Analyzing Early Smoking Initiation and Mortality

A Complete Walkthrough with 20 Years of NHANES Data

M Ehsan Karim Sadia Khan Durani

2025-08-22

Table of contents

1 Reproducing the Analysis: Early Smoking Initiation and Mortality

---

Welcome! This technical documentation provides a comprehensive and transparent guide for reproducing the analysis from the published paper:

> Karim, M. E., Hossain, M. B., & Zheng, C. (2025). Examining the Role of Race/Ethnicity and Sex in Modifying the Association Between Early Smoking Initiation and Mortality: A 20-Year NHANES Analysis. *AJPM Focus*, 4(2), 100282. https://doi.org/10.1016/j.focus.2024.100282

By documenting the complete analytical pipeline—from raw data processing to the final statistical models—this guide allows researchers to fully understand and replicate the study, or to adapt this framework for new research questions.

1.1 About This Guide

This guide is designed for researchers, students, and public health analysts interested in longitudinal data analysis using NHANES. It demonstrates a complete workflow, including:

- **Data Acquisition**: Programmatically downloading and merging multiple cycles of NHANES data.
- **Data Cleaning**: Harmonizing variables that change across survey years.
- **Complex Survey Analysis**: Correctly applying survey weights, strata, and clusters for nationally representative estimates.
- **Survival Modeling**: Implementing Kaplan-Meier curves, Cox proportional hazards models, and effect modification analysis.

1.2 Book Structure

This book is structured to guide you through each stage of the analysis:

- **Part I: Introduction and Background**
  - Sets the context for the study, provides an overview of the NHANES data, and outlines the research objectives.
- **Part II: Data Preparation**
  - Details the steps for downloading, cleaning, and merging the raw NHANES and mortality linkage files to create the final analytic dataset.
- **Part III: Statistical Analysis**
  - Elaborates on the complex survey design specification and the various statistical models employed, including survival analysis, effect modification, and sensitivity analyses.
- **Part IV: Discussion**
  - Presents the key findings of the analysis, discusses their implications in the context of the original paper, and outlines limitations and future directions.

Let's get started!

Part I

Introduction and Background

2 Overview of the Study

This section provides an overview of the original study: *Examining the Role of Race/Ethnicity and Sex in Modifying the Association Between Early Smoking Initiation and Mortality: A 20-Year NHANES Analysis*. By outlining the study's motivation, key research questions, analytical approach, and main findings, this chapter sets the stage for the detailed, step-by-step reproduction of the analysis in the chapters that follow.

2.1 How to Use This Guide

This Quarto book is designed to be a transparent and comprehensive guide for reproducing the original analysis.

- **Intended Audience**: This guide is for researchers, public health analysts, and students with a foundational understanding of R and statistical concepts like survival analysis.
- **Required Software**: To run the code in this walkthrough, you will need R (version 4.2.2 or later) and RStudio. The code relies on several key R packages, including `nhanesA`, `survey`, `survival`, and `ggplot2`, which are loaded at the beginning of each relevant chapter.
- **A Note on Reproducibility**: The primary goal is to allow other researchers to validate these findings and build upon them. By documenting the complete analytical pipeline, this guide supports the principles of open and reproducible science.

2.2 Motivation

The motive behind this study was to better understand the link between smoking initiation age and mortality across different demographic groups. Cigarette smoking remains a leading preventable cause of premature death in the U.S., accounting for over 480,000 deaths annually. Clarifying how this risk varies across populations is essential for developing targeted public health interventions that address the specific needs and risks within different communities, thereby enhancing the precision and relevance of health policies.

2.3 Research Objectives

This study addresses two primary objectives:

1. To re-assess the relationship between the initial age of cigarette smoking and overall mortality.
2. To examine how this relationship is modified by race/ethnicity and sex.

To investigate these objectives, the analysis utilizes data from U.S. adults aged 20-79 who participated in the 1999–2018 National Health and Nutrition Examination Survey (NHANES). Mortality data was provided by the National Center for Health Statistics (NCHS) through a linkage with public-use death records.

2.4 Analytical Approach

To address the research objectives, the paper employed multiple analytic approaches. The main survival analysis directly answers the two primary research questions, followed by additional analyses to explore potential mediators and validate the main conclusions.

This tutorial will replicate the following key analyses performed in the original paper:

- **Descriptive Analysis**: Reproduces Appendix Tables 1 & 2 results.
- **Main Survival Analysis**: Establishes the primary link between smoking initiation age and mortality using **Kaplan-Meier Curves**: Figure 1. **Cox Proportional Hazards Model** results are incorporated in Figure 2 in the main paper.
- **Effect Modification Analysis**: It then investigates how this relationship is modified by race/ethnicity and sex. This corresponds to Figure 2 in the main paper and Appendix Tables 4 and 5 in the supplementary material.
- **Exploratory Analysis**: Investigates the secondary relationship between the age of smoking initiation and the total duration of smoking using boxplots. This corresponds to Appendix Figures 1-3 in the supplementary material.
- **Sensitivity Analysis 1**: Adjusts for socioeconomic status (SES) proxies, such as family income and education, to check for confounding. This corresponds to Appendix Figure 5 in the supplementary material.
- **Sensitivity Analysis 2**: Repeats the analysis on data from the 2011–2018 cycles to include the "non-Hispanic Asian" category, which was introduced in the 2011 NHANES survey. This corresponds to Appendix Table 3 and Appendix Figure 4 in the supplementary material.

2.5 Summary of Findings

The final statistical analysis included **50,549 participants**. The study found that early smoking initiation was significantly associated with a higher risk of all-cause mortality across all age groups, with earlier starting ages having higher hazard ratios. Furthermore, this association was observed to differ across race/ethnicity and sex, with the interaction by sex being statistically significant.

3 NHANES Data Overview

This section provides an overview of the National Health and Nutrition Examination Survey (NHANES) program, the primary data source for this analysis. We will cover the survey's design, its core methodology, and the crucial process of linking participant data with national mortality records.

3.1 Description of NHANES

NHANES is a continuous program of studies run by the Centers for Disease Control and Prevention (CDC) that began in the 1960s. It is designed to assess the health and nutritional status of the U.S. population. Since 1999, the program has been conducted in 2-year cycles. Further details about the program can be found on the main NHANES website at the CDC.

3.1.1 Why NHANES for this study?

- **National Representativeness**: NHANES uses a complex, multi-stage probability sampling design to be representative of the entire noninstitutionalized U.S. civilian population.
- **Rich Data**: It collects a wide range of data through interviews, physical examinations, and laboratory tests.
- **Public Availability**: The data files are publicly available online, which allows researchers to easily retrieve and use the data, promoting transparency and reproducibility.

This study utilizes data from 10 consecutive cycles spanning **1999-2018**, providing a robust, long-term dataset for analysis.

3.2 Survey Design

The NHANES survey design is complex and must be accounted for in any analysis to produce valid, generalizable results. The procedure involves four main stages:

1. **Stage 1: Primary Sampling Units (PSUs)** - PSUs, which are mostly counties, are selected with a higher probability for more populated areas.
2. **Stage 2: Segments within PSUs** - Each PSU is divided into smaller geographic areas (like city blocks), and a sample of these segments is drawn.
3. **Stage 3: Households within Segments** - A sample of households is chosen, with over-sampling of certain populations (e.g., low-income persons, specific racial/ethnic groups) to ensure sufficient numbers for meaningful analysis.
4. **Stage 4: Individuals within Households** - Individuals are randomly selected from within the chosen households.

The CDC provides detailed documentation on how to correctly analyze this complex survey data. More information can be found in the NHANES Analytic Guidelines.

> **Why This Matters**
>
> This multi-stage design requires the use of special **survey weights, strata, and PSU variables** in any statistical analysis. **Ignoring these design elements would lead to biased estimates and incorrect standard errors**, potentially invalidating the study's conclusions.

3.3 Mortality Linkage

A key advantage of NHANES is the ability to link participant data with the **National Death Index (NDI)**. The NCHS periodically performs this linkage to create public-use **Linked Mortality Files (LMF)**, which provide mortality follow-up information for adult participants. The official Public-use Linked Mortality Files are available from the NCHS Data Linkage Program.

Key variables from the LMF used in this analysis include:

- **Mortality Status (`MORTSTAT`)**: Indicates whether a participant is deceased or presumed alive at the end of the follow-up period.
- **Follow-up Time (`PERMTH_INT`)**: The duration, in months, from the participant's survey interview date until the date of death or the end of the follow-up period.

By incorporating this mortality data, the cross-sectional NHANES survey is effectively transformed into a **prospective cohort study**. This powerful feature allows us to link baseline

characteristics collected at the time of the survey to long-term health outcomes like all-cause mortality, which is the foundation of our research question.

Part II

Data Preparation

4 Data Download and Merging

This part of the tutorial covers the beginning of the data preparation workflow for the project. It details the process of downloading the raw NHANES data files, selecting relevant variables, and merging the files together. While these steps prepare the data primarily for the main survival analysis, similar preparation is also required for the subsequent sensitivity and exploratory analyses.

> **Note**
>
> In this chapter, we will:
>
> - Programmatically download 20 separate data files from 10 NHANES cycles (1999-2018).
> - Select and process only the necessary variables for the analysis.
> - Merge the demographic and smoking data for each cycle.
> - Harmonize variable names that are inconsistent across different survey years.

We begin by loading the R packages required for this workflow.

4.1 Data Acquisition and Merging

- R Code Chunk 1: Load Necessary Packages

Before we begin the analysis, we import the necessary libraries to ensure we can use the certain functions. The `nhanesA` package is essential, as it provides functions to directly access and import the National Health and Nutrition Examination Survey (NHANES) datasets. In this package, data files from the 10 cycles (1999–2018) are denoted by lettered suffixes. For example, the demographics file for 1999–2000 is named `DEMO`, while subsequent cycles are named `DEMO_B`,

DEMO_C, and up to the letter J. Other key packages are loaded to support data manipulation, variable recoding and merging.

```
# Load required packages
library(nhanesA)
library(car)
library(plyr)
library(dplyr)
library(DataExplorer)
library(ggplot2)
library(grid)
library(gridExtra)
```

- R Code Chunk 2: Define NHANES Datasets by Cycle

The first step in our workflow is to define character vectors using `c()` that contain the specific names of the NHANES data files we need to download. Following the `nhanesA` package's naming convention, we create two vectors: `demo` for the demographic data files and `smoking` for the smoking questionnaire files, each covering the 10 cycles from 1999 to 2018.

```
demo <- c("DEMO", "DEMO_B", "DEMO_C", "DEMO_D", "DEMO_E",
          "DEMO_F", "DEMO_G", "DEMO_H", "DEMO_I", "DEMO_J")

smoking <- c("SMQ", "SMQ_B", "SMQ_C", "SMQ_D", "SMQ_E",
             "SMQ_F", "SMQ_G", "SMQ_H", "SMQ_I", "SMQ_J")
```

- R Code Chunk 3: Download and Process NHANES Data

This code section covers several key steps: downloading the raw data, defining the variables of interest, and then subsetting and processing the data files.

1. Download Raw Data

With the data files' names defined in the lists `demo` and `smoking`, we now use the `lapply()` function to iterate through each vector and download the corresponding data files using `nhanesA::nhanes()`. For reproducibility and to avoid re-downloading the data every time the script is run, we save these lists of raw data files as `.rds` files into a `data/` subdirectory.

> **Note**
>
> We will heavily use the `lapply()` function in this section. This is a more efficient and readable alternative to a for loop in R for applying the same operation to each element of a list (in our case, each data file name).

```
demo_list <- lapply(demo, nhanes)
demo_data_files <- demo_list
saveRDS(demo_data_files, "data/demo_data_files.rds")

smoking_list <- lapply(smoking, nhanes)
smoking_data_files <- smoking_list
saveRDS(smoking_data_files, "data/smoking_data_files.rds")
```

2. Define Variables for Selection

Next, we define the specific variables (columns) we want to keep for our analysis. We create two vectors, `demo_columns` and `smoking_columns`, listing our variables of interest.

It is important to note a key inconsistency across the NHANES cycles: the variable for the participant's country of birth changes its name over time (`DMDBORN`, `DMDBORN2`, `DMDBORN4`). We include all three variations in our list to ensure we capture this information from every cycle. This will be changed into a single variable name later.

For the demographic data, we selected the following key columns:

- `SEQN`: Respondent sequence number (unique identifier).
- `RIDAGEYR`: Age in years at screening.
- `RIAGENDR`: Gender.
- `RIDRETH1`: Race and ethnicity.
- `DMDBORN` / `DMDBORN2` / `DMDBORN4`: Country of birth.
- `SDDSRVYR`: NHANES survey cycle year.
- `WTINT2YR`: Full sample 2-year interview weight.
- `WTMEC2YR`: Full sample 2-year Mobile Examination Center (MEC) exam weight.
- `SDMVPSU`: Masked variance pseudo-Primary Sampling Unit.
- `SDMVSTRA`: Masked variance pseudo-stratum.

For the smoking data, we selected the following key columns:

- `SEQN`: Respondent sequence number.
- `SMQ020`: Indicates if respondent smoked at least 100 cigarettes in life.
- `SMD030`: Age respondent started smoking cigarettes regularly.
- `SMQ040`: Current cigarette smoking status

```
demo_columns <- c("SEQN", "RIDAGEYR", "RIAGENDR", "RIDRETH1",
                  "DMDBORN", "DMDBORN2", "DMDBORN4", "SDDSRVYR",
                  "WTINT2YR", "WTMEC2YR", "SDMVPSU", "SDMVSTRA")

smoking_columns <- c("SEQN", "SMQ020", "SMD030", "SMQ040")
```

3. Subset Data and Translate Codes

The final step in this section is to process the raw data. We iterate through each downloaded data file in the lists, `demo_data_files` and `smoking_data_files`, and perform two key operations:

- Select Columns: We use `dplyr::select()` with the `any_of()` helper to keep only the columns defined above in `demo_columns` and `smoking_columns`. Using `any_of()` prevents errors if a column name (e.g., `DMDBORN2`) doesn't exist in a particular data file.
- Translate Codes: We use `nhanesA::nhanesTranslate()` to convert the numeric codes in the data (e.g., `1` for 'Male') into more descriptive, human-readable factor labels.

