However, directly incorporating missing data handling within BART could complicate the model structure and increase the computational burden. By addressing missing values before applying BART, we reduce the complexity of the BART model and improve computational efficiency.

In conclusion, the decision to use missForest for imputation, followed by BART for modelling, is driven by their complementary strengths in handling the specific challenges of the study. missForest provided a robust and accurate imputation of missing data, ensuring a complete and reliable dataset for analysis. BART, with its ability to model complex interactions and improve statistical power, was well-suited for analyzing the heterogeneous effects of smoking and drinking on mortality. Together, these methods maintained methodological rigour while optimizing computational efficiency and ensured a rigorous and thorough analysis, leading to meaningful and actionable insights.

Given a dataset ($\mathbf{X}$) with (n) observations and (p) variables, the goal is to impute missing values and obtain a complete dataset ($\mathbf{X}^*$). The missForest algorithm accomplishes this through the following steps:

- Initialization: For each variable with missing values, the algorithm imputes the missing values with the mean of the observed values for that variable, serving as an initial imputation.

- Iterative imputation: The algorithm iteratively refines the initial imputations

using a random forest model. A random forest is trained at each iteration using the observed values as the response variable and the other variables (including the imputed values) as predictors [41]. The trained random forest is then used to predict the missing values in each variable [41]. This process is repeated for a specified number of iterations or until convergence.

In this study, the random forest model used in the missForest algorithm is a regression forest, where each tree predicts the missing values for a specific variable. The predictions are obtained by averaging the predictions from multiple trees [6].

CHAPTER 3

Methods and Outcome Estimation

This chapter provided outline details and a methodology section, focusing on outcome estimation through advanced statistical techniques, including several parts, "Bayesian Tree-Based Methods" is introduced, centering on "Bayesian Additive Regression Trees (BART)," a machine learning algorithm for non-linear modeling in a Bayesian framework. "Rationale for Using BART Methodology" justifies the choice of BART for the research, "Analytical approach" section that describes the practical implementation of BART in analyzing the data, "Convergence diagnostics for outcomes" describes the validation methodology of the reliability of the BART model outputs to ensure robust and reliable predictive modeling in research.

In recent years buzzwords such as "machine learning," "data science," "big data," "artificial intelligence," and "deep learning" have come to permeate many scientific disciplines. Fundamental to these terms is the concept of predictive modeling, which dates back to the Finetti's [42] work on exchangeability. The goal of predictive modeling is to develop a statistical model [43] using observed data that will generalize well to future or yet unseen observations. The outcome variable, is denoted as y and predictor is denoted as x. Examples of predictive modeling come from a broad range of applications [44]:

- Predict whether a patient will develop coronary artery disease based on age, sex, BMI, diet history, blood pressure, and serum cholesterol.
- Predict the survival probability of women with breast cancer based on demographics, cancer treatment history, and genes.
- Predict how much snow will fall in Iowa City, IA, in January 2020 from historical weather patterns.
- Predict the price of a house using total square footage, number of bedrooms, number of bathrooms, location, whether there is a basement [45], whether it is one-story vs. two-story, etc.

3.1 Tree-Based Methods for Predictive Modeling

Predictive modeling focuses on the conditional probability distribution $\mathbb{P}(Y|x)$ and its conditional mean $\mathbb{E}[Y|x]$. Predictions $\tilde{y}$ are made by minimizing the error in observed data $(x_i, y_i)_{i=1}^n$ using a loss function $L(y, f(\mathbf{x}))$, where f maps predictors to outcomes. For quantitative responses, the common loss function is squared error $L(Y, f(\mathbf{x})) = (Y - f(\mathbf{x}))^2$, minimized when [44]

$$f(\mathbf{x}) = \min_c \mathbb{E}_{Y|\mathbf{x}} \left((Y - c)^2 | \mathbf{X} = \mathbf{x} \right). \tag{3.1.1}$$

where $f(x)$ is the function mapping predictors to outcomes, c is a constant minimizing the squared error, E is the expectation operator, Y is the outcome variable, and X is the predictor variable.

Approximating the unknown function f varies from Generalized Linear Models (GLMs) [46] [47] to neural networks, with tree-based models being popular.

Traditional linear regression models are often inadequate due to non-linear or non-additive effects.

$$Y_i = \beta_0 + x_{i1}\beta_1 + \cdots + x_{ip}\beta_p + \varepsilon_i, \quad \varepsilon_i \sim N(0, \sigma^2) \tag{3.1.2}$$

where Y_i is the response variable for the ith observation, β_0 is the intercept term,

x_{ij} is the jth predictor for the ith observation, β_j is coefficient for the jth predictor, ϵ_i is error term, and $N(0, \sigma^2)$ is normal distribution with mean 0 and variance σ^2.

This leads to employing non-linear functions of predictors, where *basis* functions transform x for linear models. With M basis functions $g_m(x) : \mathbb{R}^p \mapsto \mathbb{R}$, the model becomes [44]

$$Y = \sum_{m=1}^{M} \beta_m g_m(\mathbf{x}) + \varepsilon, \quad \varepsilon \sim N(0, \sigma^2). \tag{3.1.3}$$

where$Y$ is the response variable, β_m is the coefficient for the mth basis function, $g_m(x)$ is mth basis function, and ϵ is the error term.

Tree-based methods, another alternative, model Y as a function of x without assuming a specific functional form. Binary decision trees partition the predictor space into hyper-rectangles, and within each partition, a simple model is fit. The fitted function of such a tree is

$$\hat{f}(\mathbf{x}) = \sum_{l=1}^{L} \mu_l \mathbb{I}(\mathbf{x} \in R_l). \tag{3.1.4}$$

where $\hat{f}(x)$ is fitted function of the tree, μ_l is mean response in region R_l, $\mathbb{I}(\mathbf{x} \in R_l)$ is indicator function, which is 1 if $x \in R_l$, otherwise 0, and L is the Number of regions/partitions.

However, individual trees can be unstable, leading to ensemble methods like bagged trees, random forests, and gradient boosted trees, which have become popular for their robustness.

3.1.1 Bayesian Tree-Based Methods

Bayesian tree-based methods differ by positing a joint probability model for observed data and parameters, leading to a posterior distribution $p(\mathcal{T}, \theta|\mathbf{y})$ used for prediction and inference. These methods develop priors for the decision tree structure and use MCMC algorithms for sampling. Chipman [48] introduced a Bayesian CART model with independent Bernoulli nodes and simple rules for tree modification.

The seminal work by Chipman in 1998 [48] introduced the Bayesian CART model, which has since inspired subsequent Bayesian tree-based models. These models aim to flexibly estimate the conditional mean given a vector of predictors $\mathbf{x} = (x_1, ..., x_{L_p})^\top$. Bayesian CART explores various decision trees to identify those that likely explain the variation in Y, For continuous responses, this aligns with the regression framework

$$Y = f(\mathbf{x}) + \epsilon, \epsilon \sim N\left(0, \sigma^2\right). \tag{3.2}$$

Here, the conditional mean $f(\mathbf{x})$=$\mathbb{E}[Y|x]$ is approximated using a binary decision tree $\mathcal{T}$.

Bayesian CART Metropolis-Hastings Algorithm

The Bayesian CART Metropolis-Hastings algorithm samples from the posterior distribution of the decision tree.tree [49]:

$$p(T|\mathbf{X}, \mathbf{y}) \propto p(\mathbf{y}|\mathbf{X}, \mathcal{T})\pi(\mathcal{T}). \tag{3.1.5}$$

where $p(T|\mathbf{X}, \mathbf{y})$ is posterior distribution of the tree, $p(\mathbf{y}|\mathbf{X}, \mathcal{T})$ is likelihood of the data given the tree, $\pi(\mathcal{T})$ is the prior distribution of the tree.

Key steps include generating a candidate tree and accepting it based on a calculated probability. The transition kernel q defines four rules: Birth, Death, Change, and Swap, which modify the tree's structure. The algorithm starts with a single-leaf tree, exploring various tree structures to determine the posterior distribution effectively.

3.1.2 Bayesian Additive Regression Trees (BART)

The development of Bayesian CART [48] was followed by the development of tree-based ensembles such as random forests [28] and gradient-boosted trees [50]. Borrowing ideas from these ensemble methods, Bayesian CART was extended by Chipman [48] to incorporate an ensemble (or sum) of decision trees rather than just

a single decision tree. Their model is referred to as Bayesian Additive Regression Trees. The key difference between Bayesian CART and BART is that the conditional mean is approximated with a sum of m decision trees [44].

$$\mathbb{E}[Y|\mathbf{x}] \approx \sum_{t=1}^{m} g(\mathbf{x}; \mathcal{T}_t, \mu_t) \tag{3.1.6}$$

where the function $g(\cdot)$ denotes a single decision tree parameterized by its structure $\mathcal{T}_i$ and vector of leaf node means $\mu_t = (\mu_{1t}, ..., \mu_{L_t t})^\top$, Y represents the response variable, $\mathbf{x}$ represents the predictor variables, m is the Number of trees in the ensemble, $\mathcal{T}_t$ is the structure of the t-th decision tree, μ_t is the vector of leaf node means for the t-th tree.

BART seeks to solve the same standard regression problem as described by 3.2. However, the regression function is comprised of m multivariate components (i.e., decision trees) that are constrained to be weak learners analogous to the methods of boosted decision trees and random forests.

Since BART inherently fits an additive model, the decision trees can be viewed as dimensionally adaptive basis functions. As such, the BART framework is a Bayesian nonparametric regression model. Whereas boosting uses a sequence of trees to iteratively fit residual variation in the response unexplained by the previous trees, BART jointly fits the tree ensemble with each tree approximating a small portion of the true regression function $f(\mathbf{x}) = \mathrm{E}[Y|\mathbf{x}]$.

BART employs the priors on each $\mathcal{T}_t$ in the ensemble and an iterative MCMC algorithm, referred to as Bayesian back fitting MCMC [51] [52], is used to obtain posterior draws of each tree in the ensemble and their corresponding leaf node parameters μ_t. Consequently, posterior draws of the re gression function $f(\mathbf{x}) = \mathbb{E}[Y|\mathbf{x}]$ are obtained. Additionally, the posterior mean at x can be obtained by averaging over the posterior draws of $f(\mathbf{x})$ along with pointwise credible intervals computed as quantiles from the posterior draws.

The posterior distribution of BART involves m decision trees, and their corresponding vector of leaf node means $\{\mathcal{T}_t, \mu_t\}_{t=1}^m$. Similar to Bayesian CART, BART

uses priors of the form [44],

$$\begin{aligned}\pi\left(\{\mathcal{T}_t, \mu_t\}_{t=1}^{m}, \sigma\right) &= \pi(\sigma^2) \prod_{t=1}^{m} \pi(\mathcal{T}_t, \mu_t) \\ &= \pi(\sigma^2) \prod_{t=1}^{m} \pi(\mu_t|\mathcal{T}_t)\pi(\mathcal{T}_t) \\ &= \pi(\sigma^2) \prod_{t=1}^{m} \prod_{l=1}^{L_t} \pi(\mu_{lt}|\mathcal{T}_t)\pi(\mathcal{T}_t).\end{aligned} \tag{3.1.7}$$

where $\pi(\cdot)$ denotes a prior distribution, σ^2 is the variance parameter for the noise in the model, L_t is the Number of leaf nodes in the t-th tree, and μ_{lt} is the mean of the l-th leaf node in the t-th tree. This formula represents the joint prior distribution of the parameters in the BART model, including the decision tree structures $\mathcal{T}_t$, the leaf node means μ_t, and the variance σ^2.