The newly processed lists of data frames are stored in `demo_data_files_2` and `smoking_data_files_2`.

```
# DEMOGRAPHICS
demo_data_files_2 <- lapply(seq_along(demo_data_files), function(i) {
  current_cycle_data <- demo_data_files[[i]]
  original <- demo[i]

  # Select Columns
  subset_data <- current_cycle_data %>%
    dplyr::select(dplyr::any_of(demo_columns))
  # Translate
  translated_data <- nhanesTranslate(original,
                                     names(subset_data),
                                     data = subset_data)
  # Return
  return(translated_data)
})
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN2 SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN2 SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN4 SDDSRVYR
```

```
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN4 SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN4 SDDSRVYR
#> Translated columns: RIAGENDR RIDRETH1 DMDBORN4 SDDSRVYR

# SMOKING
smoking_data_files_2 <- lapply(seq_along(smoking_data_files), function(i) {
  current_cycle_data <- smoking_data_files[[i]]
  original <- smoking[i]

  # Select Columns
  subset_data <- current_cycle_data %>%
    dplyr::select(dplyr::any_of(smoking_columns))
  # Translate
  translated_data <- nhanesTranslate(original,
                                     names(subset_data),
                                     data = subset_data)
  # Return
  return(translated_data)
})
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
#> Translated columns: SMQ020 SMQ040
```

- R Code Chunk 4: Merging Datasets

Now that we have two processed lists of data frames (one for demographics and one for smoking), we need to merge them for each NHANES cycle. We iterate through both lists (`demo_data_files_2` and `smoking_data_files_2`), combining each corresponding pair of cycle-specific data frames into a single data frame.

The merge is performed using `plyr::join_all()` with the unique participant identifier, `SEQN`, as the key. We use a `type = 'full'` join to ensure that all participants from both datasets are kept in the final merged data, while `NA` values will be inserted for any non-matching records.

The final output, `data_all`, is a single list containing 10 merged data frames (one for each NHANES cycle).

```
data_all <- lapply(seq_along(demo_data_files_2), function(i) {
  demo_df <- demo_data_files_2[[i]]
  smoking_df <- smoking_data_files_2[[i]]

  # Merge by SEQN
  merged_df <- join_all(list(demo_df, smoking_df),
                        by = "SEQN",
                        type = 'full')
  return(merged_df)
})
```

4.2 Variable Recoding and Cleaning

- R Code Chunk 5: Data Recoding and Cleaning

This code chunk is the most extensive data processing step of the tutorial. Here, we will loop through each of the 10 merged data frames stored in `data_all` to perform data cleaning and recoding. The goal is to create a new, standardized set of variables that are consistent across all survey cycles.

We will create clean variables for participant ID, demographics, smoking behavior, and survey design features.

1. Solve Inconsistent Column Names

As noted previously, the column name for the country of birth is inconsistent across cycles (`DMDBORN`, `DMDBORN2`, `DMDBORN4`). Before we can recode the values, we must first combine these into a single, consistent column name, `DMDBORN`. The following loop handles this standardization.

```
data_all2 <- data_all
for (i in seq_along(data_all2)) {
  df <- as.data.frame(data_all2[[i]])

  if ("DMDBORN2" %in% names(df)) {
    names(df)[names(df) == "DMDBORN2"] <- "DMDBORN"
  } else if ("DMDBORN4" %in% names(df)) {
```

```
    names(df)[names(df) == "DMDBORN4"] <- "DMDBORN"
  }
  data_all2[[i]] <- df
}
```

2. Recoding Reference Table

The table below summarizes the key new variables that will be created below. This serves as a quick reference for the data cleaning process.

New Variable	Description
`id`	Unique participant ID (from `SEQN`)
`age`	Age in years at screening (from `RIDAGEYR`)
`sex`	Biological sex (from `RIAGENDR`)
`race`	Race/ethnicity (from `RIDRETH1`), recoded into White, Black, Hispanic, Others
`born`	Country of birth (from `DMDBORN`), recoded as "Born in US" or "Other place"
`smoking`	Smoking status categorized into Never, Previous, or Current (from `SMQ020` and `SMQ040`)
`smoking.age`	Age participant started smoking (`SMD030`), with special codes 777, 999 replaced by `NA`, and `0` for never smokers
`smoked.while.child`	Derived variable indicating if smoking started at age 15 or younger
`survey.weight`	Full sample 2-year interview weight (from `WTINT2YR`)
`psu`	Masked variance pseudo-Primary Sampling Unit (from `SDMVPSU`)
`strata`	Masked variance pseudo-stratum (from `SDMVSTRA`)
`year`	Survey cycle year (from `SDDSRVYR`)

3. Recoding and Cleaning Data

Now, we will perform the recoding in a series of steps. For clarity in this tutorial, we use a separate loop for each group of variables.

First, we create a simple **ID** variable from the original `SEQN` variable.

```r
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # ID
  dat2$id <- dat2$SEQN

  # Return
  data_all2[[i]] <- dat2
}
```

Next, we recode the core **demographic** variables. This includes creating simple lowercase versions of `age` and `sex`, and collapsing the detailed categories for `race` and `born` into simpler factors using `car::recode()`.

```r
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # Age
  dat2$age <- dat2$RIDAGEYR

  # Sex
  dat2$sex <- dat2$RIAGENDR

  # Race/Ethnicity
  dat2$race <- dat2$RIDRETH1
  dat2$race <- car::recode(dat2$race, recodes = "
    'Non-Hispanic White'='White';
    'Non-Hispanic Black'='Black';
    c('Mexican American','Other Hispanic')='Hispanic';
    else='Others'")
  dat2$race <- factor(dat2$race,
                      levels = c("White", "Black",
                                 "Hispanic", "Others"))

  # Country of birth
  dat2$born <- dat2$DMDBORN
  dat2$born <- car::recode(dat2$born, recodes = "
    c('Born in Mexico','Born Elsewhere', 'Others') = 'Other place';
    c('Born in 50 US States or Washington, DC',
```

```
    'Born in 50 US states or Washington, DC') = 'Born in US';
    else = NA")
  dat2$born <- factor(dat2$born,
                      levels = c("Born in US", "Other place"))

  # Return
  data_all2[[i]] <- dat2
}
```

The tables below summarize the recoding logic for the demographic variables.

Original Variable	Original Categories	New Variable	New Categories
RIDRETH1	Non-Hispanic White	race	White
	Non-Hispanic Black		Black
	Mexican American, Other Hispanic		Hispanic
	Non-Hispanic Asian, Other Race		Others
DMDBORN	Born in 50 US States or Washington, DC	born	Born in US
	Born in Mexico, Born Elsewhere, Others		Other place

Next, we recode the **smoking** variables. A key step here is creating the three-level smoking status factor (`Never`, `Previous`, `Current`). This requires using both `SMQ020` (smoked 100+ cigarettes) and `SMQ040` (smokes at all now) to correctly identify former smokers. We also clean the `smoking.age` variable and create a new binary variable, `smoked.while.child`.

```
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # Smoking Status
  dat2$smoking <- dat2$SMQ020
  dat2$smoking <- car::recode(dat2$smoking, "
    'Yes' = 'Current smoker';
    'No' = 'Never smoker';
    else = NA")
  dat2$smoking <- factor(dat2$smoking,
                         levels = c("Never smoker",
                                    "Previous smoker",
                                    "Current smoker"))
```

```
  # Use SMQ040 to identify former smokers
  dat2$smoking[dat2$SMQ040 == "Not at all?" |
                 dat2$SMQ040 == "Not at all"] <- "Previous smoker"

  # Age Started Smoking
  dat2$smoking.age <- dat2$SMD030
  dat2$smoking.age[dat2$smoking.age %in% c(777, 999)] <- NA
  dat2$smoking.age[is.na(dat2$smoking.age) &
                     dat2$smoking == "Never smoker"] <- 0

  # Whether Smoking started age   15
  dat2$smoked.while.child <- car::recode(dat2$smoking.age,
  "0 = 'No'; 7:15 = 'Yes'; else = NA", as.factor = TRUE)

  # Return
  data_all2[[i]] <- dat2
}
```

The table below summarizes the logic for creating the final `smoking` status variable.

Step	Original Variable(s)	Logic	Resulting Category
1	`SMQ020`	`No`	`Never smoker`
2	`SMQ020`	`Yes`	`Current smoker`
3	`SMQ040`	If `SMQ020` is `Yes` AND `SMQ040` is `Not at all`, re-categorize	`Previous smoker`

For `smoking.age`, numeric codes for "Refused" (`777`) and "Don't know" (`999`) were recoded to `NA`, and a value of `0` was assigned to never smokers for clarity.

Finally, we create lowercase versions of the **survey design** variables for ease of use in later analyses.

```
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]
```

```r
  # Weight
  dat2$survey.weight <- dat2$WTINT2YR

  # PSU
  dat2$psu <- as.factor(dat2$SDMVPSU)

  # Strata
  dat2$strata <- as.factor(dat2$SDMVSTRA)

  # Survey year
  dat2$year <- dat2$SDDSRVYR

  # Return
  data_all2[[i]] <- dat2
}
```

4.3 Finalizing Datasets and Assessing Data Completeness

- R Code Chunk 6: Create, Save, and Plot Datasets

The final step is to create the analytic datasets and save them for use in subsequent chapters. The following code loops through each of the 10 cleaned data frames and performs several actions:

- Selects the final set of cleaned variables.
- Creates the final analytic sample by filtering for eligible participants.
- Generates a missing data plot for each cycle and stores it in a list.
- Saves the data for each cycle into a separate .RData file.

First, we define two vectors: `nhanes_all` contains the desired names for each cycle's data frame (e.g., `nhanes00`, `nhanes01`, etc.), and `vars` lists the set of cleaned variable names we want to keep for the analysis.

Within the loop, the code subsets the data to include only participants aged 20 years or older, matching the age criteria of the original paper. It then saves both the full data frame (all ages) and the filtered analytic data frame (ages 20+) into a single `.RData` file in the `data/` directory (e.g., `analytic00.RData`).

```r
nhanes_all <- c("nhanes00", "nhanes01", "nhanes03", "nhanes05",
                "nhanes07", "nhanes09", "nhanes11",
                "nhanes13", "nhanes15", "nhanes17")

vars <- c("id", "age", "sex", "race", "born",
          "smoking.age", "smoked.while.child", "smoking",
          "survey.weight", "psu", "strata", "year")

missing_plots <- list()


for (i in seq_along(data_all2)) {
  dat2 <- data_all2[[i]]

  nhanes_i <- nhanes_all[i]
  assign(nhanes_i, dat2[, vars], envir = .GlobalEnv)

  analytic <- subset(get(nhanes_i), age >= 20)

  # Create a temporary dataset for plotting, excluding irrelevant variables
  data_for_plot <- analytic %>%
    dplyr::select(-born, -smoked.while.child)

  # Generate a plot, add a title, and remove individual axis titles
  p <- plot_missing(data_for_plot) +
       labs(title = nhanes_i) +
       theme(
         plot.title = element_text(hjust = 0.5),
         legend.position = "none",
         axis.title = element_blank()
       )
  missing_plots[[i]] <- p

  cat("Processing:", nhanes_i, "\n")
  print(dim(analytic))

  analytic_i <- paste0("analytic", substr(nhanes_i, 7, 8))
  assign(analytic_i, analytic, envir = .GlobalEnv)

  # Create 'data' directory if it does not exist
  if (!dir.exists("data")) {
```

```
    dir.create("data")
  }

  # Save
  save(list = c(nhanes_i, analytic_i),
       file = file.path("data", paste0(analytic_i, ".RData")))
}
#> Processing: nhanes00
#> [1] 4880   12
#> Processing: nhanes01
#> [1] 5411   12
#> Processing: nhanes03
#> [1] 5041   12
#> Processing: nhanes05
#> [1] 4979   12
#> Processing: nhanes07
#> [1] 5935   12
#> Processing: nhanes09
#> [1] 6218   12
#> Processing: nhanes11
#> [1] 5560   12
#> Processing: nhanes13
#> [1] 5769   12
#> Processing: nhanes15
#> [1] 5719   12
#> Processing: nhanes17
#> [1] 5569   12
```

Finally, the 10 missing data plots are displayed in a single 5x2 grid.

```
# Define common axis labels
y_label <- textGrob("Features", gp = gpar(fontsize = 12), rot = 90)
x_label <- textGrob("Missing Rows", gp = gpar(fontsize = 12))

# Arrange the plots in a grid with common labels
grid.arrange(
  grobs = missing_plots,
  nrow = 5,
  ncol = 2,
  left = y_label,
  bottom = x_label
)
```

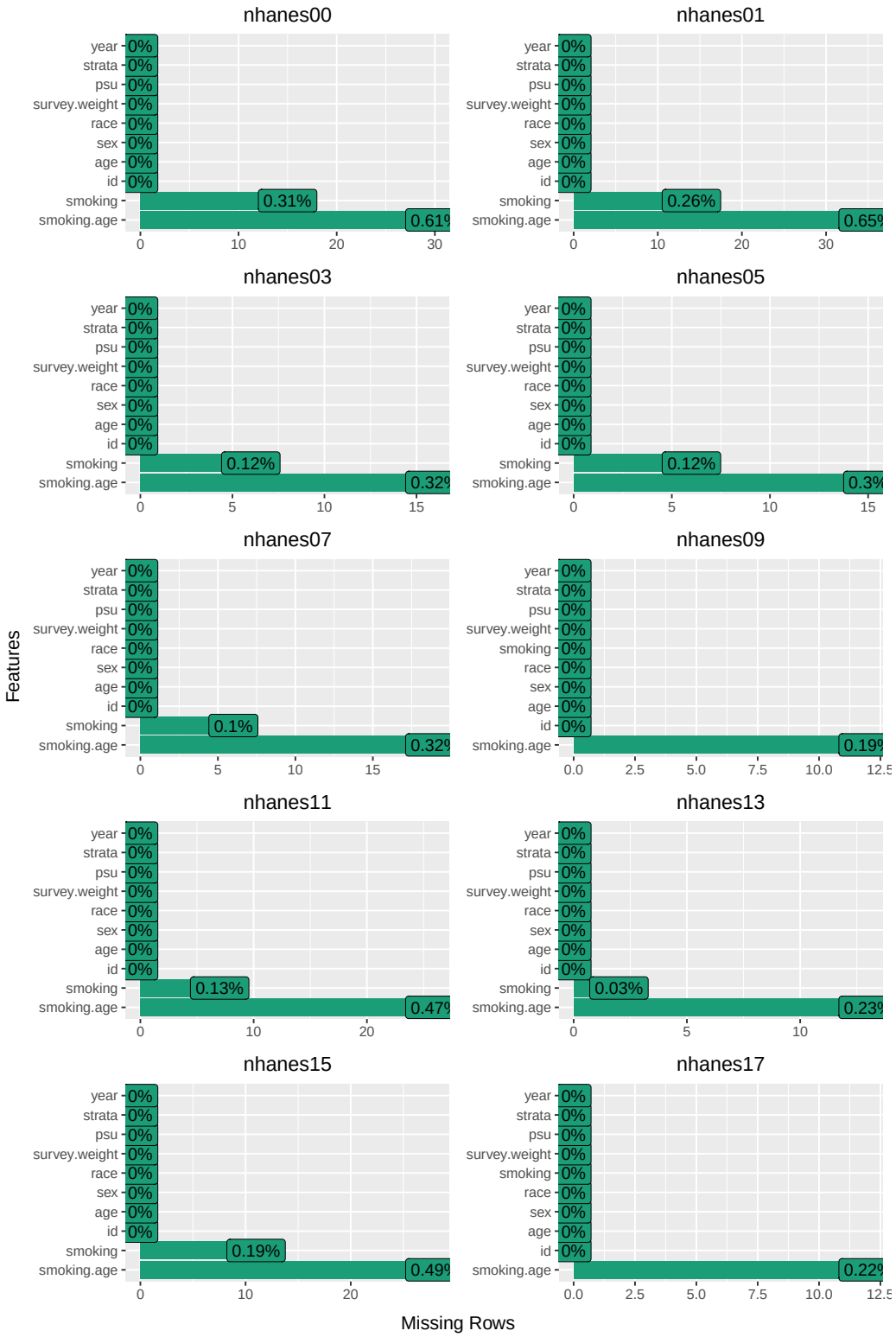


Figure 4.1: Missing data patterns for analytic variables across all 10 NHANES cycles (1999-2018). The x-axis represents the number of missing observations.

This concludes the first part of the data preparation stage.

4.4 Chapter Summary and Next Steps

We have now successfully completed the initial data acquisition and preparation phase. We have downloaded 20 raw data files, processed them into a consistent format, and merged them into a single list of 10 data frames—one for each NHANES cycle.

In the next chapter, "Mortality and NHANES Merging," we will combine these 10 data frames and link them with the public-use mortality data to create the final, comprehensive dataset for our analysis.

5 Mortality and NHANES Merging

This section continues the data preparation process. In the previous part, we created 10 cleaned datasets, one for each NHANES cycle from 1999 to 2018. Now, we will combine those 10 datasets into a single data frame and then link it with the all-cause mortality follow-up data from the National Death Index (NDI).

```
# Load required packages
library(readr)
```

5.1 Combining the NHANES Survey Cycles

- R Code Chunk 1: Load Processed NHANES Data

The following code begins by loading the 10 processed `.RData` files that were created and saved in the previous section. Each file contains the cleaned demographic and smoking variables for participants aged 20 years and older for a specific survey cycle. The `ls()` command is used to list the objects now loaded in our R environment.

```
load(file="data/analytic00.RData")
load(file="data/analytic01.RData")
load(file="data/analytic03.RData")
load(file="data/analytic05.RData")
load(file="data/analytic07.RData")
load(file="data/analytic09.RData")
load(file="data/analytic11.RData")
load(file="data/analytic13.RData")
load(file="data/analytic15.RData")
load(file="data/analytic17.RData")
ls()
#>  [1] "analytic00"      "analytic01"      "analytic03"      "analytic05"
#>  [5] "analytic07"      "analytic09"      "analytic11"      "analytic13"
#>  [9] "analytic15"      "analytic17"      "has_annotations" "nhanes00"
```

```
#> [13] "nhanes01"        "nhanes03"        "nhanes05"        "nhanes07"
#> [17] "nhanes09"        "nhanes11"        "nhanes13"        "nhanes15"
#> [21] "nhanes17"
```

- R Code Chunk 2: Combine NHANES Cycles and Adjust Weights

Now that the 10 separate NHANES datasets are loaded into the environment, we need to combine them into one single data frame and adjust the survey weights for the pooled data.

1. Combine Datasets

The following code combines the 10 individual analytic data frames into a single, large data frame named `dat.full`. The `rbind()` function is used to stack the data frames by row. After this step, we run several checks to confirm that all 10 survey cycles are present and to see the final dimensions of the combined dataset. There are now 101316 rows in the single dataframe.

```
# Full dataset : 1999-2018
dat.full <- rbind(nhanes00, nhanes01, nhanes03, nhanes05,
                  nhanes07, nhanes09, nhanes11,
                  nhanes13, nhanes15, nhanes17)

unique(dat.full$year)
#>  [1] NHANES 1999-2000 Public Release NHANES 2001-2002 Public Release
#>  [3] NHANES 2003-2004 Public Release NHANES 2005-2006 Public Release
#>  [5] NHANES 2007-2008 Public Release NHANES 2009-2010 Public Release
#>  [7] NHANES 2011-2012 public release NHANES 2013-2014 public release
#>  [9] NHANES 2015-2016 public release NHANES 2017-2018 public release
#> 10 Levels: NHANES 1999-2000 Public Release ... NHANES 2017-2018 public release
length(unique(dat.full$year))
#> [1] 10
dim(dat.full)
#> [1] 101316     12
```

2. Adjust Survey Weights

Next, the following code adjusts the survey weights. When combining multiple NHANES cycles, the original two-year survey weights must be redistributed. We create a new weight variable, `survey.weight.new`, by dividing the original weight by the number of survey cycles (10). The original weight column is then removed.

```r
# Corrected Weights
dat.full$survey.weight.new <- dat.full$survey.weight /
  length(unique(dat.full$year))

# Remove original variable
dat.full$survey.weight <- NULL

# Print Columns
names(dat.full)
#>  [1] "id"                 "age"                "sex"
#>  [4] "race"               "born"               "smoking.age"
#>  [7] "smoked.while.child" "smoking"            "psu"
#> [10] "strata"             "year"               "survey.weight.new"
```

5.2 Acquiring the Linked Mortality Data

- R Code Chunk 3: Download and Read Mortality Data

This section covers downloading the public-use mortality data linked to the NHANES participants.

1. Create Data Directories

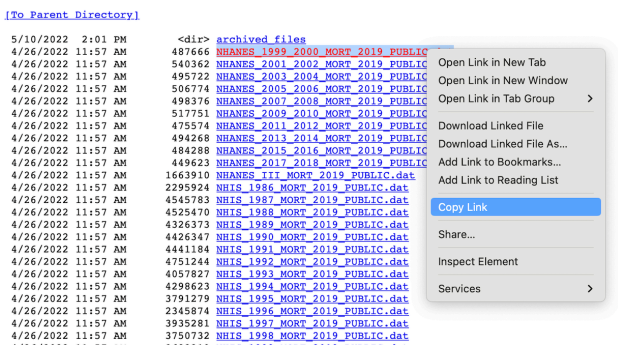
Before downloading the data, we first ensure the necessary subdirectories (`data/` and `data/Mortalitydata/`) exist in our project folder for our computer's workspace. The following code checks for these directories and creates them if they are missing, which prevents errors when saving files.

```r
# Create directories only if they don't exist
if (!dir.exists("data")) {
  dir.create("data")
}
if (!dir.exists("data/Mortalitydata")) {
  dir.create("data/Mortalitydata")
}
```

2. Read Fixed-Width Format Files

The public-use mortality data is provided by NCHS in a fixed-width text file format (.dat). In this format, each variable occupies a specific character position on each line. We need to read the file for each NHANES cycle and extract the necessary columns.

ftp.cdc.gov - /pub/Health_Statistics/NCHS/datalinkage/linked_mortality/



To read this type of file, we use the `readr::read_fwf()` function. We must provide `fwf_cols()` with the exact start and end positions for each variable we want to extract, based on the official NCHS documentation.

The following code demonstrates this process for the 1999-2000 mortality file, reading it directly from the CDC's server by copying the link (shown in image). For this tutorial, this same process was repeated for all 10 survey cycles.

5.2.1 Mortality data 1999-2000

```
mort2000 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
                              mort_permth_int = c(43,45),
                              mort_permth_exm = c(46,48)),
                     na = c("", "."))

colnames(mort2000)
#> [1] "id"                "mort_eligstat"     "mort_stat"
#> [4] "mort_ucod_leading" "mort_diabetes"     "mort_hyperten"
#> [7] "mort_permth_int"   "mort_permth_exm"
```

```
head(mort2000)
#> # A tibble: 6 x 8
#>      id mort_eligstat mort_stat mort_ucod_leading mort_diabetes mort_hyperten
#>   <int>         <int>     <int>             <int>         <int>         <int>
#> 1     1             2        NA                NA            NA            NA
#> 2     2             1         1                 6             0             0
#> 3     3             2        NA                NA            NA            NA
#> 4     4             2        NA                NA            NA            NA
#> 5     5             1         0                NA            NA            NA
#> 6     6             1         0                NA            NA            NA
#> # i 2 more variables: mort_permth_int <int>, mort_permth_exm <int>
```

5.2.2 Mortality data 2001-2002

```
mort2001 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
                              mort_permth_int = c(43,45),
                              mort_permth_exm = c(46,48)),
                     na = c("", "."))
#colnames(mort2001)
#head(mort2001)
```

5.2.3 Mortality data 2003-2004

```
mort2003 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
```

```
                        mort_permth_int = c(43,45),
                        mort_permth_exm = c(46,48)),
                    na = c("", "."))

#colnames(mort2003)
#head(mort2003)
```

5.2.4 Mortality data 2005-2006

```
mort2005 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                    col_types = "iiiiiiii",
                    fwf_cols(id = c(1,6),
                        mort_eligstat = c(15,15),
                        mort_stat = c(16,16),
                        mort_ucod_leading = c(17,19),
                        mort_diabetes = c(20,20),
                        mort_hyperten = c(21,21),
                        mort_permth_int = c(43,45),
                        mort_permth_exm = c(46,48)),
                    na = c("", "."))

#colnames(mort2005)
#head(mort2005)
```

5.2.5 Mortality data 2007-2008

```
mort2007 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                    col_types = "iiiiiiii",
                    fwf_cols(id = c(1,6),
                        mort_eligstat = c(15,15),
                        mort_stat = c(16,16),
                        mort_ucod_leading = c(17,19),
                        mort_diabetes = c(20,20),
                        mort_hyperten = c(21,21),
                        mort_permth_int = c(43,45),
                        mort_permth_exm = c(46,48)),
                    na = c("", "."))
```

```
#colnames(mort2007)
#head(mort2007)
```

5.2.6 Mortality data 2009-2010

```
mort2009 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
                              mort_permth_int = c(43,45),
                              mort_permth_exm = c(46,48)),
                     na = c("", "."))

#colnames(mort2009)
#head(mort2009)
```

5.2.7 Mortality data 2011-2012

```
mort2011 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
                              mort_permth_int = c(43,45),
                              mort_permth_exm = c(46,48)),
                     na = c("", "."))

#colnames(mort2011)
#head(mort2011)
```

5.2.8 Mortality data 2013-2014

```
mort2013 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
                              mort_permth_int = c(43,45),
                              mort_permth_exm = c(46,48)),
                     na = c("", "."))

#colnames(mort2013)
#head(mort2013)
```

5.2.9 Mortality data 2015-2016

```
mort2015 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
                              mort_permth_int = c(43,45),
                              mort_permth_exm = c(46,48)),
                     na = c("", "."))

#colnames(mort2015)
#head(mort2015)
```

5.2.10 Mortality data 2017-2018

```
mort2017 <- read_fwf(file = "https://ftp.cdc.gov/pub/Health_Statistics/NCHS/datalinkage/li
                     col_types = "iiiiiiii",
                     fwf_cols(id = c(1,6),
                              mort_eligstat = c(15,15),
                              mort_stat = c(16,16),
                              mort_ucod_leading = c(17,19),
                              mort_diabetes = c(20,20),
                              mort_hyperten = c(21,21),
                              mort_permth_int = c(43,45),
                              mort_permth_exm = c(46,48)),
                     na = c("", "."))

#colnames(mort2017)
#head(mort2017)
```

After downloading, each of the 10 mortality data frames is saved as an `.RData` file.

```
saveRDS(mort2000, file = "data/Mortalitydata/mort2000.RData")
saveRDS(mort2001, file = "data/Mortalitydata/mort2001.RData")
saveRDS(mort2003, file = "data/Mortalitydata/mort2003.RData")
saveRDS(mort2005, file = "data/Mortalitydata/mort2005.RData")
saveRDS(mort2007, file = "data/Mortalitydata/mort2007.RData")
saveRDS(mort2009, file = "data/Mortalitydata/mort2009.RData")
saveRDS(mort2011, file = "data/Mortalitydata/mort2011.RData")
saveRDS(mort2013, file = "data/Mortalitydata/mort2013.RData")
saveRDS(mort2015, file = "data/Mortalitydata/mort2015.RData")
saveRDS(mort2017, file = "data/Mortalitydata/mort2017.RData")
```

5.3 Merging NHANES with Mortality Data

- R Code Chunk 4: Combine Mortality Datasets

The following code combines the 10 individual mortality data frames into a single, large data frame named `dat.mortality`. The `rbind()` function is used to combine the data frames vertically by stacking rows. The `dim()` and `head()` functions are then used to confirm the

dimensions and inspect the first few rows of the combined dataset. There are 101,316 rows in the dataset.

```
dat.mortality <- rbind(mort2000, mort2001, mort2003, mort2005,
                       mort2007, mort2009, mort2011,
                       mort2013, mort2015, mort2017)
dim(dat.mortality)
#> [1] 101316      8
```

Examine Mortality Eligibility Status:

After combining the datasets, it's important to examine the `mort_eligstat` variable. This variable indicates the eligibility status of each participant for the mortality follow-up study. The codes are defined as follows:

`1`: Eligible for mortality follow-up.

`2`: Under age 18 and not available for public release.

`3`: Ineligible for other reasons.

`NA`: Missing eligibility status.