By using a prior of the form in 3.1.7, it is assumed that all tree structures in the ensemble are independent, $\mathcal{T}_t \perp\!\!\!\perp \mathcal{T}_{t'}$ and all leaf nodes in the ensemble are independent (within and between decision trees). This aids in the posterior computation and eliminates the need for a reversible jump MCMC sampler since the dimension of μ_t can change from iteration to iteration. Furthermore, simple analytic conjugate priors can be used for $\mu_t|\mathcal{T}_t$ and σ^2. The BART model is summarized as [44].

$$\text{Likelihood}: \quad Y_i|\mathbf{x}_i, \Theta = \left(\{\mathcal{T}_t, \mu_t\}_{t=1}^{m}, \sigma\right) \sim N\left(\sum_{t=1}^{m} g_t(x; \mathcal{T}_t, \mu_t), \sigma^2\right) \tag{3.1.8}$$

where Y_i is the response variable for the i-th observation, x_i is the predictor variables for the i-th observation, Θ is the set of all parameters, including tree structures, leaf node means, and variance, $N(\mu, \sigma^2)$ is the Normal distribution with mean μ and variance σ^2.

$$\text{Prior}: \quad \pi(\Theta) = \pi(\sigma^2) \prod_{t=1}^{m} \prod_{l=1}^{L_t} \pi(\mu_{lt}|\mathcal{T}_t)\pi(\mathcal{T}_t), \tag{3.1.9}$$

This formula represents the prior distribution over all model parameters Θ, which includes the variance σ^2, the leaf node means μ_{lt}, and the tree structures T_t. where Θ is the set of all model parameters.

3.2 MCMC within BART framework

3.2.1 Exploring Markov Chain Monte Carlo in Bayesian Context

In the realm of Bayesian analysis, when working with data X and a set of parameters θ, the posterior distribution, represented as

$$p(\theta \mid X) = \frac{p(X \mid \theta)P(\theta)}{\int p(X \mid \theta)P(\theta)\mathrm{d}\theta} \tag{3.2.1}$$

is typically elusive in its analytical form [53]. This necessitates alternative inferential strategies. One such strategy is the utilization of MCMC methods. These methods simulate a Markov chain whose equilibrium distribution aligns with the sought-after posterior, $P(\theta \mid X)$. This approach is effective due to the inherent characteristics of ergodic Markov chains, where the probability distribution over the states stabilizes irrespective of the starting point. In essence, navigating through such a chain and noting its states over time parallels drawing samples from its equilibrium distribution [54].

The key to achieving ergodicity in a Markov chain is ensuring that it is both irreducible (every state is accessible from every other state) and aperiodic (states are not revisited at predictable intervals) [55]. But the central question remains: How can we ensure that the underlying Markov chain's equilibrium distribution matches our desired posterior? The answer lies in satisfying the reversibility or detailed balance condition [56]. This condition mandates that the probability of transitioning from any state to another is equal to the probability of the reverse transition. Denoting the equilibrium distribution as $\pi(\cdot)$ and the transition kernel (the function calculating the probability of moving from state θ to $\theta^\star$) as $K(\theta^\star \mid \theta)$, the detailed balance condition is formalized as [4].

$$K(\theta^\star \mid \theta)\pi(\theta) = K(\theta \mid \theta^\star)\pi(\theta^\star). \tag{3.2.2}$$

The reason this work is discussed in previous writings on Metropolis-Hastings

or in [57] [57]. Conceptually, this can be thought of as ensuring that the proposed transition kernel is unaffected by direction or time changes [58]; what matters is the probability of occupying a specific state, as defined by $\pi(\cdot)$. In MCMC practice, we propose a distribution $\pi(\cdot)$ and a kernel $K(\cdot \mid \cdot)$ that meet the aforementioned conditions, thereby ensuring that we are traversing an ergodic Markov chain whose equilibrium distribution is $\pi(\cdot)$.

3.2.2 Metropolis-Hastings Algorithm Elucidation

The Metropolis-Hastings algorithm, a seminal approach in computational statistics, aims to generate a sequence of states aligned with a target distribution $P(x)$. This objective is achieved through a Markov process that ultimately converges to a unique stationary distribution $\pi(x)$, satisfying $\pi(x) = P(x)$ as indicated in [59].

The essence of a Markov process lies in its transition probabilities $P(x' \mid x)$, defining the likelihood of moving from a current state x to a new state x'. For the process to have a distinct stationary distribution $\pi(x)$, two key conditions must be met [59]:

1. Stationary Distribution Existence: The system must possess a stationary distribution $\pi(x)$. Detailed balance is a sufficient (but not necessary) condition for this `https://www.environmentalistsforeurope.org/what-is-a-non-equilibrium-steady-state/`, requiring that transitions between any pair of states x and x' are reversible. Mathematically, this means $\pi(x)P(x' \mid x) = \pi(x')P(x \mid x')$.

2. Stationary Distribution Uniqueness: The stationary distribution $\pi(x)$ must be singular. This uniqueness is assured by the Markov process's ergodicity, necessitating each state to be (1) aperiodic—lacking fixed interval returns to the same state; and (2) positive recurrent—having a finite expected return time to the same state.

The Metropolis-Hastings algorithm constructs a Markov process with transition probabilities designed to satisfy these conditions, thereby aligning its stationary

distribution $\pi(x)$ with the target $P(x)$. The algorithm's formulation begins with the detailed balance condition:

$$P(x' \mid x)P(x) = P(x \mid x')P(x'), \tag{3.2.3}$$

reformulated as

$$\frac{P(x' \mid x)}{P(x \mid x')} = \frac{P(x')}{P(x)}. \tag{3.2.4}$$

The process entails bifurcating the transition into two stages: proposal and acceptance-rejection. The proposal distribution $g(x' \mid x)$ specifies the conditional probability of proposing state x' from x, while the acceptance function $A(x', x)$ dictates the probability of accepting the proposed state x'. Hence, the overall transition probability is expressed as the product [44]:

$$P(x' \mid x) = g(x' \mid x)A(x', x). \tag{3.2.5}$$

Inserting this into the earlier align, we obtain [60]:

$$\frac{A(x', x)}{A(x, x')} = \frac{P(x')}{P(x)} \frac{g(x \mid x')}{g(x' \mid x)}. \tag{3.2.6}$$

subsequently, an acceptance ratio is selected that satisfies the above condition. A typical choice is the Metropolis criterion [44]:

$$A(x', x) = \min\left(1, \frac{P(x')}{P(x)} \frac{g(x \mid x')}{g(x' \mid x)}\right). \tag{3.2.7}$$

with this Metropolis acceptance function A, either $A(x', x)$ or $A(x, x')$ equals 1, fulfilling the detailed balance.

The Metropolis-Hastings algorithm procedure is as follows [44]:

1. Initialization

    1. Select an initial state x_0.

    2. Initialize counter $t = 0$.

2. Iteration

    1. Propose a new state x' based on $g(x' \mid x_t)$.
    2. Compute acceptance probability $A(x', x_t) = \min\left(1, \frac{P(x')}{P(x_t)} \frac{g(x_t|x')}{g(x'|x_t)}\right)$.
    3. Decide acceptance or rejection:
        1. Generate a random number u uniformly distributed in [0,1];
        2. If $u \leq A(x', x_t)$, accept and set $x_{t+1} = x'$;
        3. If $u > A(x', x_t)$, reject and retain $x_{t+1} = x_t$.
    4. Increment: set $t = t + 1$.

3.2.3 Formal derivation of the Metropolis-Hastings algorithm

The Metropolis-Hastings algorithm is designed for generating a sequence of states that conform to a specific distribution $P(x)$. It leverages a Markov process which, over time, converges to a unique stationary distribution $\pi(x)$, where $\pi(x) = P(x)$ as referenced in [59].

A Markov process is characterized by its transition probabilities $P(x' \mid x)$, which determine the likelihood of moving from a state x to another state x'. For $\pi(x)$ to be the unique stationary distribution of the process, two primary conditions must be satisfied: [59]

1. Stationary Distribution Existence: The process should have a stationary distribution $\pi(x)$. The detailed balance is an adequate but not mandatory condition for this, stating that for every state pair x, x', the probability of transitioning from x to x' should be equal to the probability of transitioning from x' to x [61], i.e., $\pi(x)P(x' \mid x) = \pi(x')P(x \mid x')$.

2. Stationary Distribution Uniqueness: The stationary distribution $\pi(x)$ needs to be unique, ensured by the ergodic nature of the Markov process. This requires each state to be (1) aperiodic, meaning the process doesn't revisit the same state at predictable intervals, and (2) positive recurrent, implying a finite expected return time to each state [62].

The Metropolis-Hastings algorithm constructs a Markov process by defining transition probabilities that meet the above conditions, aligning $\pi(x)$ with the intended distribution $P(x)$. The derivation begins with the detailed balance condition [44]:

$$P(x' \mid x)P(x) = P(x \mid x')P(x'), \tag{3.2.8}$$

which can be rephrased as [44]:

$$\frac{P(x' \mid x)}{P(x \mid x')} = \frac{P(x')}{P(x)}. \tag{3.2.9}$$

The transition involves two phases: proposing a new state and deciding on its acceptance. The proposal distribution $g(x' \mid x)$ indicates the likelihood of suggesting a state x' given the current state x, while the acceptance probability $A(x,' x)$ determines the chance of accepting x'. The transition probability is their product [63]:

$$P(x' \mid x) = g(x' \mid x)A(x', x). \tag{3.2.10}$$

integrating this into the previous equation, we get:

$$\frac{A(x', x)}{A(x, x')} = \frac{P(x')}{P(x)} \frac{g(x \mid x')}{g(x' \mid x)}. \tag{3.2.11}$$

choosing an acceptance ratio that satisfies this relation is crucial. The common choice is the Metropolis criterion [64]:

$$A(x', x) = \min\left(1, \frac{P(x')}{P(x)} \frac{g(x \mid x')}{g(x' \mid x)}\right). \tag{3.2.12}$$

with this criterion, $A(x', x) = 1$ or $A(x, x') = 1$, ensuring compliance with the condition.

The algorithmic steps of Metropolis-Hastings are as follows [44]:

1. Initialization

    1. Choose an initial state x_0.

    2. Initialize iteration counter $t = 0$.

2. Iterative Process

    1. Propose a new state x' as per $g(x' \mid x_t)$.
    2. Compute the acceptance probability $A(x', x_t) = \min\left(1, \frac{P(x')}{P(x_t)} \frac{g(x_t|x')}{g(x'|x_t)}\right)$.
    3. Decide to accept or reject:
        1. Generate a uniform random number u in the range [0, 1].
        2. If $u \leq A(x', x_t)$, accept and set $x_{t+1} = x'$.
        3. If $u > A(x', x_t)$, reject and retain $x_{t+1} = x_t$.
    4. Increment the counter by setting $t = t + 1$.

Under these conditions, the empirical distribution of the collected states $x_0, \ldots, x_T$ will converge to $P(x)$. The number of iterations T necessary for an effective estimation of $P(x)$ depends on various factors, including the relationship between $P(x)$ and the proposal distribution g, and the desired accuracy of the estimation [65]. For discrete state spaces, the iteration count should approximate the autocorrelation time of the Markov process [66].

It is crucial to note that the optimal choice of the proposal distribution $g(x' \mid x)$ and the required number of iterations is not predetermined in general problems; they are method parameters that must be tailored to the specifics of each case [67].