The `table()` function is used to see the distribution of participants across these categories.

```
table(dat.mortality$mort_eligstat, useNA = "always")
#>
#>     1     2     3  <NA>
#> 59064 42112   140     0
```

- R Code Chunk 5: Final Merge of NHANES and Mortality Data

The following code performs the final merge, joining the combined NHANES dataset (`dat.full`) with the combined mortality dataset (`dat.mortality`). The two datasets are linked using the unique participant identifier, `id`. By setting the argument `all.x = TRUE`, all participants from the combined NHANES dataset are kept, regardless of whether they have a matching record in the combined mortality dataset.

```
dat.full.with.mortality <- merge(dat.full, dat.mortality,
                                 by = "id", all.x = TRUE)

dim(dat.full.with.mortality)
#> [1] 101316     19
```

- R Code Chunk 6: Save Final Dataset

Finally, the complete, merged dataset is saved as an RDS file. This file, `dat.full.with.mortality.RDS`, will be the starting point for the statistical analysis in the next chapter.

```
saveRDS(dat.full.with.mortality,
        file = "data/dat.full.with.mortality.RDS")
```

This concludes the second part of the data preparation stage.

5.4 Chapter Summary and Next Steps

This chapter concluded the data preparation phase. We first combined the 10 individual NHANES survey cycles into a single, comprehensive dataset and adjusted the survey weights for the pooled 20-year period.

Subsequently, we acquired the public-use mortality follow-up data and merged it with the NHANES data by participant ID. The final output of this chapter is a single, complete dataset that links baseline survey data to long-term mortality outcomes, forming the foundation for all statistical analyses in the chapters that follow.

6 Variable Definitions and Cleaning

This section continues the data preparation process using the combined NHANES and mortality dataset created in the previous section, `dat.full.with.mortality`. Here, we will construct the final variables required for our survival analysis. This includes defining the exposure, outcome, and follow-up time, as well as creating the final analytic dataset by applying inclusion criteria and handling missing data.

```
# Load required packages
library(car)
#> Loading required package: carData
library(DataExplorer)
```

- R Code Chunk 1: Load Complete Merged Data

The following code begins by loading the `dat.full.with.mortality.RDS` file. This dataset is the complete, merged result from the previous chapter, containing the processed demographic, smoking, and mortality data for all participants. The `readRDS()` function loads the file from the `data/` directory, and the `colnames()` function is used to display the names of all its columns.

```
# Load previous data file
dat.full.with.mortality <-
  readRDS("data/dat.full.with.mortality.RDS")

colnames(dat.full.with.mortality)
#>  [1] "id"                 "age"                "sex"
#>  [4] "race"               "born"               "smoking.age"
#>  [7] "smoked.while.child" "smoking"            "psu"
#> [10] "strata"             "year"               "survey.weight.new"
#> [13] "mort_eligstat"      "mort_stat"          "mort_ucod_leading"
#> [16] "mort_diabetes"      "mort_hyperten"      "mort_permth_int"
```

```
#> [19] "mort_permth_exm"
```

6.1 Creating Key Analysis Variables

- R Code Chunk 2: Variable Construction

This section details the construction of the key variables essential for our survival analysis. This includes defining the exposure variable (age at smoking initiation), the survival outcome, and the follow-up time.

1. Exposure Variable: Age at Smoking Initiation Categories (exposure.cat)

The following code creates the primary exposure variable for our analysis, `exposure.cat`. The numeric `smoking.age` variable (`SMD030` from NHANES) is categorized into six levels based on age ranges used in the original paper. This categorization is based on established age ranges for smoking initiation used in previous research, and aims to capture distinct developmental and addiction risk profiles. We first inspect the raw distribution of `smoking.age` before using `car::recode()` to create the new factor and explicitly setting the level order to ensure "Never smoked" is the reference category.

```
# Display initial distribution of raw smoking.age
table(dat.full.with.mortality$smoking.age, useNA = "always")
#>
#>     0     5     6     7     8     9    10    11    12    13    14    15    16
#> 31915     4     3   106   100   157   261   240   841  1207  1597  2354  2973
#>    17    18    19    20    21    22    23    24    25    26    27    28    29
#>  2399  3591  1510  1755   995   708   415   263   760   166   173   156    63
#>    30    31    32    33    34    35    36    37    38    39    40    41    42
#>   370    26    71    29    22   143    34    24    28    19    98     9    18
#>    43    44    45    46    47    48    49    50    51    52    53    54    55
#>    14    11    44    11     7     6     5    25     4     2     1     4     8
#>    56    57    58    59    60    62    63    64    65    68    70    71    72
#>     2     1     5     2     2     2     1     1     4     4     1     1     4
#>    74    76    77  <NA>
#>     2     1     1 45537

# Recode smoking.age into categorical exposure.cat
dat.full.with.mortality$exposure.cat <- car::recode(
dat.full.with.mortality$smoking.age, "
```

```
0 = 'Never smoked'; 1:9 = 'Started before 10';
10:14 = 'Started at 10-14'; 15:17 = 'Started at 15-17';
18:20 = 'Started at 18-20'; 21:80 = 'Started after 20';
else = NA ", as.factor = TRUE)

# Explicitly set the order of factor levels
dat.full.with.mortality$exposure.cat <-
  factor(dat.full.with.mortality$exposure.cat,
         levels = c("Never smoked", 'Started before 10',
                    'Started at 10-14', "Started at 15-17",
                    "Started at 18-20", "Started after 20"))
```

The following tables are used to inspect the distribution of the newly created exposure variable, both overall and broken down by mortality status.

```
# Display distributions of the newly created exposure variable
table(dat.full.with.mortality$exposure.cat, useNA = "always")
#>
#>      Never smoked Started before 10  Started at 10-14  Started at 15-17
#>             31915               370              4146              7726
#>  Started at 18-20  Started after 20              <NA>
#>              6856              4766             45537


# Display distributions of exposure.cat level
# broken down by mort_stat variable
table(dat.full.with.mortality$exposure.cat,
      dat.full.with.mortality$mort_stat, useNA = "always")
#>
#>                          0     1  <NA>
#>   Never smoked       27823  3994    98
#>   Started before 10    254   116     0
#>   Started at 10-14    3217   922     7
#>   Started at 15-17    6193  1522    11
#>   Started at 18-20    5438  1405    13
#>   Started after 20    3618  1140     8
#>   <NA>                3272   150 42115
```

2. Outcome and Follow-up Time Variables

The following code constructs the variables for the survival analysis. As noted in the original paper, the time of birth is designated as the baseline to prevent differential start times for exposed and unexposed participants. The primary survival outcome is therefore the time from

birth to all-cause mortality. Specifically, this is calculated by combining the participant's age at the interview (from the RIDAGEYR variable) with their mortality follow-up time from the interview date (from the PERMTH_INT variable). We also define the `status_all` variable from the raw mortality data to indicate the participant's final mortality status.

```
# Survival time Person-Months of Follow-up from Interview date
# (PERMTH_INT)
dat.full.with.mortality$stime.since.interview <-
  dat.full.with.mortality$mort_permth_int
summary(dat.full.with.mortality$stime.since.interview)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>     0.0    59.0   113.0   117.9   172.0   250.0   42252


# Age in month at NHANES interview
# (RIDAGEYR)
dat.full.with.mortality$age.month <- dat.full.with.mortality$age * 12
summary(dat.full.with.mortality$age.month)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>     0.0   120.0   288.0   373.5   624.0  1020.0


# Survival time - Person-Months of Follow-up from birth = age at
# screening (months) + person-months of follow-up from interview.
# This combines the age at interview with the time from interview
# to death or end of follow-up.
dat.full.with.mortality$stime.since.birth <-
  with(dat.full.with.mortality, age.month + stime.since.interview)


# Convert to years for consistent interpretation with age
dat.full.with.mortality$stime.since.birth <- dat.full.with.mortality$stime.since.birth/12


# Ensure NA values are consistent if stime.since.interview was NA
dat.full.with.mortality$stime.since.birth[
  is.na(dat.full.with.mortality$stime.since.interview)] <- NA
summary(dat.full.with.mortality$stime.since.birth)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>   19.08   41.50   57.17   57.56   72.92  105.67   42252
```

```r
# All-cause mortality status (PRIMARY outcome)
dat.full.with.mortality$status_all <-
  dat.full.with.mortality$mort_stat
table(dat.full.with.mortality$status_all, useNA = "always")
#>
#>     0     1  <NA>
#> 49815  9249 42252
```

3. Final Variable Cleaning

The following code drops mortality-related variables not needed for the final analytic dataset.

```r
# Remove unneeded columns
dat.full.with.mortality$mort_diabetes <- NULL
dat.full.with.mortality$mort_hyperten <- NULL
dat.full.with.mortality$mort_permth_int <- NULL
dat.full.with.mortality$mort_permth_exm <- NULL
dat.full.with.mortality$mort_ucod_leading <- NULL
dat.full.with.mortality$mort_eligstat <- NULL
```

4. Recoding Survey Year: year.cat

The `year` variable, representing the NHANES survey cycle is recoded into a more descriptive categorical variable, `year.cat`.

```r
# Recode year variable into a factor with descriptive labels
dat.full.with.mortality$year.cat <- dat.full.with.mortality$year

new_levels <- c(
  "1999-2000", "2001-2002", "2003-2004",
  "2005-2006", "2007-2008", "2009-2010",
  "2011-2012", "2013-2014", "2015-2016",
  "2017-2018"
)

levels(dat.full.with.mortality$year.cat) <- new_levels

# Display the unique levels of the new year.cat variable
levels(dat.full.with.mortality$year.cat)
#>  [1] "1999-2000" "2001-2002" "2003-2004" "2005-2006" "2007-2008" "2009-2010"
#>  [7] "2011-2012" "2013-2014" "2015-2016" "2017-2018"
```

6.2 Constructing the Final Analytic Cohort

- R Code Chunk 3: Final Analytic Dataset Construction

The following code creates the final analytic dataset by applying the specific inclusion and exclusion criteria used in the published paper. The two main steps are restricting the participant age range to 20–79 years and performing a complete-case analysis by excluding observations with missing data for key variables.

1. Apply Age and Variable Exclusions

First, we apply the age restriction, keeping only participants between 20 and 79 years old. We also drop several intermediate or raw variables that are no longer needed for the final analysis. We will remove the `smoked.while.child` variable. While this variable was created as a simple binary indicator for early smoking, the final analysis relies on the more granular, multi-level `exposure.cat` variable.

> **Note**
>
> The `exposure.cat` variable (e.g., 'Started before 10', 'Started at 10-14') is superior for our research question because it allows us to investigate a dose-response relationship—that is, to see if the risk of mortality changes incrementally with different ages of initiation. A simple binary variable like `smoked.while.child` would only allow us to compare "early starters" to everyone else, masking the important nuances in risk across different age groups. Therefore, it is excluded from the final analytic dataset.

```
# Drop the following column not being used
dat.full.with.mortality$smoked.while.child <- NULL

### Analytic Dataset - age 20 - 79

# Inclusion criteria: adults aged between 20 and 79 years
dat.analytic <- subset(dat.full.with.mortality, age>=20 & age < 80)

# Dimensions before age filtering
dim(dat.full.with.mortality)
#> [1] 101316     18

# Dimensions after age filtering
dim(dat.analytic)
#> [1] 50824    18

# Calculate number of participants dropped due to age restriction
```

```r
dim(dat.full.with.mortality)[1] - dim(dat.analytic)[1]
#> [1] 50492
```

Drop other variables that are not being used in the final analysis.

```r
dat.analytic$born <- NULL
dat.analytic$age <- NULL
dat.analytic$age.month <- NULL

# Dimension
dim(dat.analytic)
#> [1] 50824    15
```

2. Apply Complete-Case Analysis

Next, we exclude participants who have missing data for the primary outcome (survival time) or the primary exposure (smoking initiation category). This is consistent with the original paper, which notes that about 0.5% of the sample (275 observations) was discarded for this reason. Also, there were no missing values in the covariates variables considered in the main analysis.

```r
# Participants with incomplete data on smoking status
# or mortality were excluded
dat.analytic1 <- dat.analytic[
  complete.cases(dat.analytic$stime.since.birth),]
dim(dat.analytic1) # Dimension after removing missing survival time
#> [1] 50690    15

dat.analytic2 <- dat.analytic1[
  complete.cases(dat.analytic1$exposure.cat),]
dim(dat.analytic2) # Dimension after removing missing exposure var
#> [1] 50549    15

# Complete case analysis: remove missing values in covariates
dat.complete <- na.omit(dat.analytic2) # No missing values
dim(dat.complete)
#> [1] 50549    15
```

```r
# Profile missingness after initial filtering
profile_missing(dat.analytic2)
#>                  feature num_missing pct_missing
#> 1                     id           0           0
```

```
#> 2                   sex           0           0
#> 3                  race           0           0
#> 4            smoking.age          0           0
#> 5                smoking          0           0
#> 6                   psu           0           0
#> 7                strata           0           0
#> 8                  year           0           0
#> 9      survey.weight.new          0           0
#> 10             mort_stat          0           0
#> 11          exposure.cat          0           0
#> 12 stime.since.interview          0           0
#> 13     stime.since.birth          0           0
#> 14            status_all          0           0
#> 15              year.cat          0           0
```

3. Reconcile Sample Sizes The following code calculates the number of participants dropped at each stage of the filtering process. This helps confirm that our data cleaning has resulted in the same final sample size as the original study.

```
# Participants dropped - overall
nrow(dat.full.with.mortality) - nrow(dat.complete)
#> [1] 50767

# Participants dropped - due to missing exposure or outcome
nrow(dat.analytic) - nrow(dat.analytic2)
#> [1] 275

# Participants dropped - due to missing covariates
nrow(dat.analytic2) - nrow(dat.complete)
#> [1] 0
```

6.3 Exporting Final Datasets

- R Code Chunk 4: Save Final Datasets

The final step is to save our processed datasets. We save three different versions of the data to create useful checkpoints for the analysis:

- `dat.full.with.mortality.RData`: The updated, merged dataset before any filtering was applied in this chapter.
- `dat.analytic2.RData`: The dataset after applying age restrictions and removing participants with missing outcome or exposure data.
- `dat.complete.RData`: The final, clean analytic dataset containing only complete cases.

```
# Save
save(dat.full.with.mortality,
     file = "data/dat.full.with.mortality.RData")

save(dat.analytic2,
     file = "data/dat.analytic2.RData")

save(dat.complete,
     file = "data/dat.complete.RData")
```

This completes the variable definitions and cleaning part of the data preparation stage.

6.4 Chapter Summary and Next Steps

In this chapter, we constructed the core variables for the analysis, including the categorized smoking initiation exposure (`exposure.cat`) and the survival outcome (time from birth to all-cause mortality). We then applied the study's inclusion and exclusion criteria to create the final, complete-case analytic dataset with 50,549 participants, matching the sample size of the original paper.

With a clean dataset ready, the next chapter, "Cohort Definition," will formally summarize the criteria used to define our study population.

7 Cohort Definition

This section details the criteria used to define the study cohort, including participant inclusion and exclusion criteria, and the approach taken to handle missing data.

7.1 Inclusion/Exclusion Criteria

The study sample was derived from 10 cycles of the National Health and Nutrition Examination Surveys (NHANES), spanning from 1999 to 2018. The participants included in this study from the NHANES data were adults aged between 20 and 79 years old.

Key exclusion criteria were as follows:

- Individuals younger than 20 or older than 79 years were excluded from the study.
- Individuals with incomplete data concerning their smoking status or mortality outcomes were also excluded.
- Some participants were not included in the public-use mortality files because their records lacked the minimum identifying data required for a successful linkage to the National Death Index

7.2 Missing Data Handling

The study performed a complete-case analysis, meaning only participants with complete data for all variables in the main analysis were included. The final data set comprised a sample size of 50,549 individuals. A total of 275 participants (about 0.5% of the sample) were removed due to missing information on the primary exposure (smoking initiation age) or the outcome (mortality). The main covariates used in the analysis such as race/ethnicity, sex, and survey cycle had no missing values. However, other variables related to socioeconomic status, like family income ratio and education level, were not included in the primary analysis due to a high degree of missing data. The authors chose not to impute these missing values, believing that a reliable model could not be built from the available data.

7.3 Chapter Summary and Next Steps

This chapter formally outlined the inclusion and exclusion criteria that defined our final study cohort of 50,549 participants. We also reviewed the complete-case approach used to handle missing data for the primary exposure and outcome variables.

Now that our cohort is clearly defined, we will proceed to "Survey Design Specification," a critical step where we tell R how to correctly account for the complex sampling design of the NHANES data.

Part III

Statistical Analysis

8 Survey Design Specification

This section begins the statistical portion of the analysis by specifying the complex survey design of the NHANES data. In the previous chapters, the raw NHANES and mortality data were downloaded, cleaned, and merged, resulting in the `dat.full.with.mortality` and `dat.complete` datasets. Now, we will use the survey package in R to correctly account for the complex sampling design, which is a critical step for all subsequent analyses.

```
# Load required packages
library(survey)
options(survey.want.obsolete=TRUE)
```

8.1 Loading the Analytic Datasets

- R Code Chunk 1: Load Data

The following code loads the datasets created in the previous steps. `dat.full.with.mortality` contains the complete merged data, while `dat.complete` and `dat.analytic2` contain the filtered subsets that will be used to define the analytic sub-cohort for the survey design.

```
# Load the full merged dataset
load(file = "data/dat.full.with.mortality.RData")

# Load the intermediate analytic dataset
# Used for subsetting the design
load(file = "data/dat.analytic2.RData")
load(file = "data/dat.complete.RData")

# Check dimensions
dim(dat.full.with.mortality)
#> [1] 101316     18
dim(dat.analytic2)
#> [1] 50549    15
```

```r
dim(dat.complete)
#> [1] 50549    15
```

Note: The `dat.complete` and `dat.analytic2` datasets are identical. This confirms that after participants with missing exposure or outcome data were removed, no missing values remained for the covariates used in the main analysis, as stated in the original paper

---

8.2 Specifying the Survey Design

- R Code Chunk 2: Specify the Survey Design Object

To correctly analyze NHANES data, we must account for its complex sampling design, which includes stratification, clustering, and unequal weighting. The following code uses the `survey` package to create a survey design object. This step is critical for obtaining accurate standard errors and p-values in all subsequent analyses.

> **Note**
>
> Accounting for the complex sampling design is not just a formality; it is the most critical step for ensuring our results are statistically valid.
> We will use the `svydesign()` function from the survey package to create a "survey design object." This object bundles our dataset with the information on how the data was collected (the weights, strata, and PSUs). All of our subsequent analyses will be performed on this object, not on the raw data frame.
> Failing to perform this step and instead running analyses on the raw data would lead to severely underestimated standard errors and incorrect p-values, likely resulting in false positive findings. By specifying the design, we ensure our results are nationally representative and that our statistical inferences are trustworthy.

1. The `svydesign` Function

We use the `svydesign()` function to combine the dataset with its design information. The key arguments are:

- `ids = ~psu`: Specifies the primary sampling unit (PSU) variable.
- `strata = ~strata`: Specifies the stratification variable.
- `weights = ~survey.weight.new`: Specifies the adjusted survey weight variable.
- `nest = TRUE`: Correctly handles PSUs that are nested within strata.

2. Subsetting the Design Object

For correct variance estimation, the `survey` package requires us to define the design on the full dataset first and then specify the subset to be analyzed. We do this by creating a `miss` variable on the `dat.full.with.mortality` dataset. We then create the full design object (`w.design`) and use the `subset()` function to create the final analytic design object (`w.design0`). This final object includes only our cohort of interest (where `miss == 0`) and participants with positive survey weights.

```
# Set up the 'miss' indicator for subsetting the design
# Initialize 'miss' as 1 (excluded) for all
# in dat.full.with.mortality
dat.full.with.mortality$miss <- 1

# Set 'miss' to 0 (included) for participants whose 'id'
# is in dat.analytic2
dat.full.with.mortality$miss[dat.full.with.mortality$id
                             %in% dat.analytic2$id] <- 0

# Display the distribution of the 'miss' variable
table(dat.full.with.mortality$miss)
#>
#>     0     1
#> 50549 50767

# Create the full survey design object
w.design <- svydesign(ids = ~psu,
                      strata = ~strata,
                      weights = ~survey.weight.new,
                      data = dat.full.with.mortality,
                      nest = TRUE)

# Subset the design to the analytic cohort (miss == 0) and
# positive weights
w.design0 <- subset(w.design, miss == 0 & survey.weight.new > 0)

# Verify the number of observations in the subsetted design
cat("Number of observations in w.design0 (subsetted design):",
    nrow(w.design0), "\n")
#> Number of observations in w.design0 (subsetted design): 50549
```

8.3 Saving the Survey Design Object

- R Code Chunk 3: Save Survey Design Object

Finally, we save the completed survey design object (`w.design0`) as an `.rds` file. This allows us to load this object directly in the next chapters without needing to re-run the design specification code, making our workflow more efficient.

```
# Save the final survey design object
saveRDS(w.design0, file = "data/w.design0.rds")
```

This completes the survey design specification section of the statistical analysis stage.

8.4 Chapter Summary and Next Steps

We have now completed one of the most critical methodological steps in this analysis. By creating a survey design object (`w.design0`), we have properly accounted for the complex stratification, clustering, and weighting of the NHANES data. This ensures that all subsequent analyses will produce statistically valid, nationally representative results.

With the survey design specified, we are ready to begin our analysis. In the next chapter, "Descriptive Analysis," we will generate the first set of results: weighted summary statistics of our cohort.

9 Descriptive Analysis

This section presents the descriptive statistics of the final analytic cohort, providing an overview of participant characteristics. Cohort characteristics are summarized using weighted proportions and sample counts, accounting for the complex survey design of NHANES. This section aligns with the "Results" section of the paper, particularly referencing content related to Appendix Tables 1 and 2.

This chapter will generate the following key descriptive tables:

Table Description	R Packages Used	Weighting	Purpose
Unweighted Summary	`tableone`	Unweighted	Initial exploration of raw sample counts.
Detailed Unweighted Tables	`table1`	Unweighted	Visually appealing, detailed breakdowns of the cohort.
Appendix Table 1	`survey`	**Weighted**	**Reproduces paper**: Participant characteristics by mortality status.
Appendix Table 2	`survey`	**Weighted**	**Reproduces paper**: Participant characteristics by smoking exposure.

```r
# Load required packages
library(tableone)
library(table1)
library(survey)
options(survey.want.obsolete=TRUE)
require(knitr)
require(kableExtra)
library(expss)
```

- R Code Chunk 1: Load Analytic Data and Survey Design

We begin by loading the `dat.analytic2` and `dat.complete` datasets, and the final survey design object, `w.design0`. These are essential for generating accurate weighted descriptive statistics, as required for nationally representative results.

```
# Load the complete case analytic dataset
load(file = "data/dat.analytic2.RData")
load(file = "data/dat.complete.RData")

# Load the subsetted survey design object
w.design0 <- readRDS(file = "data/w.design0.rds")
```

9.1 Exploratory Analysis: Unweighted Summaries

- R Code Chunk 2: Create Unweighted Descriptive Data Table

The following code creates a basic unweighted descriptive data table, referred to as "Table 1", using the `CreateTableOne()` function from the `tableone` package. This table provides unweighted counts and percentages for our key variables, stratified by mortality status. While the paper's main results are based on weighted proportions, this unweighted table can be useful for initial data exploration and understanding the raw counts of variables within your specific analytic sample.

```
# Unweighted Table 1 summarizing exposure, race,
# and sex, stratified by mortality status
tab1 <- CreateTableOne(vars = c("exposure.cat", "race", "sex"),
                                  strata = "status_all",
                                  data = dat.analytic2,
                                  addOverall = TRUE,
                                  test = TRUE)

# View the summarized version of the table
print(tab1$CatTable)
#>                        Stratified by status_all
#>                         Overall       0             1            p      test
#>   n                     50549         44377         6172
#>   exposure.cat (%)                                                <0.001
#>      Never smoked       28593 (56.6)  26235 (59.1)  2358 (38.2)
```

```
#>      Started before 10   337 ( 0.7)    244 ( 0.5)    93 ( 1.5)
#>      Started at 10-14   3903 ( 7.7)   3145 ( 7.1)   758 (12.3)
#>      Started at 15-17   7189 (14.2)   6003 (13.5)  1186 (19.2)
#>      Started at 18-20   6254 (12.4)   5250 (11.8)  1004 (16.3)
#>      Started after 20   4273 ( 8.5)   3500 ( 7.9)   773 (12.5)
#>   race (%)                                                      <0.001
#>      White             21069 (41.7)  17889 (40.3)  3180 (51.5)
#>      Black             10977 (21.7)   9471 (21.3)  1506 (24.4)
#>      Hispanic          13592 (26.9)  12342 (27.8)  1250 (20.3)
#>      Others             4911 ( 9.7)   4675 (10.5)   236 ( 3.8)
#>   sex = Female (%)     26158 (51.7)  23544 (53.1)  2614 (42.4)  <0.001
```

- R Code Chunk 3: More Detailed Unweighted Tables

The following code uses the `table1` package to generate more visually appealing unweighted descriptive tables. These tables offer flexibility in stratifying variables and are useful for detailed unweighted breakdowns of your cohort's characteristics, though they do not account for survey weights.

1. Create and Display Tables

The following code chunk demonstrates this by creating four different tables, each exploring a different combination of demographic variables, smoking exposure categories, and mortality status. Each table is then rendered as a clean HTML table using `table1::t1kable()`.

```r
# Table of exposure.cat stratified by race and mortality status
tab11 <- table1::table1(~ exposure.cat | race * status_all ,
                        data = dat.analytic2)
# Display as a formatted HTML table in Quarto
table1::t1kable(tab11)
```

	White		Black		Hispanic	
	0	1	0	1	0	1
	(N=17889)	(N=3180)	(N=9471)	(N=1506)	(N=12342)	(N=1250)
exposure.cat						
Never smoked	9153 (51.2%)	1041 (32.7%)	5729 (60.5%)	605 (40.2%)	8054 (65.3%)	600 (48.0%
Started before 10	134 (0.7%)	58 (1.8%)	31 (0.3%)	13 (0.9%)	55 (0.4%)	18 (1.4%)
Started at 10-14	1685 (9.4%)	430 (13.5%)	503 (5.3%)	166 (11.0%)	788 (6.4%)	135 (10.8%
Started at 15-17	3198 (17.9%)	725 (22.8%)	1082 (11.4%)	267 (17.7%)	1348 (10.9%)	163 (13.0%
Started at 18-20	2458 (13.7%)	561 (17.6%)	1116 (11.8%)	240 (15.9%)	1249 (10.1%)	174 (13.9%
Started after 20	1261 (7.0%)	365 (11.5%)	1010 (10.7%)	215 (14.3%)	848 (6.9%)	160 (12.8%

Unweighted: Smoking initiation categories stratified by race and mortality status.

```
# Table of exposure.cat stratified by sex and mortality status
tab12 <- table1::table1(~ exposure.cat| sex * status_all ,
                        data = dat.analytic2)
# Display as a formatted HTML table in Quarto
table1::t1kable(tab12)
```

	Male		Female		Overall	
	0	1	0	1	0	1
	(N=20833)	(N=3558)	(N=23544)	(N=2614)	(N=44377)	(N=61
exposure.cat						
Never smoked	10504 (50.4%)	1030 (28.9%)	15731 (66.8%)	1328 (50.8%)	26235 (59.1%)	2358 (
Started before 10	177 (0.8%)	80 (2.2%)	67 (0.3%)	13 (0.5%)	244 (0.5%)	93 (1.5
Started at 10-14	1923 (9.2%)	572 (16.1%)	1222 (5.2%)	186 (7.1%)	3145 (7.1%)	758 (1
Started at 15-17	3414 (16.4%)	823 (23.1%)	2589 (11.0%)	363 (13.9%)	6003 (13.5%)	1186 (
Started at 18-20	3005 (14.4%)	656 (18.4%)	2245 (9.5%)	348 (13.3%)	5250 (11.8%)	1004 (
Started after 20	1810 (8.7%)	397 (11.2%)	1690 (7.2%)	376 (14.4%)	3500 (7.9%)	773 (1

Unweighted: Smoking initiation categories stratified by sex and mortality status.

```
# Table of demographics and survey year stratified by mortality status
tab13 <- table1::table1(~ exposure.cat + race + sex + year.cat |
                          status_all , data = dat.analytic2)
# Display as a formatted HTML table in Quarto
table1::t1kable(tab13)
```