3.2.4 Gibbs Sampling Methodology

Gibbs sampling represents a unique instance of the Metropolis-Hastings algorithm where every proposed state is accepted with a probability of one [68]. Understanding this is straightforward upon grasping the algorithm. Imagine a multidimensional posterior distribution with parameters $\theta = (\theta_1, \ldots, \theta_D)$. The core principle of Gibbs sampling involves iteratively sampling from the conditional distribution $P(\theta_d \mid X, \theta_{-d})$, where θ_{-d} denotes the set of all parameters except the dth:

Gibbs Sampling Procedure: For each iteration $t = 1, \ldots, T$, execute [69]

$$\begin{aligned}
\theta_1^{(t+1)} &:= \theta_1^\star \sim P(\theta_1^{(t)} \mid X, \theta_2^{(t)}, \theta_3^{(t)}, \ldots, \theta_D^{(t)}) \\
\theta_2^{(t+1)} &:= \theta_2^\star \sim P(\theta_2^{(t)} \mid X, \theta_1^{(t+1)}, \theta_3^{(t)}, \ldots, \theta_D^{(t)}) \\
&\vdots \\
\theta_D^{(t+1)} &:= \theta_D^\star \sim P(\theta_D^{(t)} \mid X, \theta_1^{(t+1)}, \theta_2^{(t+1)}, \ldots, \theta_{D-1}^{(t+1)})
\end{aligned} \tag{3.2.13}$$

To comprehend its effectiveness, note that

$$P(\theta \mid X) = P(\theta_d, \theta_{-d} \mid X) = P(\theta_d \mid X, \theta_{-d}) P(\theta_{-d} \mid X). \tag{3.2.14}$$

Disregarding the iteration labels, the transition probability is given by [70]:

$$\begin{aligned}
\alpha(\theta^\star \mid \theta) &= \min\left\{1, \frac{P(\theta^\star \mid X) P(\theta_d \mid X, \theta_{-d})}{P(\theta \mid X) P(\theta_d^\star \mid X, \theta_{-d}^\star)}\right\} \\
&= 1.
\end{aligned} \tag{3.2.15}$$

In (3.2.15), the terms effectively cancel out due to the unique characteristic of Gibbs sampling, where $\theta_{-d}^\star = \theta_{-d}$. Thus, each Gibbs sampling step can be seen as a Metropolis-Hastings walk with guaranteed state acceptance.

The primary advantage of Gibbs sampling is its guaranteed acceptance of proposals, but it requires the ability to compute the specific conditional distributions mentioned above. This approach is feasible when $P(\theta_d)$ is conjugate to the posterior.

3.3 Rationale for Using BART Methodology

The concept of Heterogeneous Treatment Effects (HTE) indicates that within a cohort, individual responses to a given treatment can vary markedly. In epidemiological and clinical studies, the focus is often on assessing average effects, which might mask the nuances of individual responses. Traditional methods examining HTE typically analyze individual characteristics in isolation, potentially overlooking the complexity of interactions between these characteristics. This approach can lead to a lack of statistical power and a failure to recognize synergistic effects, resulting in

an incomplete understanding of treatment responses.

Bayesian Additive Regression Trees (BART) offer a compelling solution to these challenges. BART inherently automates the discovery of nonlinear relationships and interactions, prioritizing these based on their significance. This automation reduces the risk of model misspecification and inherent bias, which is common in traditional interaction testing.

Moreover, BART's integration into the counterfactual framework for investigating HTE enables the estimation of conditional average treatment effects based on various covariates. BART's efficacy has been demonstrated in extensive simulation studies, outperforming competing methods. This makes it an ideal tool for exploring HTE, aiding in hypothesis generation and informing future confirmatory analyses in trials.

In the current study, BART is employed to examine multivariable HTE and estimate conditional average effects within the Comprehensive Longitudinal Health and Longevity Study (CLHLS). This application underscores BART's robustness in unraveling the complexities of HTE, illuminating the intricate interplay of variables influencing treatment responses within the study.

The study employed the missForest algorithm for imputing missing data, ensuring robust handling of non-random missingness. Bayesian Additive Regression Trees (BART) were then utilized to model the complex interactions between smoking, drinking, and mortality. The BART approach allowed for detecting non-linear relationships and interactions among variables, improving the analysis's statistical power and depth. The iterative MCMC algorithm used within the BART framework refined the model iteratively, providing posterior draws and credible intervals for the estimated effects.

The benefits of using this combined methodology are significant, Firstly, the missForest algorithm provides high-quality imputation, maintaining the data's integrity and variability. BART's ability to model complex interactions offers a comprehensive understanding of how smoking and drinking behaviours influence mortality across different subgroups. Secondly, by handling the imputation separately, we ensured that the BART model could focus on identifying and modelling the intricate interac-

tions between smoking, drinking, and mortality without the additional complexity of simultaneously imputing missing values. In conclusion, this comprehensive methodology leveraged the strengths of both techniques and enabled the identification of significant variations in mortality risk across different subgroups, highlighting the nuanced impacts of smoking and drinking behaviours on health outcomes.

3.4 Analytical Approach

This study utilizes a binarized mortality outcome as the binary mortality outcome and assesses Heterogeneous Treatment Effects (HTE) on an absolute scale, specifically evaluating the mean difference in mortality over five years. The analysis unfolds in two stages: first, BART models estimate conditional average treatment effects based on covariates. Then, the logit BART model analyzes the binary mortality outcome. The second stage adopts a "fit-the-fit" approach, applying the estimated effects as dependent variables in classification tree models to identify differential effects in covariate-defined subgroups, with a maximum tree depth of three for interpretability.

BART models, typically fitted with 200 trees and specific hyperparameters, underwent extensive hyperparameter evaluation through 10-fold cross-validation. This included exploring combinations of power priors (1, 2, or 3), base priors (0.25, 0.5, or 0.95), and tree numbers (50, 200, or 400), with model diagnostics confirming robust performance.

3.4.1 Causal inference framework

The goal of a trial is to estimate the effect of an intervention, denoted as Z, on an outcome, Y. Under the counterfactual outcome framework, we assume individual i has two potential outcomes: the outcome $Y_i(1)$ we would observe under intervention ($Z = 1$), and the outcome $Y_i(0)$ we would observe under control ($Z = 0$).

Treatment Z has a causal effect for participant i if $Y_i(1) \neq Y_i(0)$ (i.e., the potential outcomes differ under intervention vs. control). Since individual i cannot experience both potential outcomes, and only one of the two potential outcomes

can be observed for each individual in a specific trial, it cannot calculate the causal effect for participant i with the observed data.

However, the randomization allows estimation of the Average Treatment Effect (ATE) [71] across individuals without confounding. The Average Treatment Effect (ATE) is a measure used to evaluate the effect of a treatment across the entire population. It is defined as the difference in the expected outcomes between the treated and untreated groups. The ATE provides a population-level estimate of the treatment effect, which helps in understanding the general impact of the treatment across the entire study sample.

$$\Delta ATE = E[Y_i(1) - Y_i(0)] \tag{3.4.1}$$

where $Y_i(1)$ is the potential outcome for individual i if they receive the treatment, $Y_i(0)$ is the potential outcome for individual i if they do not receive the treatment, $Y_i(1) - Y_i(0)$ is the individual treatment effect for individual i, representing the difference in outcomes between receiving and not receiving the treatment.

In our studies, ΔATE represents the average effect of smoking or drinking on mortality across all individuals in the study. This measures the overall impact of the treatment (smoking or drinking) on the outcome (mortality).

Beyond an overall summary, the treatment effect measure for each specific participant with baseline covariate vector x (the Conditional Average Treatment Effect, CATE) are also estimable from the observed data [72]. CATE can measure the average effect of a treatment for specific subpopulations defined by certain characteristics or covariates. It provides a more granular view of the treatment effect, which can vary across different subgroups.

$$\Delta CATE(x) = E[Y_i(1) - Y_i(0)|X_i = x] \tag{3.4.2}$$

where $\Delta CATE(x)$ is the conditional average treatment effect for individuals with covariates $X_i = x$, X_i is Covariates or characteristics of individual i, x is the specific value of the covariates.

In our studies, $\Delta CATE(x)$ represents the effect of smoking or drinking on mor-

tality for specific subgroups defined by characteristics such as age, gender, health status, etc.

Heterogeneous Treatment Effects (HTE) are inferred if the CATE:

$$\Delta CATE(x_i) = \Delta CATE(x_j) \tag{3.4.3}$$

for two different covariate vectors x_i and x_j.

3.4.2 Technical Details: BART

BART estimates CATEs non-parametrically. It uses an ensemble of binary trees, with prior distributions regularizing the fit. The mean function is approximated by the sum of these trees' outputs [73]. The model parameters include the trees T_j and node values M_j, with the mean function represented as `https://jmloyola.github.io/posts/2019/06/introduction-to-bart`:

$$Y_i = \sum_{j=1}^{m} g\left(x_i; T_j, M_j\right) + \epsilon_i \tag{3.4.4}$$

where Y_i is the observed outcome for individual i, $g(x_i; T_j, M_j)$ is the prediction from the j-th tree for individual i, parameterized by tree structure T_j and terminal node parameters M_j, m is the Number of trees in the BART model, ϵ_i is the Error term, typically assumed to be normally distributed with mean zero and variance σ^2.

3.4.3 Sensitivity analysis

Three sensitivity analyses were performed: 1) using best/worst-case imputation, 2) using worst/best-case imputation, and 3) dropping the participants who died in six months after the baseline survey. These strategies are designed to test the robustness of the results under extreme scenarios `https://cyberinsight.co/what-are-offensive-cyber-security-strategies/`. The details of the best/worst-case and worst/best-case imputation are as follows:

Best/Worst-Case Imputation: In the best/worst-case scenario, the analysis assumes an extreme scenario that is favorable to the smoking or drinking group: All

missing mortality outcome data for participants in the smoking or drinking group are imputed as 'alive at five years' [42]. Conversely, all missing mortality outcome data for participants in the non-smoking or drinking group are imputed as 'dead at five years.' This approach tests the robustness of the results under a scenario that is most favorable to the hypothesis that smoking or drinking does not increase mortality risk. The analysis is repeated with these imputations, and the impact on the results is observed [74].

Worst/Best-Case Imputation: The worst/best-case scenario is the opposite of the best/worst-case: All missing mortality outcome data for participants in the smoking or drinking group are imputed as 'dead at five years.' All missing mortality outcome data for participants in the non-smoking or drinking group are imputed as 'alive at five years' [75]. This scenario tests the strength of the findings under the most unfavorable conditions for the smoking or drinking group [76]. It assesses whether the results still hold when assuming that missing data indicates the worst possible outcome for the smoking or drinking group and the best possible outcome for the non-smoking or drinking group [77].

Mediator Imputation: In both scenarios, after setting the mortality outcomes, the mediators in the dataset (if any) are imputed using multiple imputations by chained equations (MICE), a robust method for handling missing data in epidemiological studies. This step ensures that the analysis accounts for missing information in other relevant variables that could mediate the relationship between smoking or drinking and mortality [70].

Drop the participants who died in six months after baseline research: those participants who died in six months after the baseline were frail and there would be measurement bias within their answers [78].

All BART models were fit using R statistical computing software v. 4.1.226 with the 'BART' package v. 2.924, and all CART models were fit using the 'report' package v. 4.1.1627. The descriptive statistics were performed on STATA v1.8. All statistical code is made available at `https://github.com/TingjiaoCui/HTE_for_smoking_and_drinking`.

3.5 Convergence Diagnostics for Outcomes

Convergence in BART models is evaluated through specific methods for continuous outcomes and probit and Multinomial BART models. The logit BART model uses auxiliary latent variables for convergence diagnostics. Traditional MCMC diagnostic tools, including Geweke's method, are applied to ensure robust convergence monitoring [79].

The Geweke diagnostic, a key tool in Bayesian analysis, assesses the convergence of MCMC chains by segmenting the chain into parts and comparing means using a normality test [80]. Geweke's diagnostic evaluates convergence by comparing measures of two subsequences of a parameter. Given subsequences θ_A and θ_B, the diagnostic uses a Z-score: the difference between the two sample means divided by the estimated standard error. Geweke proposed in 1992 that when the chain is stationary, the means of two subsequences are equal, and Geweke's statistic shows an asymptotically standard normal distribution [81],

$$Z_{AB} = \frac{\bar{\theta}_A - \bar{\theta}_B}{\sqrt{\frac{1}{n_A}\hat{S}_A + \frac{1}{n_B}\hat{S}_B}} \xrightarrow{d} N(0,1) \tag{3.5.1}$$

Here,$\bar{\theta}_A$ and $\bar{\theta}_A$ are the means, $\hat{S}_A$ and $\hat{S}_A$ are the variances, and n_A and n_B are the iteration counts of the subsequences [81].