	0	1	Overall
	(N=44377)	(N=6172)	(N=50549)
exposure.cat			
Never smoked	26235 (59.1%)	2358 (38.2%)	28593 (56.6%)
Started before 10	244 (0.5%)	93 (1.5%)	337 (0.7%)
Started at 10-14	3145 (7.1%)	758 (12.3%)	3903 (7.7%)
Started at 15-17	6003 (13.5%)	1186 (19.2%)	7189 (14.2%)
Started at 18-20	5250 (11.8%)	1004 (16.3%)	6254 (12.4%)
Started after 20	3500 (7.9%)	773 (12.5%)	4273 (8.5%)
race			
White	17889 (40.3%)	3180 (51.5%)	21069 (41.7%)
Black	9471 (21.3%)	1506 (24.4%)	10977 (21.7%)
Hispanic	12342 (27.8%)	1250 (20.3%)	13592 (26.9%)
Others	4675 (10.5%)	236 (3.8%)	4911 (9.7%)
sex			
Male	20833 (46.9%)	3558 (57.6%)	24391 (48.3%)
Female	23544 (53.1%)	2614 (42.4%)	26158 (51.7%)
year.cat			
1999-2000	3188 (7.2%)	1247 (20.2%)	4435 (8.8%)
2001-2002	3772 (8.5%)	1073 (17.4%)	4845 (9.6%)
2003-2004	3583 (8.1%)	921 (14.9%)	4504 (8.9%)
2005-2006	3907 (8.8%)	667 (10.8%)	4574 (9.0%)
2007-2008	4705 (10.6%)	769 (12.5%)	5474 (10.8%)
2009-2010	5204 (11.7%)	563 (9.1%)	5767 (11.4%)
2011-2012	4770 (10.7%)	392 (6.4%)	5162 (10.2%)
2013-2014	5111 (11.5%)	285 (4.6%)	5396 (10.7%)
2015-2016	5132 (11.6%)	170 (2.8%)	5302 (10.5%)
2017-2018	5005 (11.3%)	85 (1.4%)	5090 (10.1%)

Unweighted: Full cohort characteristics stratified by mortality status.

```
# Table of race and sex stratified by exposure.cat
tab14 <- table1::table1(~ race + sex | exposure.cat ,
                        data = dat.analytic2)
# Display as a formatted HTML table in Quarto
table1::t1kable(tab14)
```

	Never smoked	Started before 10	Started at 10-14	Started at 15-17	Started at 18-20	Sta
	(N=28593)	(N=337)	(N=3903)	(N=7189)	(N=6254)	(N=
race						
White	10194 (35.7%)	192 (57.0%)	2115 (54.2%)	3923 (54.6%)	3019 (48.3%)	162
Black	6334 (22.2%)	44 (13.1%)	669 (17.1%)	1349 (18.8%)	1356 (21.7%)	122
Hispanic	8654 (30.3%)	73 (21.7%)	923 (23.6%)	1511 (21.0%)	1423 (22.8%)	100
Others	3411 (11.9%)	28 (8.3%)	196 (5.0%)	406 (5.6%)	456 (7.3%)	414
sex						
Male	11534 (40.3%)	257 (76.3%)	2495 (63.9%)	4237 (58.9%)	3661 (58.5%)	220
Female	17059 (59.7%)	80 (23.7%)	1408 (36.1%)	2952 (41.1%)	2593 (41.5%)	206

Unweighted: Demographic characteristics stratified by smoking initiation category.

2. Save Tables to Excel (Optional)

The tables created with the `table1` package can also be saved to an external file, such as an Excel spreadsheet, for further review outside of this book. The following code demonstrates how to save two of these tables using the `expss` package.

```
# Table 3
t13 <- as.data.frame(tab13)
expss::xl_write_file(t13, filename = "data/t13.xlsx")

# Table 4
t14 <- as.data.frame(tab14)
expss::xl_write_file(t14, filename = "data/t14.xlsx")
```

9.2 Weighted Descriptive Statistics (Paper Reproduction)

- R Code Chunk 4: Weighted Descriptive Tables

The following code generates the primary descriptive tables for the analysis. Unlike the previous examples, these tables correctly account for the complex survey design by using the `svyCreateTableOne()` function on our survey design object (`w.design0`). This produces nationally representative, weighted percentages.

9.2.1 Appendix Table 1: Characteristics by Mortality Status

This first table summarizes participant characteristics (smoking exposure, race, sex, and survey year), stratified by their mortality status. The output is designed to directly reproduce the results shown in Appendix Table 1 of the supplementary material where percentages were in brackets. Those percentages were calculated by accounting for sampling (interview) weights.

```
# Create weighted Table 1
# stratified by outcome status (all-cause mortality)
tab13_weighted <- svyCreateTableOne(vars = c("exposure.cat", "race",
                                             "sex", "year.cat"),
                                    strata = "status_all",
                                    data = w.design0, # CRITICAL
                                    addOverall = TRUE,
                                    test = TRUE)

# Print the table with weighted proportions,
# specified decimal places, and all factor levels
tab13p_weighted <- print(tab13_weighted,
                         format = "p",
                         catDigits = 2,
                         showAllLevels = TRUE,
                         smd = TRUE)
# Re-label
colnames(tab13p_weighted)[colnames(tab13p_weighted)
                          == "0"] <- "Alive"
colnames(tab13p_weighted)[colnames(tab13p_weighted)
                          == "1"] <- "Dead"
# Order
new_order <- c("level", "Alive", "Dead",
               "Overall", "p", "test", "SMD")
# Apply the new order to the table object
tab13p_weighted <- tab13p_weighted[, new_order]

# Save the weighted table to CSV
write.csv(tab13p_weighted, file = "data/Table_App_1_Weighted_Mortality.csv")
```

```
# Display the formatted table using kable for a clean Quarto output
kable(tab13p_weighted, caption = "Weighted Characteristics by
      All-Cause Mortality Status (Analogous to Appendix
      Table 1)") %>%
```

```
kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

Table 9.2: Weighted Characteristics by All-Cause Mortality Status (Analogous to Appendix Table 1)

	level	Alive	Dead	Overall	p	test	SMD
n		187497900.24	18761712.99	206259613.23			
exposure.cat (%)	Never smoked	57.99	36.57	56.04	<0.001		0.450
	Started before 10	0.49	1.48	0.58			
	Started at 10-14	7.17	12.54	7.66			
	Started at 15-17	15.02	21.01	15.57			
	Started at 18-20	12.24	16.78	12.65			
	Started after 20	7.09	11.61	7.50			
race (%)	White	66.42	74.81	67.18	<0.001		0.266
	Black	11.35	12.90	11.50			
	Hispanic	14.81	7.87	14.18			
	Others	7.42	4.42	7.15			
sex (%)	Male	47.88	55.16	48.54	<0.001		0.146
	Female	52.12	44.84	51.46			
year.cat (%)	1999-2000	7.83	19.75	8.92	<0.001		0.793
	2001-2002	8.42	17.12	9.21			
	2003-2004	8.91	15.36	9.49			
	2005-2006	9.46	12.42	9.73			
	2007-2008	9.84	10.81	9.93			
	2009-2010	10.28	8.31	10.10			
	2011-2012	10.65	6.73	10.29			
	2013-2014	11.11	5.36	10.59			
	2015-2016	11.57	2.79	10.78			
	2017-2018	11.92	1.34	10.96			

9.2.2 Appendix Table 2: Characteristics by Smoking Initiation Exposure

This second table reproduces the results from Appendix Table 2 of the supplementary material. It summarizes the demographic characteristics (`race` and `sex`) of the cohort, stratified by the `exposure.cat` variable (smoking initiation categories). This provides the nationally representative, weighted demographic composition within each smoking exposure group as it accounts for the complex survey design.

```r
# Create weighted Table 2 stratified by smoking initiation categories
tab14_weighted <- svyCreateTableOne(vars = c("race", "sex"),
                                    strata = "exposure.cat",
                                    data = w.design0, # CRITICAL
                                    addOverall = TRUE,
                                    test = TRUE)

# Print the table with weighted proportions
# and specified decimal places
tab14p_weighted <- print(tab14_weighted,
                         format = "p",
                         catDigits = 2,
                         showAllLevels = TRUE,
                         smd = TRUE)

# Define the desired column order
new_order_t2 <- c("level", "Never smoked", "Started before 10",
                  "Started at 10-14", "Started at 15-17",
                  "Started at 18-20", "Started after 20",
                  "Overall", "p", "test", "SMD")
# Apply the new order to the table object
tab14p_weighted <- tab14p_weighted[, new_order_t2]

# Save the weighted table to CSV
write.csv(tab14p_weighted, file = "data/Table_App_2_Weighted_Exposure.csv")
```

```r
# Display the formatted table using kable for a clean Quarto output
kable(tab14p_weighted, caption = "Weighted Characteristics by
      Smoking Initiation Categories (Analogous to Appendix
      Table 2)") %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

Table 9.3: Weighted Characteristics by Smoking Initiation Categories (A
Table 2)

	level	Never smoked	Started before 10	Started at 10-14	Started at 15-17	Started at 18-
n		115595521.42	1188555.71	15799589.69	32111621.33	26098412.76
race (%)	White	62.54	74.05	75.24	76.89	72.65
	Black	12.53	6.92	8.07	8.47	10.39
	Hispanic	16.50	10.22	12.37	10.31	11.09
	Others	8.42	8.81	4.32	4.34	5.87

sex (%)	Male	42.98	73.85	58.70	56.37	55.14
	Female	57.02	26.15	41.30	43.63	44.86

This completes the descriptive analysis section of the statistical analysis stage.

9.3 Chapter Summary and Next Steps

In this chapter, we generated the primary descriptive statistics for the study cohort. By using the survey design object, we successfully reproduced the weighted characteristics of the participants, creating tables analogous to Appendix Tables 1 and 2 from the paper's supplementary material. This gives us a clear, nationally representative picture of our study population.

Now that we understand the characteristics of our cohort, we will move on to the core analysis in the next chapter, "Survival Analysis," where we will investigate the primary relationship between smoking initiation and mortality.

10 Survival Analysis

This section performs the main survival analyses to investigate the association between early smoking initiation and all-cause mortality. We will first visualize survival probabilities using Kaplan-Meier curves and then estimate hazard ratios (HRs) with survey-weighted Cox Proportional Hazards models. This section directly reproduces the "Statistical Analysis" and "Results" sections of the published paper. Specifically, Figure 1 (Kaplan-Meier curve) and Figure 2 (the crude/adjusted HRs) of the original paper.

This chapter will reproduce the paper's core survival analyses using the following methods:

Analysis/Visualization	Key R Function(s)	Purpose	Reproduces
Kaplan-Meier Curve	`svykm()`, `svylogrank()`	Visualize and test differences in survival probabilities between exposure groups.	**Figure 1**
Crude Cox Model	`svycoxph()`	Estimate the unadjusted association between smoking initiation and mortality.	**Figure 2** (Crude HRs)
Adjusted Cox Model	`svycoxph()`	Estimate the association while controlling for confounders (sex, race, survey year).	**Figure 2** (Adjusted HRs)

```r
# Load required packages
library(Publish)
library(survey)
options(survey.want.obsolete=TRUE)
library(survival)
require(knitr)
require(kableExtra)
```

```
require(stringr)
```

- R Code Chunk 1: Load Data and Survey Design Object

We begin by loading the final, complete-case analytic dataset (`dat.complete`) and the survey design object (`w.design0`) that were created in the previous chapters. The `w.design0` object is essential for ensuring our analysis accounts for the complex NHANES survey design.

```
# Load the complete case analytic dataset
load(file = "data/dat.complete.RData")

# Load the subsetted survey design object
w.design0 <- readRDS(file = "data/w.design0.rds")

# Verify the dimensions of the loaded survey design object
dim(w.design0)
#> [1] 50549    19
```

10.1 Kaplan-Meier Survival Analysis (Figure 1)

- R Code Chunk 2: Kaplan-Meier Curves and Log-Rank Test

Kaplan-Meier Curves

The following code visualizes the survival probabilities over time using Kaplan-Meier (KM) curves, stratified by the `exposure.cat` (age at smoking initiation) variable. The KM curves illustrate the proportion of participants surviving over their follow-up time from birth. The `svykm()` function from the `survey` package is used to calculate the survival estimates while accounting for the complex survey design. The resulting plot reproduces the trends shown in Figure 1 of the original paper.

```
# Define the survival formula
formulax0 <- as.formula(Surv(stime.since.birth, status_all)
                        ~ exposure.cat)

# Calculate survey-weighted Kaplan-Meier curves
```

```r
sA <- svykm(formulax0, design = w.design0)
saveRDS(sA, file = "data/sA.rds")

# Dynamically set the number of colors for the legend
dummy <- length(unique(as.factor(w.design0$variables$exposure.cat)))

# Plot the Kaplan-Meier curves
plot(sA, pars = list(col = c(1:dummy)),
     xlab = "Time",
     ylab = "Proportion surviving",
     main = "Survey-featured Kaplan-Meier Curve:
     All-Cause Mortality")
legend("bottomleft",
       levels(as.factor(w.design0$variables$exposure.cat)),
       col = (1:dummy), lty = c(1,1),
       bty = "n")
```

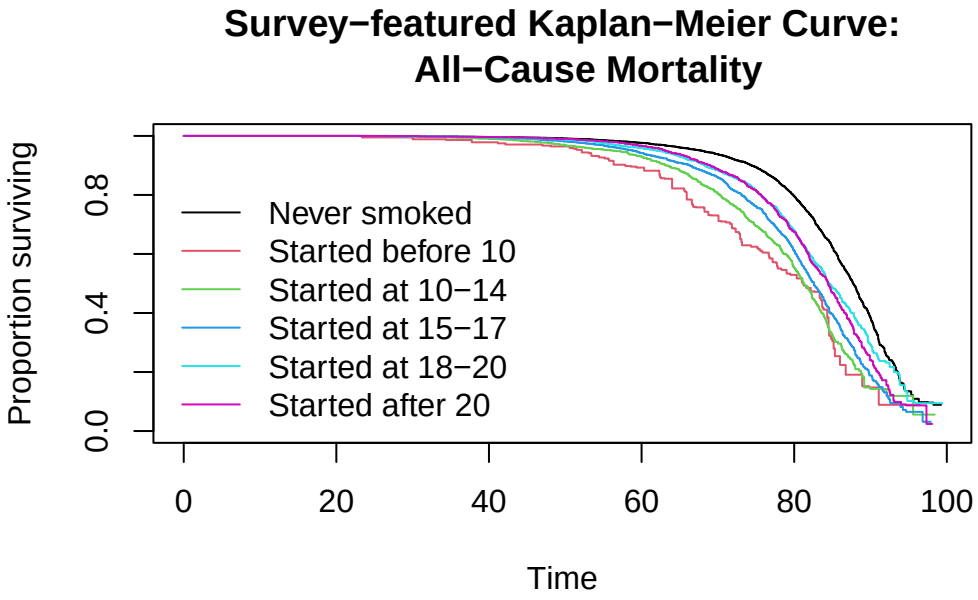


Figure 10.1: Survey-weighted Kaplan-Meier curves for all-cause mortality, stratified by age of smoking initiation (reproduces Figure 1 from the paper).