In conclusion, my study leveraged BART to identify heterogeneous effects, expanding upon the well-established ensemble Bayesian learning method. By focusing on the identification of variables and cut-points that induce significant differences in hazard ratios, my approach provides a nuanced perspective on potential effect modifiers contributing to the heterogeneity within the analyzed association. Stratifying the analysis by age groups acknowledges the impactful role of age in the relationship under investigation, aligning with best practices in epidemiological research [82]. The utilization of R and Stata for algorithm development and statistical analyses, respectively, reflects my commitment to transparency, reproducibility, and robustness in my methodology. Through these methodological choices and considerations, my study contributes valuable insights to the realm of biomedicine, advancing my

understanding of complex associations and enhancing my ability to uncover heterogeneity within them.

CHAPTER 4

Results

In this chapter, I presented the results of the analysis, including four parts: 1) Baseline characters, which describe the demographic information of the cohort; 2) Heterogeneous effect identified by BART, which presents the heterogeneous effect identified by BART and its effect in each subgroup, also some statistical test to validate the model; 3) Subgroup analysis, which use COX regression with the adjustment of residency and age to further validate the effect in each predefined subgroup identified by BART; 4) Sensitivity analysis, which apply different methods to validate the robustness of the model.

4.1 Baseline characters

Table 4.1 presents the baseline characteristics of 43,487 participants [83] categorized into three age groups: 9,202 aged between 65-80 years, 23,732 aged 81-100 years, and 10,553 aged over 100. The table illustrates significant demographic and health-related variations across these age groups, as indicated by the statistical p-value of <0.001 for all measured parameters, suggesting strong evidence of difference across age categories. In terms of gender distribution, females were more prevalent

in the oldest age group, representing 80.3% of those over 100 years old, compared to 53.7% in the 81-100 years group and 47.0% in the 65-80 years group. This reflects a higher female survival rate into advanced age. The participants predominantly resided in rural areas, especially among the oldest, where 60.6% lived in rural settings. Health-related characteristics varied significantly with age.

Smoking or drinking status showed a lower prevalence of current smokers or drinkers in the oldest group (33.6%) compared to those aged 65-80 years (49.8%). Cognitive function, assessed by the Mini-Mental State Examination (MMSE) [84], showed a decrease in mean scores with increasing age, from 27.67 in the youngest group to 13.83 in the oldest. Social activity scores also demonstrated a decrease with age. Blood pressure readings indicated a trend of lower diastolic blood pressure with advancing age. The ability to perform daily activities, as reflected by the sum of the Activity of Daily Living score, showed an increase in dependency with age. Fruit intake was least frequent in the oldest age group, with 30.0% reporting they rarely ate fruit, compared to 26.8% and 41.2% in the 81-100 and 65-80 age groups, respectively.

The comprehensive data highlight the shifting demographic and health-related characteristics within an aging population and underline the importance of age-specific considerations in epidemiological research. Specifically, the smoking or drinking behaviour, significantly varied by age, which may indicate that its heterogeneous effect may vary by age groups.

Table 4.1: Baseline characteristics of the 43487 participants by age group

	Age groups				*p-value*
	65-80 (N=9202)	81-100(N = 23732)	>100(N =10553)	Total(N =43487)	
Age					<0.001
Mean(SD)	70.5(5.2)	89.0(5.5)	101.6(1.9)	88.1(11.5)	
Median(Q1,Q3)	70.0(66.0,75.0)	90.0(84.0,93.0)	101.0(100.0,102.0)	90.0(81.0,99.0)	
Gender					<0.001
Female	4325(47.0)	12752(53.7)	8476(80.3)	25553(58.8)	
Male	4877(53.0)	10980(46.3)	2077(19.7)	17934(41.2)	
Residence					<0.001
Urban	3804(41.3)	10285(43.3)	4157(39.4)	18246(42.0)	
Rural	5398(58.7)	13447(56.7)	6396(60.6)	25241(58.0)	
Smoking or drinking status					<0.001
Current smoking or drinking	4579(49.8)	10868(45.8)	3551(33.6)	18998(43.7)	
Non-current	4623(50.2)	12864(54.2)	7002(66.4)	24489(56.3)	
MMSE score					<0.001
Mean(SD)	27.67(3.62)	22.07(8.54)	13.83(10.21)	21.26(9.51)	
Median(Q1,Q3)	29.0(27.0,30.0)	25.0(19.0,29.0)	15.0(3.0,23.0)	25.0(17.0,29.0)	
Social activity score					<0.001
Mean(SD)	13.52(2.48)	10.53(2.69)	8.75(2.03)	10.73(2.97)	
Median(Q1,Q3)	14.0(12.0,15.0)	10.0(8.0,12.0)	8.0(7.0,10.0)	10.0(8.0,13.0)	
Diastolic blood pressure					<0.001
Mean(SD)	83.92(11.97)	82.71(12.81)	81.55(12.44)	82.69(12.57)	
Median(Q1,Q3)	81.0(78.0,90.0)	80.0(75.0,90.0)	80.0(73.2,90.0)	80.0(75.0,90.0)	
Sum of Activity of daily living					<0.001
Mean(SD)	0.10(0.54)	0.66(1.43)	1.76(2.05)	0.81(1.59)	
Median(Q1,Q3)	0.0(0.0,0.0)	0.0(0.0,1.0)	1.0(0.0,3.0)	0.0(0.0,1.0)	
Fruit intake					<0.001
Rarely	3791(41.2)	6360(26.8)	3169(30.0)	13320(30.6)	
Occasionally	3443(37.4)	10879(45.8)	4399(41.7)	18721(43.0)	
Everyday	1968(21.4)	6493(27.4)	2985(28.3)	11446(26.3)	

4.2 Heterogeneous effect identified by BART

Figure 4.1 shows a decision tree used to illustrate the results of my study on detecting heterogeneous effects; it shows model results for the binary outcome mortality of drinking alcohol or smoking, where percentages indicate the proportion of individuals within each subgroup. The top value in each box is the estimated treatment effect in the subgroup, which has corresponding covariate values. The bottom value in each box is the proportion of the trial sample belonging to the subset. The upper effect is the effect of the binary outcome of drinking and death within five years in this group. The percentage below represents the number of people in this group.

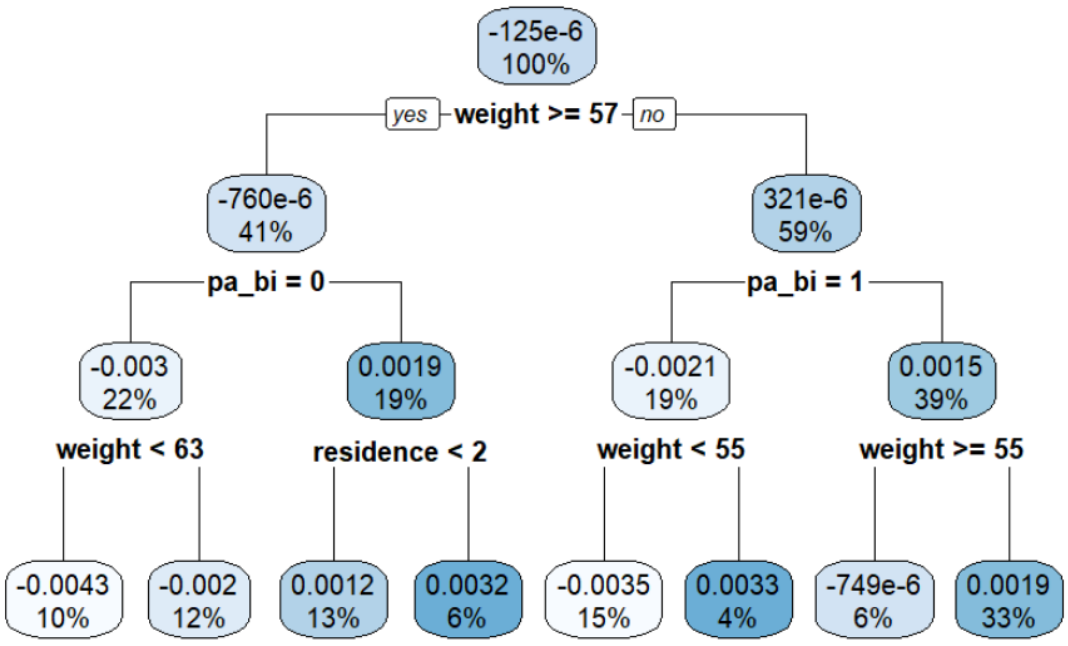

Figure 4.1: Model results for the binary outcome mortality of drinking or smoking

It is imperative to highlight that the variable pa_bi indicates physical activity, with pa_bi=0 denoting individuals who do not exercise regularly. Furthermore, the variable "residence" distinguishes between urban and rural habitation statuses. The variable named pa_bi is the level of physical activity, where '1' indicates a high level of physical activity, and '0' shows a low level of physical activity. The variable

'Weight' is the body weight by kilogram, and the 'residency' is the participants' living location, where '0' and '1' indicate living in a city and town, and '2' means living in a rural.

Starting from the root, the population is first split based on weight, with one branch for individuals with weight $\geq$57 (41% of the total population) and another for those with weight $<$ 57 (59%). In the $\geq$57 weight category, the further subdivision is based on the physical activity level, creating two paths: low physical activity level (pa_bi = 0, 22%) and high physical activity level (pa_bi = 1, 19%). The "pa_bi = 0" path is split again by weight into $<$ 63 and $\geq$63, with respective proportions of 10% to 13% and effect sizes ranging from -0.0043 to 0.0032. The "pa_bi = 1" branch shows a single effect size of -0.0021.

In the $<$ 57 weight category, the tree also divides according to "pa_bi," resulting in "pa_bi = 0" (19%) and "pa_bi = 1" (39%) paths. Further splits are made based on residence ($<$ 2 and $\geq$2) with effect sizes of -0.0035 and 0.0033, and weight ($<$ 55 and $\geq$55) with effect sizes of -7.49e-6 and 0.0019, corresponding to 15%, 4%, 6%, and 33% of the subgroups respectively. These branches reflect the study's detailed population stratification by weight, residence, and physical activity level, showing the effect size in each subgroup to understand the heterogeneous effects.

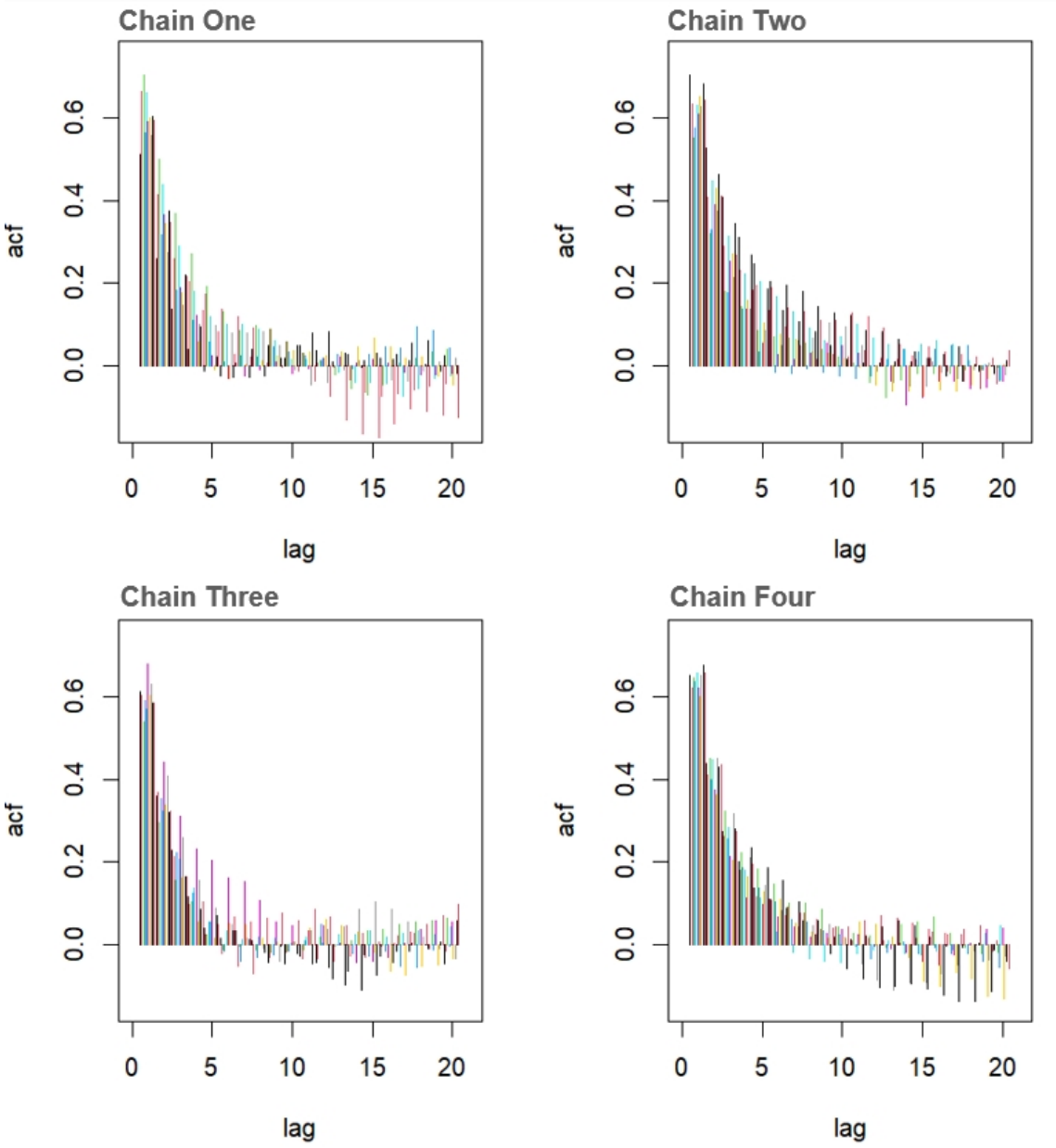

Figure 4.2: The autocorrelation of the estimated response surface from each BART chain: the auto-correlations of $f_{(x_i)}$ for randomly selected x_i where i indexes subjects

Figure 4.2 presented autocorrelation function (ACF) plots for four separate Markov chains in a BART model analysis. These plots are used to diagnose the independence of samples within each chain. Ideally, the autocorrelation should drop off quickly to near zero, indicating that the samples are not correlated and the chain is mixing well, suggesting effective exploration of the posterior distribution.

Each plot corresponds to one of the chains and depicts how the samples' autocorrelations decrease with increasing lag. The X-axis represents the lag, and the Y-axis represents the autocorrelation coefficient, ranging from -0.2 to 0.8. The colored bars at each lag point may represent the distribution of ACF values at different quantiles, providing a visual representation of the variability in autocorrelation across lags.

A sharp decline in ACF at the initial lags, seen across all chains, indicates quick decorrelation and sound mixing. This suggests that the chains efficiently explore the parameter space and generate independent samples. This pattern of rapid decrease in autocorrelation is a positive sign of convergence, implying that the posterior samples drawn from these chains likely represent the actual posterior distribution. This is essential for ensuring the robustness of the BART model's estimations.

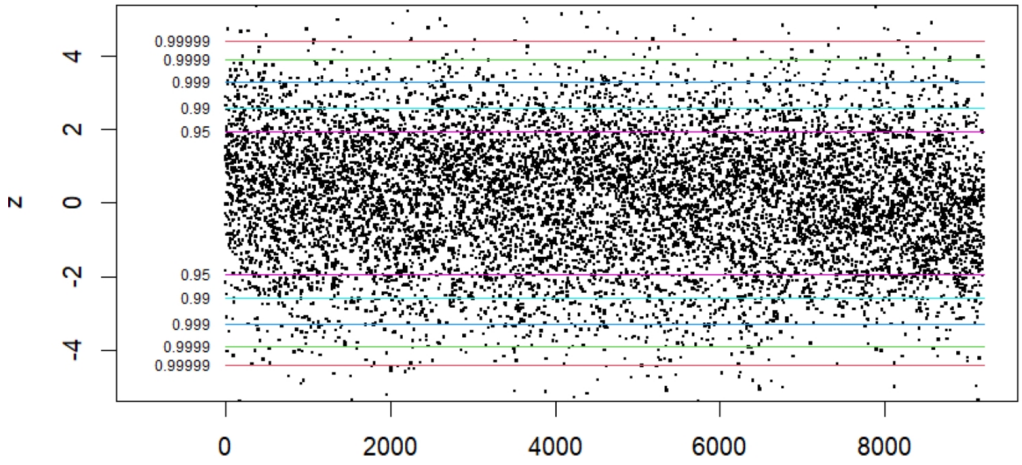

Figure 4.3: Geweke convergence diagnostics for probit BART: plot the Geweke Z statistics for each subject. The Z exceeds the 95% limits a handful of times. Based on this figure, we can conclude that the chains have converged.

Figure 4.3 illustrates a convergence diagnostic for a BART model using a sequence of Z-scores plotted over iterations of an MCMC simulation. The horizontal axis indexes each iteration of the MCMC chain, extending from 0 to beyond 8000. The vertical axis presents the computed Z-scores at each iteration. Horizontal lines indicate thresholds for various confidence levels, typically corresponding to standard normal distribution critical values for 95%, 99%, 99.9%, and 99.99% confidence intervals.

The Z-scores are densely concentrated around the horizontal line representing

zero, which implies that a significant proportion of the sampled values are near the expected mean of the posterior distribution. This concentration around zero suggests that the samples are from an equilibrium distribution. The majority of Z-scores are contained within the outermost horizontal lines, likely representing the 99.999% confidence level, indicating that the values are within a range expected for a converging MCMC chain. Additionally, the lack of discernible patterns or trends in the Z-scores across iterations, such as systematic drifts, suggests the absence of non-stationarity, further supporting the indication of convergence.

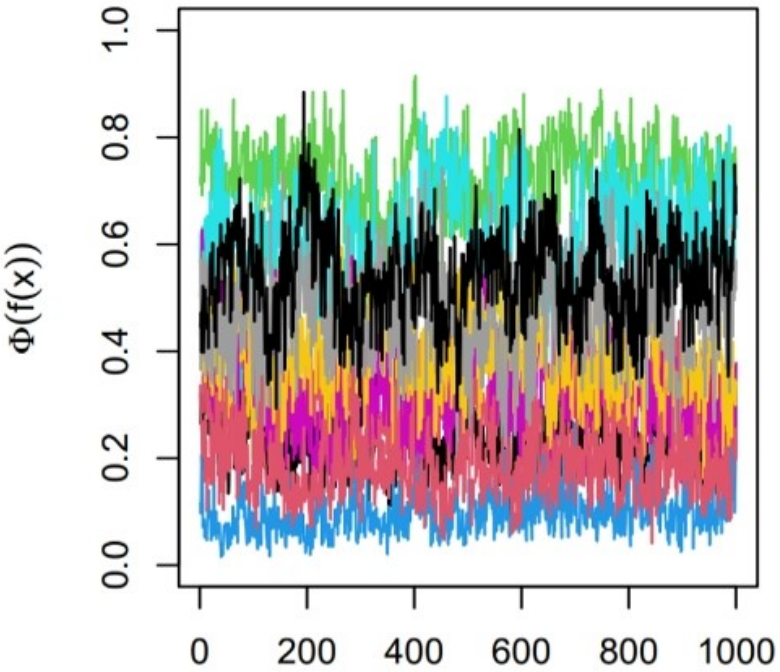

Figure 4.4: Trace plots for the BART Model: The traces demonstrate that samples of $f(x_i)$ appear to adequately traverse the sample space.

Figure 4.4 depicts a stacked area plot, often used in the visualization of the results from the BART model, particularly to show the distribution of probabilities or contributions of different components over a series of iterations. Each colored area represents the proportion of a certain feature or component (Y-axis) across the iterations of the model (X-axis). The X-axis, ranging from 0 to 1000, indicates the number of iterations or a subset of iterations during the MCMC simulation. The Y-axis, showing values from 0 to 1.0, represents the estimated probability or

proportion contributed by each component at each iteration.

The plot shows multiple layers of colors stacked on top of each other, with each color corresponding to one of the 48 groups within the model. This indicates that the model is examining multiple features or predictors simultaneously.

Such plots are useful for examining how the contribution of each component changes throughout the MCMC process and can help identify which features are consistently contributing more to the model's output. If the layers remain parallel and consistent, it suggests that the relative contributions of each component are stable across iterations, which could imply that the model has reached a stable solution.

The plot is a visualization of a BART model's components' probabilities or contributions over MCMC iterations, indicating the model's behavior and the stability of the features' effects within the model.

Table 4.2: The Association between Smoking or Drinking and All-cause Mortality within Pre-defined Subgroup by BART

Group		**No. of Participants**	**Hazard Ratio (95% CI) Smoking or Drinking**
Level 1 mediator	**Level 2 mediator**		
Weight $\geq$ 57	**High level physical activity level**	9588	**1.09 (0.98 – 1.11)**
	Low level physical activity level	8474	**1.12 (1.05 – 1.16)**
Weight $<$ 57	**High level physical activity level**	8466	**1.12 (1.05 – 1.16)**
	Low level physical activity level	16959	**1.03 (0.98 – 1.08)**

Table 4.2 displays a BART model's analysis results valid by COX regression, showing hazard ratios for all-cause mortality in relation to smoking or drinking across subgroups defined by weight and physical activity levels. It divides participants into groups based on their weight (either greater than or equal to 57 or less than 57) and their physical activity level (high or low). For each subgroup, the number of participants and the hazard ratio for smoking or drinking is provided, along with the 95% confidence interval (CI).

The hazard ratios suggest the relative risk of mortality associated with smoking or drinking for individuals within these subgroups. A hazard ratio greater than 1 indicates a higher risk compared to the baseline risk [85], which is not explicitly stated here but would be the risk for non-smokers or non-drinkers. For example,

individuals with a weight of 57 or more and a low level of physical activity have a hazard ratio of 1.12, suggesting that they have a 12% higher risk of mortality associated with smoking or drinking compared to the baseline group. The confidence interval of 1.05 – 1.16 indicates that this estimate is statistically significant with 95% certainty.

Subgroups include individuals with weight $\geq$ with high physical activity (9588 participants, hazard ratio 1.09, 95% CI 0.98–1.11), weight $\geq$ with low physical activity (8474 participants, hazard ratio 1.12, 95% CI 1.05–1.16), weight < 57 with high physical activity (8466 participants, hazard ratio 1.12, 95% CI 1.05–1.16), and weight < 57 with low physical activity (16959 participants, hazard ratio 1.03, 95% CI 0.98–1.08), suggesting that higher physical activity may be associated with a slight increase in the hazard ratio for mortality from smoking or drinking, though the confidence intervals indicate a degree of uncertainty in these estimates.

Overall, this table communicates that body weight and physical activity levels are significant mediators in the relationship between smoking or drinking and mortality, and these relationships are quantified by the BART model.

4.3 Subgroup analysis

The analysis investigated the association between smoking or drinking and mortality in four subgroups based on body weight and physical activity level predefined by the analysis from my heterogeneous effect analysis. The hazard ratios (HRs) [86] and corresponding 95 percentage confidence intervals (CIs) were calculated for each subgroup. All the models were adjusted from gender and age [87].

In the subgroup of individuals with a weight of 57 kilograms or more, a high level of physical activity was associated with an HR of 1.09 (95% CI: 0.98 – 1.11), while a low level of physical activity was associated with an HR of 1.12 (95% CI: 1.05 – 1.16).

Among individuals with a weight less than 57 kilograms, a high level of physical activity was associated with an HR of 1.12 (95% CI: 1.05 – 1.16), while a low level of physical activity had an HR of 1.03 (95% CI: 0.98 – 1.08).