Log-Rank Test

After plotting the curves, a survey-weighted log-rank test is performed using `svylogrank()` to statistically determine if there are significant differences in survival curves among the exposure groups. A significant p-value from this test would indicate a statistically significant difference in survival experiences between the age groups of smoking initiation. The paper states that the log-rank test was significant.

```
# Perform the survey-weighted log-rank test
lrt <- svylogrank(formulax0, design = subset(w.design0,
                                             survey.weight.new > 0))

# Display the rounded p-value from the log-rank test
round(lrt[[2]],2)
#>  Chisq      p
#> 293.72   0.00
```

To present these results more formally, the following code uses `knitr::kable()` to display the full test output, including the Chi-squared statistic and degrees of freedom, in a clean table.

```
kable(lrt[[1]],
      booktabs = TRUE, digits = 2,
      col.names = c("Test Statistic", "Standard Error", "Z-value", "p-value"),
      caption="Survey-weighted log-rank test for main analysis comparing survival distribu
  kable_styling(latex_options = "hold_position")
```

Table 10.2: Survey-weighted log-rank test for main analysis comparing survival distributions by age of smoking initiation.

Test Statistic	Standard Error	Z-value	p-value
146771.4	33970.03	4.32	0.00
1057040.8	110866.82	9.53	0.00
1215922.9	137852.17	8.82	0.00
259155.7	128215.04	2.02	0.04
235676.6	91478.11	2.58	0.01

10.2 Cox Proportional Hazards Models (Figure 2)

10.2.1 Unadjusted Model (Crude HRs)

- R Code Chunk 3: Unadjusted Cox Model (for Comparison)

The following code demonstrates a standard, un-weighted Cox proportional hazards model using base R's `coxph()` function on the `dat.complete` dataset. This model does not account for the complex survey design of NHANES and is included as a baseline, primarily for comparison or to understand how unweighted models would be run. The `publish()` function is used to format the model output for easier interpretation.

```
# Fit an un-weighted Cox Proportional Hazards model
fit0 <- coxph(Surv(stime.since.birth, status_all) ~ exposure.cat,
              data = dat.complete)

# Use 'publish' to format the output for display
publish(fit0)
#>      Variable             Units HazardRatio       CI.95 p-value
#>  exposure.cat      Never smoked         Ref
#>                Started before 10        2.44 [1.99;3.00]  <0.001
#>                 Started at 10-14        2.37 [2.19;2.58]  <0.001
#>                 Started at 15-17        1.92 [1.79;2.06]  <0.001
#>                 Started at 18-20        1.51 [1.40;1.62]  <0.001
#>                 Started after 20        1.51 [1.39;1.64]  <0.001
```

- R Code Chunk 4: Unadjusted Survey-Weighted Cox Proportional Hazards Model

The following code fits the unadjusted Cox proportional hazards model, but it uses `svycoxph()` from the `survey` package to account for the complex NHANES survey design. This model examines the crude association between smoking initiation categories and all-cause mortality, producing Hazard Ratios. The `publish()` function is used to format the model output for easier interpretation. The HRs from this model contribute to the "Crude" estimates shown in Figure 2 of the paper.

```
# Fit the unadjusted survey-weighted Cox Proportional Hazards model
fit0 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat,
                 design = w.design0)

# Use 'publish' to format the output for display
f0 <- publish(fit0)
#> Stratified 1 - level Cluster Sampling design (with replacement)
#> With (301) clusters.
#> subset(w.design, miss == 0 & survey.weight.new > 0)
#>      Variable             Units HazardRatio       CI.95 p-value
#>  exposure.cat      Never smoked         Ref
```

```
#>                Started before 10        3.01 [2.29;3.95]  <0.001
#>                 Started at 10-14        2.60 [2.32;2.92]  <0.001
#>                 Started at 15-17        2.07 [1.88;2.28]  <0.001
#>                 Started at 18-20        1.55 [1.40;1.72]  <0.001
#>                 Started after 20        1.60 [1.45;1.76]  <0.001
```

- R Code Chunk 5: Process Unadjusted Hazard Ratios for Plotting

The output from `publish()` is a formatted table, but for plotting, we need a clean data frame of the results.

The following code extracts the Hazard Ratios and their 95% Confidence Intervals from the unadjusted survey-weighted Cox Model (`f0`). The `stringr` package is used to parse the confidence interval strings into separate lower and upper bound numeric columns.

```r
# Select HR and CI columns for exposure categories from 'f0'
f0r <- f0$regressionTable[2:6,c("HazardRatio","CI.95")]

# Add a 'group' column to label these results as "Crude" for plotting
f0r$group <- "Crude"

# Extract lower and upper bounds from the CI string
ci <- str_extract_all(f0r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f0r$CI.l <- as.numeric(as.character(ci[,1])) # Lower bound
f0r$CI.u <- as.numeric(as.character(ci[,2])) # Upper bound

# Rename columns for consistency in plotting
names(f0r) <- c("mean","CI.95","group","lower","upper")

# Display the processed data frame
f0r
#>   mean       CI.95 group lower upper
#> 2 3.01 [2.29;3.95] Crude  2.29  3.95
#> 3 2.60 [2.32;2.92] Crude  2.32  2.92
#> 4 2.07 [1.88;2.28] Crude  1.88  2.28
#> 5 1.55 [1.40;1.72] Crude  1.40  1.72
#> 6 1.60 [1.45;1.76] Crude  1.45  1.76
```

10.2.2 Adjusted Model (Adjusted HRs)

- R Code Chunk 6: Adjusted Survey-Weighted Cox Proportional Hazards Model

The following code fits the adjusted Cox proportional hazards model, accounting for the complex survey design and controlling for **key covariates**. As described in the paper, the model is adjusted to account for **sex, race/ethnicity, and survey cycle**. This model estimates the adjusted Hazard Ratios, which contribute to the "Adjusted" estimates shown in Figure 2 of the paper.

```
# Fit the adjusted survey-weighted Cox Proportional Hazards model
fit1 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat +
                   sex + race + year.cat, design = w.design0)

# Use 'publish' to format the output for display
f1 <- publish(fit1) # Store the published output
#> Stratified 1 - level Cluster Sampling design (with replacement)
#> With (301) clusters.
#> subset(w.design, miss == 0 & survey.weight.new > 0)
#>      Variable             Units HazardRatio       CI.95   p-value
#>  exposure.cat      Never smoked         Ref
#>                Started before 10        2.71 [2.05;3.58]   < 0.001
#>                 Started at 10-14        2.38 [2.11;2.68]   < 0.001
#>                 Started at 15-17        1.96 [1.77;2.16]   < 0.001
#>                 Started at 18-20        1.48 [1.33;1.64]   < 0.001
#>                 Started after 20        1.54 [1.39;1.70]   < 0.001
#>           sex              Male         Ref
#>                          Female        0.75 [0.70;0.80]   < 0.001
#>          race             White         Ref
#>                           Black        1.59 [1.47;1.73]   < 0.001
#>                        Hispanic        1.03 [0.93;1.15]   0.55581
#>                          Others        1.15 [0.97;1.37]   0.10079
#>      year.cat         1999-2000         Ref
#>                       2001-2002        0.96 [0.86;1.08]   0.50457
#>                       2003-2004        0.84 [0.75;0.94]   0.00328
#>                       2005-2006        0.76 [0.67;0.85]   < 0.001
#>                       2007-2008        0.77 [0.67;0.90]   < 0.001
#>                       2009-2010        0.68 [0.59;0.78]   < 0.001
#>                       2011-2012        0.62 [0.51;0.76]   < 0.001
#>                       2013-2014        0.58 [0.48;0.69]   < 0.001
#>                       2015-2016        0.35 [0.27;0.45]   < 0.001
#>                       2017-2018        0.20 [0.13;0.29]   < 0.001
```

- R Code Chunk 7: Process Adjusted Hazard Ratios for Plotting

The following code follows the similar processing done for the unadjusted model, but applies it to the results of the adjusted survey-weighted Cox model.

```
# Select Hazard Ratio and CI columns for exposure categories from 'f1'
f1r <- f1$regressionTable[2:6,c("HazardRatio","CI.95")]

# Add a 'group' column to label these results as "Adjusted"
f1r$group <- "Adjusted"

# Extract lower and upper bounds from the CI string
ci <- str_extract_all(f1r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f1r$CI.l <- as.numeric(as.character(ci[,1]))
f1r$CI.u <- as.numeric(as.character(ci[,2]))

# Rename columns for consistency in plotting
names(f1r) <- c("mean","CI.95","group","lower","upper")

# Display the processed data frame
f1r
#>   mean       CI.95    group lower upper
#> 2 2.71 [2.05;3.58] Adjusted  2.05  3.58
#> 3 2.38 [2.11;2.68] Adjusted  2.11  2.68
#> 4 1.96 [1.77;2.16] Adjusted  1.77  2.16
#> 5 1.48 [1.33;1.64] Adjusted  1.33  1.64
#> 6 1.54 [1.39;1.70] Adjusted  1.39  1.70
```

10.3 Saving Model Results

- R Code Chunk 8: Save Processed Results

Finally, we save the two processed data frames (`f0r` and `f1r`) containing the crude and adjusted results. These files will be loaded in the next section to create the plot that visualizes these findings (Figure 2 of paper).

```r
# Save the processed HR data frames for the next chapter
saveRDS(f0r, file = "data/f0r.rds")
saveRDS(f1r, file = "data/f1r.rds")
```

This completes the survival analysis section of the statistical analysis stage.

10.4 Chapter Summary and Next Steps

We have now completed the main survival analysis. We successfully reproduced the Kaplan-Meier curves (Figure 1 from the paper) to visualize survival probabilities and fitted both crude and adjusted survey-weighted Cox models to estimate the hazard ratios, replicating the core findings presented in Figure 2.

The results show a clear association between early smoking initiation and mortality. The next logical step, covered in the "Effect Modification Analysis" chapter, is to investigate whether this association differs across key demographic groups.

11 Effect Modification Analysis

This section explains the effect modification analysis, which investigates whether the association between early smoking initiation and all-cause mortality is changed by key demographic variables like race/ethnicity and sex. Effect modification occurs when the strength or direction of an association differs across levels of another variable.

This is investigated through 2 ways.

1. Interaction Terms : Including interaction terms in our survey-weighted Cox proportional hazards models and performing statistical tests for these interaction terms.
2. RERI (Relative Excess Risk due to Interaction) : Calculating this measure to assess additive interaction, which provides a different perspective on how factors combine their effects.

The key methods are summarized below:

Analysis	Key R Function(s)	Purpose	Reproduces
Effect Modification by Race/Ethnicity	`svycoxph()`	Estimate stratum-specific HRs for each racial/ethnic group.	**Figure 2** (Race-specific HRs)
Effect Modification by Sex	`svycoxph()`	Estimate stratum-specific HRs for males and females.	**Figure 2** (Sex-specific HRs)
Interaction Tests	`svyglm()`, `regTermTest()`	Formally test if the effect of smoking initiation differs across groups.	**Paper's p-values** for interaction
Forest Plot	`ggplot()`	Visualize and compare all crude, adjusted, and stratum-specific HRs.	**Figure 2** (Complete Plot)

Analysis	Key R Function(s)	Purpose	Reproduces
Additive Interaction (RERI)	`reri.f()` (custom)	Quantify the excess risk due to the combined presence of two factors.	**Appendix Tables 4 & 5**

```r
# Load required packages
library(Publish)
library(survey)
options(survey.want.obsolete=TRUE)
library(survival)
require(knitr)
require(kableExtra)
library(ggplot2)
library(stringr)
library(expss)
```

- R Code Chunk 1: Load Data and Processed Results

The following code loads the essential `w.design0` survey design object, which was created and saved in the previous section. This object is important for all subsequent survey-weighted analyses coming up. Additionally, it ensures that `f0r` and `f1r` (the processed Hazard Ratios from previous section) are available for binding and plotting.

```r
# Load the subsetted survey design object
w.design0 <- readRDS(file = "data/w.design0.rds")

# Load f0r and f1r from previous section
f0r <- readRDS("data/f0r.rds")
f1r <- readRDS("data/f1r.rds")
```