These results indicate that the association between smoking or drinking and mortality varied based on body weight and physical activity level significantly. In the subgroup of individuals with a weight of 57 kilograms or more, both high and low physical activity levels were associated with a higher risk [88] of mortality [89]. Among individuals with a weight of less than 57 kilograms, only a high level of physical activity was associated with an increased risk of mortality [90].

These findings highlight the importance of considering lifestyle factors, such as smoking or drinking, along with body weight and physical activity [74], when examining the risk of mortality. Further research is needed to understand better the underlying mechanisms and potential confounding factors that may influence these associations [91].

4.4 Sensitivity analysis

Sensitivity analyses dropping those participants who died in six months or under best- and worst-case imputation of the missing outcomes resulted in very similar final decision trees that selected the same covariates and resulted in nearly identical subgroups and conclusions [92].

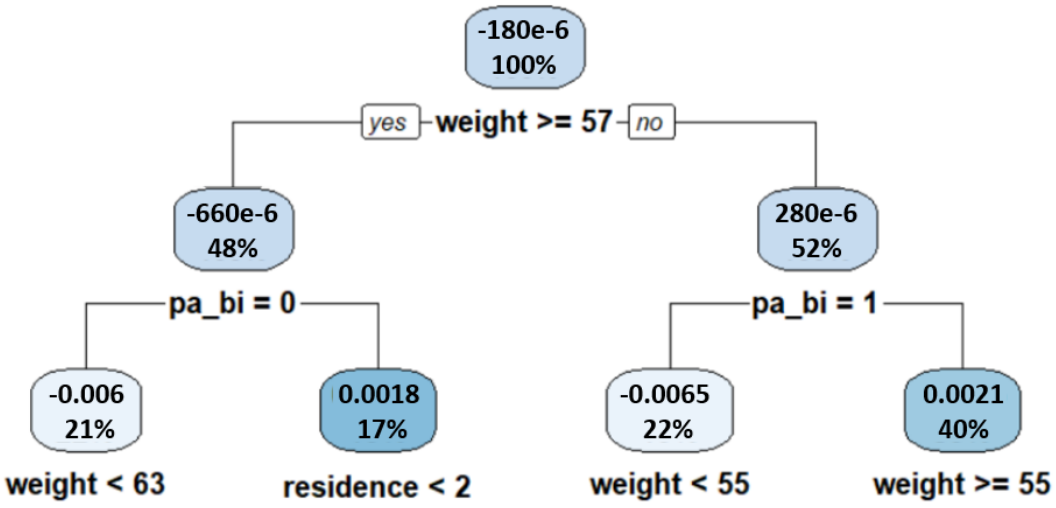

Figure 4.5: Sensitivity analysis: Best-worse outcome imputation

Figure 4.5 presents the results of the sensitivity analysis by best-worse outcome imputation of mortality. The results only have minor changes compared with the main analysis, which indicates a good robustness of the analysis.

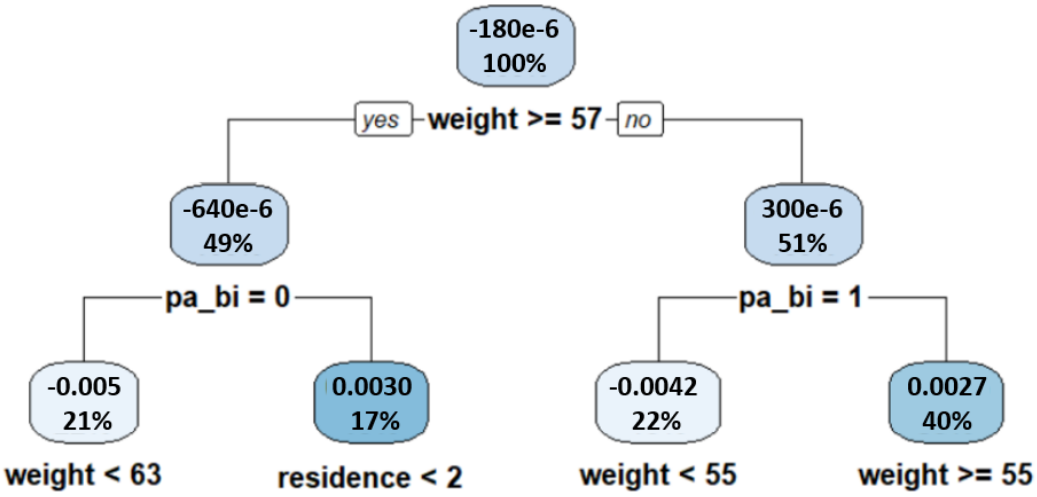

Figure 4.6: Sensitivity analysis: Worse-best outcome imputation

Figure 4.6 presents the results of the sensitivity analysis by worse-best outcome

imputation of mortality. The results only have minor changes compared with the main analysis, which indicates a good robustness of the analysis.

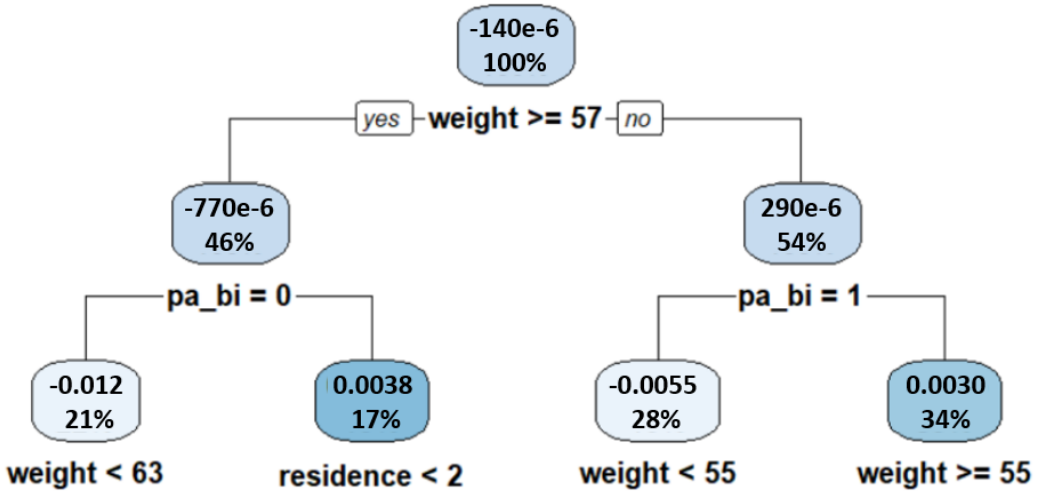

Figure 4.7: Sensitivity analysis: Removing those nearly death

Figure 4.7 presents the results of the sensitivity analysis by removing those nearly dying. The results only have minor changes compared with the main analysis, which indicates a good robustness of the analysis.

These three images, titled "Sensitivity Analysis," describe the robustness of the study's findings by testing the results against different scenarios to assess if and how conclusions may change under various conditions [93]. Sensitivity analysis is crucial in research to ensure that the results are not unduly influenced by specific assumptions or data treatment methods [94].

We conducted three different sensitivity analyses:

(1)Using best/worst-case imputation, see Figure 4.5;

(2)Using worst/best-case imputation, see Figure 4.6;

(3)Dropping the participants who died within six months after the baseline survey, see Figure 4.7 [95].

For each type of analysis, a corresponding decision tree is presented to show how the outcome variable (possibly the participants' mortality) is affected by variables such as weight and physical activity (PA), indicated as "pa_bi".

The decision trees under each method show the probabilities and impacts of different participant characteristics on the outcome. These trees can help identify which factors are most influential and robust across various methods of handling missing or uncertain data. For example, the first tree under the best/worst-case imputation method shows that for participants with a weight of 63 or less and a residence less than 2, there's a specific probability (shown in the box) of the outcome occurring. Similarly, other trees show different probabilities based on the weight criteria and the physical activity index [96].

This type of analysis helps to confirm the reliability of the study's conclusions by showing that the significant findings hold true even when the data is manipulated to account for potential uncertainties or biases [97].

The sensitivity analysis helps confirm the reliability of the study's conclusions by showing that the significant findings hold true even when the data is manipulated to account for potential uncertainties or biases. I have done three types of sensitivity analysis, and the results only show minor changes, which indicate the robustness of my study.

CHAPTER 5

Discussion

5.1 Conclusion and implication

This study investigates the heterogeneous effects of mortality associated with smoking and drinking using novel machine learning-based methods, specifically by combining the missForest algorithm and BART. The study's implications are significant, as it challenges traditional notions of the risks associated with these behaviours and, if further validated by trials, offers a more nuanced understanding of their impact on different subgroups within the elderly population.

Two distinct subgroups were identified where smoking and drinking considerably increase mortality risks, characterized by specific combinations of body weight, physical activity levels, and residency. In the geriatric population, these concurrent risk factors suggest a synergistic increase in mortality risks. In contrast, other subgroups showed no traditional correlation between smoking, drinking, and increased mortality risks, indicating a more complex relationship that varies with individual characteristics.

The findings imply that public health interventions should be tailored to address the specific needs of different subgroups. Smoking and drinking cessation

programs may offer substantial benefits to certain elderly subgroups, particularly those with identified concurrent risk factors. However, for other subgroups, alternative strategies that do not focus solely on cessation may be more appropriate. This study's use of BART methods to identify heterogeneous effects demonstrates the technique's effectiveness and potential for broader application in public health and precision medicine research.

The research's nuanced insights into the relationship between smoking, drinking, and mortality underscore the importance of individualized public health strategies. It encourages the development of interventions that are responsive to the varying impacts of these behaviors across different segments of the elderly population. This approach could lead to more effective public health campaigns and potentially reduce the mortality rates associated with smoking and drinking among the elderly.

In conclusion, the study contributes to a deeper understanding of the complex interplay between smoking, drinking, and mortality in the elderly. It advocates for a shift towards personalized public health interventions that consider the unique characteristics of each subgroup, which could significantly improve the effectiveness of smoking and drinking cessation efforts and enhance the well-being of the elderly population.

5.2 Limitation and future work

While this study provides valuable insights into the relationship between smoking, drinking, and mortality using the CLHLS cohort, it recognizes certain limitations that warrant consideration. The CLHLS cohort, although extensive for the Asian elderly population, could be further strengthened by external validation across multiple cohorts from diverse ethnic backgrounds, including African and Caucasian populations. Such validation could bolster the generalizability and applicability of the findings. Additionally, conducting clinical trials would be an important step in corroborating the associations found and testing the efficacy of interventions based on these insights.

Looking ahead, the establishment of a benchmark for comparing the efficacy of

various machine learning algorithms in this domain is a critical task. This benchmark would enable a systematic evaluation of different models, including but not limited to tree-based approaches for detecting heterogeneous effect. The exploration of graph-based models and reinforcement-learning based causal networks presents a promising direction for future research. These advanced methods could offer new perspectives and deeper insights into the causal relationships between lifestyle factors and mortality.

Further research could also delve into the development of personalized public health strategies based on machine learning predictions. Tailoring interventions to individual risk profiles identified through predictive modeling could lead to more effective health outcomes. Additionally, investigating the mechanisms through which smoking and drinking impact health could lead to the discovery of novel therapeutic targets or preventative measures.

In terms of data analysis, future work could involve integrating longitudinal data analysis techniques to account for time-varying effects and the potential evolution of risk factors over time. The role of genetic predispositions and their corresponding heterogeneous effect is another avenue that could be explored to provide a more comprehensive understanding of mortality risks.

Lastly, the ethical implications of using machine learning in public health should be carefully considered. The development of models that are not only accurate but also fair and interpretable will be essential in ensuring the responsible application of these technologies in healthcare settings. Accordingly, the methodology to increase the model's interpretability is vital in the future.

In summary, the study lays the groundwork for a wide array of future research opportunities that could significantly advance our understanding of mortality risks and contribute to improving public health strategies for the elderly population worldwide.

Bibliography

[1] D. M. Kent, P. M. Rothwell, J. P. Ioannidis, D. G. Altman, and R. A. Hayward, "Assessing and reporting heterogeneity in treatment effects in clinical trials: a proposal," *Trials*, vol. 11, pp. 1–11, 2010. 1

[2] A. Rekkas, J. K. Paulus, G. Raman, J. B. Wong, E. W. Steyerberg, P. R. Rijnbeek, D. M. Kent, and D. van Klaveren, "Predictive approaches to heterogeneous treatment effects: a scoping review," *BMC Medical Research Methodology*, vol. 20, no. 1, pp. 1–12, 2020. 1, 1.1

[3] M. C. Odden, A. M. Rawlings, A. Khodadadi, X. Fern, M. G. Shlipak, K. Bibbins-Domingo, K. Covinsky, A. M. Kanaya, A. Lee, M. N. Haan, *et al.*, "Heterogeneous exposure associations in observational cohort studies: the example of blood pressure in older adults," *American journal of epidemiology*, vol. 189, no. 1, pp. 55–67, 2020. 1

[4] R. Entezari, *Bayesian Computations via MCMC, with Applications to Big Data and Spatial Data*. University of Toronto (Canada), 2018. 1, 1.1, 3.2.1

[5] C. L. Hart, G. Davey Smith, L. Gruer, and G. C. Watt, "The combined effect of smoking tobacco and drinking alcohol on cause-specific mortality: a 30 year cohort study," *BMC public health*, vol. 10, pp. 1–11, 2010. 1

[6] S. R. Robinson, *Essays in Health and Public Economics*. University of California, Santa Barbara, 2023. 1, 2.5

[7] A. Linero, D. Sinha, and S. Lipsitz, "Semiparametric mixed-scale models using shared bayesian forests. arxiv e-prints," *arXiv preprint arXiv:1809.08521*, 2018. 1

[8] S. Negi, "A blockchain technology for improving financial flows in humanitarian supply chains: benefits and challenges," *Journal of Humanitarian Logistics and Supply Chain Management*, 2024. 1

[9] T. Cao, L. Lu, and T. Jiang, "Robust regression in environmental modeling based on bayesian additive regression trees," *Environmental Modeling & Assessment*, pp. 1–13, 2023. 1

[10] G. W. Gundersen, *Practical Algorithms for Latent Variable Models*. Princeton University, 2021. 1.1

[11] P. Dadvand, J. Parker, M. L. Bell, M. Bonzini, M. Brauer, L. A. Darrow, U. Gehring, S. V. Glinianaia, N. Gouveia, E.-h. Ha, *et al.*, "Maternal exposure to particulate air pollution and term birth weight: a multi-country evaluation of effect and heterogeneity," *Environmental health perspectives*, vol. 121, no. 3, pp. 267–373, 2013. 1.1

[12] K. Lee, D. S. Small, and F. Dominici, "Discovering heterogeneous exposure effects using randomization inference in air pollution studies," *Journal of the American Statistical Association*, vol. 116, no. 534, pp. 569–580, 2021. 1.1

[13] D. T. Lackland, "Racial differences in hypertension: implications for high blood pressure management," *The American journal of the medical sciences*, vol. 348, no. 2, pp. 135–138, 2014. 1.1

[14] M. Shadrina, E. A. Bondarenko, and P. A. Slominsky, "Genetics factors in major depression disease," *Frontiers in psychiatry*, vol. 9, p. 334, 2018. 1.1

[15] R. K. McHugh and R. D. Weiss, "Alcohol use disorder and depressive disorders," *Alcohol research: current reviews*, vol. 40, no. 1, 2019. 1.1

[16] T. M. Barber, P. Hanson, S. Kabisch, A. F. Pfeiffer, and M. O. Weickert, "The low-carbohydrate diet: Short-term metabolic efficacy versus longer-term limitations," *Nutrients*, vol. 13, no. 4, p. 1187, 2021. 1.1

[17] C. De Chaisemartin and X. d'Haultfoeuille, "Two-way fixed effects estimators with heterogeneous treatment effects," *American Economic Review*, vol. 110, no. 9, pp. 2964–2996, 2020. 1.1

[18] Y. Yao, K. Cao, K. Zhang, T. Zhu, D. Yue, H. Zhang, J. Zhang, X. Jin, and Y. Zeng, "Residential proximity to major roadways and prevalent hypertension among older women and men: results from the chinese longitudinal healthy longevity survey," *Frontiers in cardiovascular medicine*, vol. 7, p. 587222, 2020. 2.1

[19] F. Meyer, I. Bairati, A. Fortin, M. Gélinas, A. Nabid, F. Brochet, and B. Têtu, "Interaction between antioxidant vitamin supplementation and cigarette smoking during radiation therapy in relation to long-term effects on recurrence and mortality: a randomized trial among head and neck cancer patients," *International journal of cancer*, vol. 122, no. 7, pp. 1679–1683, 2008. 2.1

[20] H.-X. Wang, A. Karp, B. Winblad, and L. Fratiglioni, "Late-life engagement in social and leisure activities is associated with a decreased risk of dementia: a longitudinal study from the kungsholmen project," *American journal of epidemiology*, vol. 155, no. 12, pp. 1081–1087, 2002. 2.1

[21] D. Gu, "General data quality assessment of the clhls," *Healthy longevity in China: Demographic, socioeconomic, and psychological dimensions*, pp. 39–60, 2008. 2.1

[22] X. Jin, W. He, Y. Zhang, E. Gong, Z. Niu, J. Ji, Y. Li, Y. Zeng, and L. L. Yan, "Association of apoe $\varepsilon 4$ genotype and lifestyle with cognitive function among chinese adults aged 80 years and older: A cross-sectional study," *PLoS medicine*, vol. 18, no. 6, p. e1003597, 2021. 2.1, 2.3

[23] L. Zhu, M. Lei, L. Tan, and M. Zou, "Sex difference in the association between bmi and cognitive impairment in chinese older adults," *Journal of Affective Disorders*, 2024. 2.1

[24] Y. Zeng, Q. Feng, D. Gu, and J. W. Vaupel, "Demographics, phenotypic health characteristics and genetic analysis of centenarians in china," *Mechanisms of ageing and development*, vol. 165, pp. 86–97, 2017. 2.1

[25] A. Cappelleri, N. Bussmann, S. Harvey, P. T. Levy, O. Franklin, and E.-K. Afif, "Myocardial function in late preterm infants during the transitional period: comprehensive appraisal with deformation mechanics and non-invasive cardiac output monitoring," *Cardiology in the Young*, vol. 30, no. 2, pp. 249–255, 2020. 2.1

[26] X. Jin, S. Xiong, S.-Y. Ju, Y. Zeng, L. L. Yan, and Y. Yao, "Serum 25-hydroxyvitamin d, albumin, and mortality among chinese older adults: a population-based longitudinal study," *The Journal of Clinical Endocrinology & Metabolism*, vol. 105, no. 8, pp. 2762–2770, 2020. 2.2

[27] S. Xiong, Z. Wang, B. Lee, Q. Guo, N. Peoples, X. Jin, E. Gong, Y. Li, X. Chen, Z. He, *et al.*, "The association between self-rated health and all-cause mortality and explanatory factors in china's oldest-old population," *Journal of Global Health*, vol. 12, 2022. 2.2

[28] L. Breiman, "Using iterated bagging to debias regressions," *Machine Learning*, vol. 45, pp. 261–277, 2001. 2.2, 3.1.2

[29] H. Zhu and D. Gu, "The protective effect of marriage on health and survival: Does it persist at oldest-old ages?," *Journal of Population Ageing*, vol. 3, pp. 161–182, 2010. 2.2

[30] O. V. Marchenko and N. V. Katenka, "Quantitative methods in pharmaceutical research and development: Concepts and applications," 2020. 2.2

[31] L. Qiu, J. Sautter, Y. Liu, and D. Gu, "Age and gender differences in linkages of sleep with subsequent mortality and health among very old chinese," *Sleep medicine*, vol. 12, no. 10, pp. 1008–1017, 2011. 2.2

[32] H. Zhu, Q. Feng, and D. Gu, "Self-rated health and interviewer-rated health: differentials in predictive power for mortality among subgroups of chinese elders," 2017. 2.2

[33] R. Quinlan, "Bagging, boosting, and c4.5," in *Proceedings of the Thirteenth National Conference on Artificial Intelligence*, 1996. 2.3

[34] S. Wu, X. Lv, J. Shen, H. Chen, Y. Ma, X. Jin, J. Yang, Y. Cao, G. Zong, H. Wang, *et al.*, "Association between body mass index, its change and cognitive impairment among chinese older adults: a community-based, 9-year prospective cohort study," *European Journal of Epidemiology*, vol. 36, no. 10, pp. 1043–1054, 2021. 2.3

[35] S. Fisher, "Doctor-patient communication: a social and micro-political performance," *Sociology of health & illness*, vol. 6, no. 1, pp. 1–29, 1984. 2.3

[36] G. Huang, F. Guo, and G. Chen, "Educational differences of healthy life expectancy among the older adults in china: a multidimensional examination using the multistate life table method," *Educational Gerontology*, vol. 45, no. 10, pp. 624–635, 2019. 2.3

[37] W. He, T. T. Hu, *et al.*, "The influence of cultural and policy intervention on medical values: Based on chinese theories and empirical research," 2020. 2.3

[38] S. Casucci, L. Lin, and A. Nikolaev, "Modeling the impact of care transition programs on patient outcomes and 30 day hospital readmissions," *Socio-Economic Planning Sciences*, vol. 63, pp. 70–79, 2018. 2.4

[39] X. Jin, S. Xiong, C. Yuan, E. Gong, X. Zhang, Y. Yao, Y. Leng, Z. Niu, Y. Zeng, and L. L. Yan, "Apolipoprotein e genotype, meat, fish, and egg intake in relation to mortality among older adults: a longitudinal analysis in china," *Frontiers in Medicine*, p. 1114, 2021. 2.4

[40] L. Weiss, *Modeling participation in citizen science: Recreational fishermen in Massachusetts*. University of Rhode Island, 2015. 2.4

[41] D. Eddington, A. F. D. Correa, G. Wolf, and K. R. Moon, "Data imputation with an autoencoder and magic," in *2023 International Conference on Sampling Theory and Applications (SampTA)*, pp. 1–5, IEEE, 2023. 2.5

[42] B. De Finetti, "La prévision: ses lois logiques, ses sources subjectives," in *Annales de l'institut Henri Poincaré*, vol. 7, pp. 1–68, 1937. 3, 3.4.3

[43] J. Brownlee, "How to choose the right test options when evaluating machine learning algorithms," *MachineLearningMastery. com. Available online: https://machinelearningmastery. com/how-to-choose-the-right-test-options-when-evaluating-machine-learning-algorithms/(accessed on 31 January 2023)*, 2014. 3

[44] B. D. Butcher, *MCMC diagnostics for Bayesian additive regression trees and methods for flexible modeling of predictors*. PhD thesis, The University of Iowa, 2020. 3, 3.1, 3.1, 3.1.2, 3.1.2, 3.1.2, 3.2.2, 3.2.2, 3.2.2, 3.2.3, 3.2.3, 3.2.3

[45] D. A. Kaufman and N. R. Cloutier, "The impact of small brownfields and greenspaces on residential property values," *The Journal of Real Estate Finance and Economics*, vol. 33, pp. 19–30, 2006. 3

[46] L. Beusch, L. Foresti, M. Gabella, and U. Hamann, "Satellite-based rainfall retrieval: From generalized linear models to artificial neural networks," *Remote Sensing*, vol. 10, no. 6, p. 939, 2018. 3.1

[47] O. Gencel, F. Kocabas, M. S. Gok, and F. Koksal, "Comparison of artificial neural networks and general linear model approaches for the analysis of abrasive wear of concrete," *Construction and building materials*, vol. 25, no. 8, pp. 3486–3494, 2011. 3.1

[48] H. A. Chipman, E. I. George, and R. E. McCulloch, "Bart: Bayesian additive regression trees," 2010. 3.1.1, 3.1.2

[49] W. Ebeling, V. E. Fortov, and V. Filinov, *Quantum Statistics of Dense Gases and Nonideal Plasmas*. Springer, 2017. 3.1.1

[50] J. H. Friedman, "Greedy function approximation: a gradient boosting machine," *Annals of statistics*, pp. 1189–1232, 2001. 3.1.2

[51] T. Hastie, R. Tibshirani, J. Friedman, T. Hastie, R. Tibshirani, and J. Friedman, "Model inference and averaging," *The elements of statistical learning: Data mining, inference, and prediction*, pp. 261–294, 2009. 3.1.2

[52] T. Hastie and R. Tibshirani, "Bayesian backfitting (with comments and a rejoinder by the authors," *Statistical Science*, vol. 15, no. 3, pp. 196–223, 2000. 3.1.2

[53] S. Brooks, "Markov chain monte carlo method and its application," *Journal of the royal statistical society: series D (the Statistician)*, vol. 47, no. 1, pp. 69–100, 1998. 3.2.1

[54] B. P. Carlin and S. Chib, "Bayesian model choice via markov chain monte carlo methods," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 57, no. 3, pp. 473–484, 1995. 3.2.1

[55] P. L. Green and K. Worden, "Bayesian and markov chain monte carlo methods for identifying nonlinear systems in the presence of uncertainty," *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, vol. 373, no. 2051, p. 20140405, 2015. 3.2.1

[56] M. Johnson, T. L. Griffiths, and S. Goldwater, "Bayesian inference for pcfgs via markov chain monte carlo," in *Human Language Technologies 2007: The Conference of the North American Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference*, pp. 139–146, 2007. 3.2.1

[57] S. Chib and E. Greenberg, "Understanding the metropolis-hastings algorithm," *The American Statistician*, vol. 49, no. 4, pp. 327–335, 1995. 3.2.1

[58] R. M. Neal, "Probabilistic inference using markov chain monte carlo methods," 1993. 3.2.1

[59] C. Robert and G. Casella, *Monte Carlo Statistical Methods*. Springer, 2004. 3.2.2, 3.2.3

[60] D. Sejdinovic, H. Strathmann, M. L. Garcia, C. Andrieu, and A. Gretton, "Kernel adaptive metropolis-hastings," in *International conference on machine learning*, pp. 1665–1673, PMLR, 2014. 3.2.2

[61] T. Neupane, Z. Zhang, C. Madsen, H. Zheng, and C. J. Myers, "Approximation techniques for stochastic analysis of biological systems," *Automated Reasoning for Systems Biology and Medicine*, pp. 327–348, 2019. 1

[62] J. Bellettiere, M. J. LaMonte, G. N. Healy, S. Liles, K. R. Evenson, C. Di, J. Kerr, I.-M. Lee, E. Rillamas-Sun, D. Buchner, *et al.*, "Sedentary behavior and diabetes risk among women over the age of 65 years: the opach study," *Diabetes care*, vol. 44, no. 2, pp. 563–570, 2021. 2

[63] J. Starling, J. Murray, P. Lohr, A. Aiken, C. Carvalho, and J. Scott, "Targeted smooth bayesian causal forests: An analysis of heterogeneous treatment effects for simultaneous versus interval medical abortion regimens over gestation," *arXiv:1905.09405 [stat.AP]*, 2019. 3.2.3

[64] R. Sparapani, N. Dabbouseh, J. Gutterman, Zhang, H. Chen, D. Bluemke, J. Lima, G. Burke, and E. Soliman, "Detection of left ventricular hypertrophy using bayesian additive regression trees: The mesa (multi-ethnic study of atherosclerosis)," *Journal of the American Heart Association*, vol. 8, no. 5, 2019. 3.2.3

[65] A. E. Raftery and S. Lewis, "How many iterations in the gibbs sampler?," in *Bayesian Statistics 4*, 1992. 3.2.3

[66] M. E. J. Newman and G. T. Barkema, *Monte Carlo Methods in Statistical Physics*. USA: Oxford University Press, 1999. 3.2.3

[67] A. Basu, "Directed acyclic graphs to explore causality in epidemiological study designs, part i: an introduction to dags," *Qeios*, 2020. 3.2.3

[68] I. Yildirim, "Bayesian inference: Gibbs sampling," *Technical Note, University of Rochester*, 2012. 3.2.4

[69] A. E. Gelfand, "Gibbs sampling," *Journal of the American statistical Association*, vol. 95, no. 452, pp. 1300–1304, 2000. 3.2.4

[70] I. Porteous, D. Newman, A. Ihler, A. Asuncion, P. Smyth, and M. Welling, "Fast collapsed gibbs sampling for latent dirichlet allocation," in *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 569–577, 2008. 3.2.4, 3.4.3

[71] P. Ding, A. Feller, and L. Miratrix, "Randomization inference for treatment effect variation," *Journal of the Royal Statistical Society Series B: Statistical Methodology*, vol. 78, no. 3, pp. 655–671, 2016. 3.4.1

[72] A. Coppock, T. J. Leeper, and K. J. Mullinix, "Generalizability of heterogeneous treatment effect estimates across samples," *Proceedings of the National Academy of Sciences*, vol. 115, no. 49, pp. 12441–12446, 2018. 3.4.1

[73] D. P. Green and H. L. Kern, "Modeling heterogeneous treatment effects in large-scale experiments using bayesian additive regression trees," in *The annual summer meeting of the society of political methodology*, pp. 100–110, 2010. 3.4.2

[74] C. Geissler and H. J. Powers, *Human nutrition*. Oxford University Press, 2017. 3.4.3, 4.3

[75] C. Daskalopoulou, B. Stubbs, C. Kralj, A. Koukounari, M. Prince, and A. M. Prina, "Associations of smoking and alcohol consumption with healthy ageing: a systematic review and meta-analysis of longitudinal studies," *BMJ open*, vol. 8, no. 4, p. e019540, 2018. 3.4.3

[76] D. Gu, Q. Feng, and Y. Zeng, "Chinese longitudinal healthy longevity study," *Encyclopedia of Geropsychology. Singapore: Springer*, pp. 469–82, 2017. 3.4.3

[77] J. Hill, A. Linero, and J. Murray, "Bayesian additive regression trees: A review and look forward," *Annual Review of Statistics and Its Application*, vol. 7, pp. 251–278, 2020. 3.4.3

[78] A. Raftery and S. Lewis, "One long run with diagnostics: Implementation strategies for markov chain monte carlo," *Statistical Science*, vol. 7, pp. 493–497, 1992. 3.4.3

[79] A. Kapelner and J. Bleich, "Bartmachine: Machine learning with bayesian additive regression trees," *arXiv preprint arXiv:1312.2171*, 2013. 3.5

[80] R. A. Sparapani, B. R. Logan, R. E. McCulloch, and P. W. Laud, "Nonparametric survival analysis using bayesian additive regression trees (bart)," *Statistics in medicine*, vol. 35, no. 16, pp. 2741–2753, 2016. 3.5

[81] H. Du, Z. Ke, G. Jiang, and S. Huang, "The performances of gelman-rubin and geweke's convergence diagnostics of monte carlo markov chains in bayesian analysis," *Journal of Behavioral Data Science*, vol. 2, no. 2, pp. 47–72, 2022. 3.5, 3.5

[82] R. Sparapani, C. Spanbauer, and R. McCulloch, "Nonparametric machine learning and efficient computation with bayesian additive regression trees: the bart r package," *Journal of Statistical Software*, vol. 97, pp. 1–66, 2021. 3.5

[83] M. Goicoechea, F. Gomez-Preciado, S. Benito, J. Torras, R. Torra, A. Huerta, A. Restrepo, J. Ugalde, D. E. Astudillo, I. Agraz, *et al.*, "Predictors of outcome in a spanish cohort of patients with fabry disease on enzyme replacement therapy," *nefrologia*, vol. 41, no. 6, pp. 652–660, 2021. 4.1

[84] J. S. Henrique, P. L. G. Braga, S. S. d. Almeida, N. S. P. Nunes, I. D. Benfato, R. M. Arida, C. A. M. de Oliveira, and S. Gomes da Silva, "Effect of the actn-3 gene polymorphism on functional fitness and executive function of elderly," *Frontiers in Aging Neuroscience*, vol. 14, p. 943934, 2022. 4.1

[85] A. Lee, G. Cheung, and M. Wong, "Long-term outcome of primary non-surgical root canal treatment," *Clinical oral investigations*, vol. 16, pp. 1607–1617, 2012. 4.2

[86] F. Wolfe, L. Caplan, and K. Michaud, "Treatment for rheumatoid arthritis and the risk of hospitalization for pneumonia: associations with prednisone, disease-modifying antirheumatic drugs, and anti–tumor necrosis factor therapy," *Arthritis & Rheumatism: Official Journal of the American College of Rheumatology*, vol. 54, no. 2, pp. 628–634, 2006. 4.3

[87] E. Svensson, E. Horváth-Puhó, R. W. Thomsen, J. C. Djurhuus, L. Pedersen, P. Borghammer, and H. T. Sørensen, "Vagotomy and subsequent risk of p arkinson's disease," *Annals of neurology*, vol. 78, no. 4, pp. 522–529, 2015. 4.3

[88] M. d. S. V. Fernandes, T. M. V. d. Silva, P. R. e. S. Noll, A. A. d. Almeida, and M. Noll, "Depressive symptoms and their associated factors in vocational–technical school students during the covid-19 pandemic," *International journal of environmental research and public health*, vol. 19, no. 6, p. 3735, 2022. 4.3

[89] W. Li, C. P. Wen, W. Li, Z. Ying, S. Pan, Y. Li, Z. Zhu, M. Yang, H. Tu, Y. Guo, *et al.*, "6-year trajectory of fasting plasma glucose (fpg) and mortality risk among individuals with normal fpg at baseline: a prospective cohort study," *Diabetology & Metabolic Syndrome*, vol. 15, no. 1, p. 169, 2023. 4.3

[90] S. Natarajan, S. R. Lipsitz, and E. Rimm, "A simple method of determining confidence intervals for population attributable risk from complex surveys," *Statistics in medicine*, vol. 26, no. 17, pp. 3229–3239, 2007. 4.3

[91] H. Uno, J. Wittes, H. Fu, S. D. Solomon, B. Claggett, L. Tian, T. Cai, M. A. Pfeffer, S. R. Evans, and L.-J. Wei, "Alternatives to hazard ratios for comparing the efficacy or safety of therapies in noninferiority studies," *Annals of internal medicine*, vol. 163, no. 2, pp. 127–134, 2015. 4.3

[92] C. Gamble and S. Hollis, "Uncertainty method improved on best–worst case analysis in a binary meta-analysis," *Journal of clinical epidemiology*, vol. 58, no. 6, pp. 579–588, 2005. 4.4

[93] C. Vincent, É. Tremblay-Wragg, and I. Plante, "Effects of a participation in a structured writing retreat on doctoral mental health: An experimental and comprehensive study," *International Journal of Environmental Research and Public Health*, vol. 20, no. 20, p. 6953, 2023. 4.4

[94] J. C. Jakobsen, C. Gluud, J. Wetterslev, and P. Winkel, "When and how should multiple imputation be used for handling missing data in randomised clinical trials–a practical guide with flowcharts," *BMC medical research methodology*, vol. 17, no. 1, pp. 1–10, 2017. 4.4

[95] J. P. Higgins, I. R. White, and A. M. Wood, "Imputation methods for missing outcome data in meta-analysis of clinical trials," *Clinical trials*, vol. 5, no. 3, pp. 225–239, 2008. 4.4

[96] I. Rombach, R. Knight, N. Peckham, J. R. Stokes, and J. A. Cook, "Current practice in analysing and reporting binary outcome data—a review of randomised controlled trial reports," *BMC medicine*, vol. 18, no. 1, pp. 1–8, 2020. 4.4

[97] M. L. Antonucci, *Bringing reading strategies home from a family literacy workshop: Two case studies of parents and their children reading together*. University of Massachusetts Amherst, 2005. 4.4