11.1 Investigating Effect Modification by Race/Ethnicity

- R Code Chunk 2: Effect Modification by Race/Ethnicity

The following code fits a survey-weighted Cox proportional hazards model, including an interaction term between `exposure.cat` (smoking initiation age) and `race`. This model is used to assess if the association between early smoking initiation and mortality differs across various racial/ethnic levels or groups.

```
# Effect modification by race
fit2 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat*race
                 + sex + year.cat, design = w.design0)

# Publish the results for later processing
f2 <- publish(fit2)
#> Stratified 1 - level Cluster Sampling design (with replacement)
#> With (301) clusters.
#> subset(w.design, miss == 0 & survey.weight.new > 0)
#>                                                          Variable     Units HazardRatio
#>                                                               sex      Male         Ref
#>                                                                      Female        0.74
#>                                                          year.cat 1999-2000         Ref
#>                                                                   2001-2002        0.96
#>                                                                   2003-2004        0.84
#>                                                                   2005-2006        0.75
#>                                                                   2007-2008        0.77
#>                                                                   2009-2010        0.68
#>                                                                   2011-2012        0.62
#>                                                                   2013-2014        0.58
#>                                                                   2015-2016        0.34
#>                                                                   2017-2018        0.19
#>                 exposure.cat(Never smoked): race(Black vs White)                    1.82
#>               exposure.cat(Never smoked): race(Hispanic vs White)                  1.24
#>                 exposure.cat(Never smoked): race(Others vs White)                  1.21
#>             exposure.cat(Started before 10): race(Black vs White)                  1.11
#>           exposure.cat(Started before 10): race(Hispanic vs White)                 0.40
#>             exposure.cat(Started before 10): race(Others vs White)                 1.74
#>               exposure.cat(Started at 10-14): race(Black vs White)                 1.44
#>            exposure.cat(Started at 10-14): race(Hispanic vs White)                 0.78
#>              exposure.cat(Started at 10-14): race(Others vs White)                 1.44
#>               exposure.cat(Started at 15-17): race(Black vs White)                 1.43
#>            exposure.cat(Started at 15-17): race(Hispanic vs White)                 0.89
#>              exposure.cat(Started at 15-17): race(Others vs White)                 1.57
#>               exposure.cat(Started at 18-20): race(Black vs White)                 1.72
#>            exposure.cat(Started at 18-20): race(Hispanic vs White)                 1.05
#>              exposure.cat(Started at 18-20): race(Others vs White)                 0.96
```

```
#>             exposure.cat(Started after 20): race(Black vs White)                1.32
#>           exposure.cat(Started after 20): race(Hispanic vs White)               0.98
#>             exposure.cat(Started after 20): race(Others vs White)               0.77
#>     race(White): exposure.cat(Started before 10 vs Never smoked)                3.13
#>      race(White): exposure.cat(Started at 10-14 vs Never smoked)                2.53
#>      race(White): exposure.cat(Started at 15-17 vs Never smoked)                2.06
#>      race(White): exposure.cat(Started at 18-20 vs Never smoked)                1.53
#>      race(White): exposure.cat(Started after 20 vs Never smoked)                1.70
#>     race(Black): exposure.cat(Started before 10 vs Never smoked)                1.91
#>      race(Black): exposure.cat(Started at 10-14 vs Never smoked)                2.00
#>      race(Black): exposure.cat(Started at 15-17 vs Never smoked)                1.62
#>      race(Black): exposure.cat(Started at 18-20 vs Never smoked)                1.45
#>      race(Black): exposure.cat(Started after 20 vs Never smoked)                1.22
#>  race(Hispanic): exposure.cat(Started before 10 vs Never smoked)                1.00
#>   race(Hispanic): exposure.cat(Started at 10-14 vs Never smoked)                1.58
#>   race(Hispanic): exposure.cat(Started at 15-17 vs Never smoked)                1.47
#>   race(Hispanic): exposure.cat(Started at 18-20 vs Never smoked)                1.29
#>   race(Hispanic): exposure.cat(Started after 20 vs Never smoked)                1.33
#>    race(Others): exposure.cat(Started before 10 vs Never smoked)                4.51
#>     race(Others): exposure.cat(Started at 10-14 vs Never smoked)                3.01
#>     race(Others): exposure.cat(Started at 15-17 vs Never smoked)                2.67
#>     race(Others): exposure.cat(Started at 18-20 vs Never smoked)                1.22
#>     race(Others): exposure.cat(Started after 20 vs Never smoked)                1.08
```

- R Code Chunk 3: Testing Race/Ethnicity Interaction (Poisson Approximation)

The following code performs statistical tests to formally assess the significance of the interaction term between `exposure.cat` and `race`. It uses `svyglm` with a Poisson family (as an approximation to the Cox model) to allow for `anova()` comparisons, and `regTermTest()` is used to perform a Likelihood Ratio Test (LRT) directly on the interaction term. This helps determine if the observed variations in association by race are statistically significant.

```r
# Fit the baseline Poisson model (no interaction)
fit1a <- svyglm(status_all ~ offset(log(stime.since.birth)) +
                  exposure.cat + race + sex  + year.cat,
                design = w.design0, family=poisson)

# Fit the full Poisson model (with interaction)
fit2a <- svyglm(status_all ~ offset(log(stime.since.birth)) +
                  exposure.cat*race + sex  + year.cat,
```

```
                design = w.design0, family=poisson)

# Compare the two models using a Rao-Scott test (via anova)
anova(fit2a,fit1a)
#> Working (Rao-Scott+F) LRT for exposure.cat:race
#>  in svyglm(formula = status_all ~ offset(log(stime.since.birth)) +
#>     exposure.cat * race + sex + year.cat, design = w.design0,
#>     family = poisson)
#> Working 2logLR =  21.45753 p= 0.16075
#> (scale factors:  2 1.9 1.8 1.5 1.4 0.98 0.93 0.76 0.73 0.62 0.57 0.54 0.45 0.37 0.36 );

# Directly test the interaction term using a Likelihood Ratio Test
regTermTest(fit2a, test.terms = "exposure.cat:race", method = "LRT")
#> Working (Rao-Scott+F) LRT for exposure.cat:race
#>  in svyglm(formula = status_all ~ offset(log(stime.since.birth)) +
#>     exposure.cat * race + sex + year.cat, design = w.design0,
#>     family = poisson)
#> Working 2logLR =  21.45753 p= 0.16075
#> (scale factors:  2 1.9 1.8 1.5 1.4 0.98 0.93 0.76 0.73 0.62 0.57 0.54 0.45 0.37 0.36 );
```

- R Code Chunk 4: Process Race-Specific Hazard Ratios for Plotting

Although the overall interaction term was not statistically significant, the original paper still presented the stratum-specific Hazard Ratios (HRs) to visualize the trends within each racial/ethnic group. So the following code extracts stratum-specific Hazard Ratios (HRs) and their 95% Confidence Intervals (CIs) for each racial/ethnic group from the `f2` object (the published output of fit2). These HRs reflect the effect of smoking initiation within each racial/ethnic stratum, allowing for a detailed examination of potential effect modification. The extracted data is formatted into a dataframe, `f2r`.

```
# Extract HRs and CIs for each race subgroup from the published output
f2rW <- f2$regressionTable[31:35,c("HazardRatio","CI.95")]
f2rB <- f2$regressionTable[36:40,c("HazardRatio","CI.95")]
f2rH <- f2$regressionTable[41:45,c("HazardRatio","CI.95")]

# Assign group labels and combine into a single data frame
f2rW$group <- "White"
f2rB$group <- "Black"
f2rH$group <- "Hispanic"
f2r <- rbind(f2rW, f2rB, f2rH)
```

```r
# Process CIs: Extract lower and upper bounds from the CI string
ci <- str_extract_all(f2r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f2r$CI.l <- as.numeric(as.character(ci[,1]))
f2r$CI.u <- as.numeric(as.character(ci[,2]))
names(f2r) <- c("mean","CI.95","group","lower","upper")

# Display the final processed data frame
f2r
#>    mean       CI.95    group lower upper
#> 31 3.13 [2.27;4.31]    White  2.27  4.31
#> 32 2.53 [2.16;2.97]    White  2.16  2.97
#> 33 2.06 [1.84;2.31]    White  1.84  2.31
#> 34 1.53 [1.35;1.74]    White  1.35  1.74
#> 35 1.70 [1.50;1.92]    White  1.50  1.92
#> 36 1.91 [0.99;3.69]    Black  0.99  3.69
#> 37 2.00 [1.65;2.42]    Black  1.65  2.42
#> 38 1.62 [1.37;1.92]    Black  1.37  1.92
#> 39 1.45 [1.20;1.75]    Black  1.20  1.75
#> 40 1.22 [1.03;1.46]    Black  1.03  1.46
#> 41 1.00 [0.47;2.15] Hispanic  0.47  2.15
#> 42 1.58 [1.22;2.05] Hispanic  1.22  2.05
#> 43 1.47 [1.08;2.00] Hispanic  1.08  2.00
#> 44 1.29 [1.00;1.67] Hispanic  1.00  1.67
#> 45 1.33 [1.00;1.78] Hispanic  1.00  1.78
```

11.2 Investigating Effect Modification by Sex

- R Code Chunk 5: Effect Modification by Sex

The following code fits another survey-weighted Cox proportional hazards model, this time including an interaction term between exposure.cat and sex. This model is used to investigate whether the association between early smoking initiation and mortality varies between male and female participants.

```r
# Effect modification by sex
fit3 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat*sex
                 + race + year.cat, design = w.design0)
```

```
# Publish the results for later processing
f3 <- publish(fit3)
#> Stratified 1 - level Cluster Sampling design (with replacement)
#> With (301) clusters.
#> subset(w.design, miss == 0 & survey.weight.new > 0)
#>                                                      Variable     Units HazardRatio
#>                                                          race     White         Ref
#>                                                                   Black        1.60 [1.
#>                                                                Hispanic        1.04 [0.
#>                                                                  Others        1.16 [0.
#>                                                     year.cat 1999-2000         Ref
#>                                                              2001-2002        0.96 [0.
#>                                                              2003-2004        0.84 [0.
#>                                                              2005-2006        0.75 [0.
#>                                                              2007-2008        0.77 [0.
#>                                                              2009-2010        0.68 [0.
#>                                                              2011-2012        0.62 [0.
#>                                                              2013-2014        0.58 [0.
#>                                                              2015-2016        0.35 [0.
#>                                                              2017-2018        0.20 [0.
#>              exposure.cat(Never smoked): sex(Female vs Male)                   0.67 [0.
#>         exposure.cat(Started before 10): sex(Female vs Male)                   0.77 [0.
#>          exposure.cat(Started at 10-14): sex(Female vs Male)                   1.03 [0.
#>          exposure.cat(Started at 15-17): sex(Female vs Male)                   0.76 [0.
#>          exposure.cat(Started at 18-20): sex(Female vs Male)                   0.73 [0.
#>          exposure.cat(Started after 20): sex(Female vs Male)                   0.78 [0.
#>    sex(Male): exposure.cat(Started before 10 vs Never smoked)                  2.54 [1.
#>     sex(Male): exposure.cat(Started at 10-14 vs Never smoked)                  2.05 [1.
#>     sex(Male): exposure.cat(Started at 15-17 vs Never smoked)                  1.83 [1.
#>     sex(Male): exposure.cat(Started at 18-20 vs Never smoked)                  1.40 [1.
#>     sex(Male): exposure.cat(Started after 20 vs Never smoked)                  1.42 [1.
#>  sex(Female): exposure.cat(Started before 10 vs Never smoked)                  2.90 [1.
#>   sex(Female): exposure.cat(Started at 10-14 vs Never smoked)                  3.15 [2.
#>   sex(Female): exposure.cat(Started at 15-17 vs Never smoked)                  2.07 [1.
#>   sex(Female): exposure.cat(Started at 18-20 vs Never smoked)                  1.53 [1.
#>   sex(Female): exposure.cat(Started after 20 vs Never smoked)                  1.63 [1.
```

- R Code Chunk 6: Testing Sex Interaction (Poisson Approximation and AIC)

Similar to the race interaction, we now formally test the significance of the interaction between

smoking initiation and sex. We use the same Poisson approximation approach to compare the model with the interaction term (`fit3a`) against the baseline model without it (`fit1a`).

> **Note**
>
> The Akaike Information Criterion (AIC) is calculated for this model and compared with previous models to evaluate overall model fit, as discussed in the paper. A lower AIC value indicates a better model fit. The paper reports that the sex interaction was statistically significant (p=0.001) and that this model provided a superior fit (lower AIC) compared to the others.

```
# Fit the full Poisson model (with sex interaction)
fit3a <- svyglm(status_all ~ offset(log(stime.since.birth)) +
                  exposure.cat*sex + race + year.cat,
                design = w.design0, family=poisson)

# Compare the sex-interaction model to the baseline model (fit1a)
anova(fit3a,fit1a)
#> Working (Rao-Scott+F) LRT for exposure.cat:sex
#>  in svyglm(formula = status_all ~ offset(log(stime.since.birth)) +
#>     exposure.cat * sex + race + year.cat, design = w.design0,
#>     family = poisson)
#> Working 2logLR =  21.44332 p= 0.0014171
#> (scale factors:  1.2 1.1 1 0.87 0.76 );  denominator df= 130

# Directly test the interaction term
regTermTest(fit3a, test.terms = "exposure.cat:sex", method = "LRT")
#> Working (Rao-Scott+F) LRT for exposure.cat:sex
#>  in svyglm(formula = status_all ~ offset(log(stime.since.birth)) +
#>     exposure.cat * sex + race + year.cat, design = w.design0,
#>     family = poisson)
#> Working 2logLR =  21.44332 p= 0.0014171
#> (scale factors:  1.2 1.1 1 0.87 0.76 );  denominator df= 130

# Compare AIC values across all three models for overall fit
AIC(fit1a, fit2a, fit3a)
#>         eff.p      AIC deltabar
#> [1,] 24.73244 17872.62 1.374024
#> [2,] 34.63283 17878.17 1.049480
#> [3,] 30.80173 17858.35 1.339206
```

- R Code Chunk 7: Process Sex-Specific Hazard Ratios for Plotting

The following code extracts the Hazard Ratios (HRs) and their 95% Confidence Intervals (CIs) for both male and female subgroups from the `f3` object (the published output of fit3). These HRs reflect the effect of smoking initiation within each sex level The extracted data is formatted into data frame, `f3r`.

```
# Extract HRs and CIs for male and female subgroups from the published output
f3rM <- f3$regressionTable[21:25,c("HazardRatio","CI.95")]
f3rF <- f3$regressionTable[26:30,c("HazardRatio","CI.95")]

# Assign group labels and combine into a single data frame
f3rM$group <- "Male"
f3rF$group <- "Female"
f3r <- rbind(f3rM, f3rF)

# Process CIs: Extract lower and upper bounds from the CI string
ci <- str_extract_all(f3r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f3r$CI.l <- as.numeric(as.character(ci[,1]))
f3r$CI.u <- as.numeric(as.character(ci[,2]))
names(f3r) <- c("mean","CI.95","group","lower","upper")

# Display the final processed data frame
f3r
#>    mean       CI.95  group lower upper
#> 21 2.54 [1.86;3.47]   Male  1.86  3.47
#> 22 2.05 [1.81;2.32]   Male  1.81  2.32
#> 23 1.83 [1.61;2.09]   Male  1.61  2.09
#> 24 1.40 [1.22;1.61]   Male  1.22  1.61
#> 25 1.42 [1.22;1.64]   Male  1.22  1.64
#> 26 2.90 [1.21;6.92] Female  1.21  6.92
#> 27 3.15 [2.49;3.97] Female  2.49  3.97
#> 28 2.07 [1.81;2.37] Female  1.81  2.37
#> 29 1.53 [1.32;1.78] Female  1.32  1.78
#> 30 1.63 [1.44;1.85] Female  1.44  1.85
```

11.3 Visualizing the Results (Figure 2)

- R Code Chunk 8: Forest Plot Visualization of Hazard Ratios

The following code prepares the data for the comprehensive forest plot. It binds together the Hazard Ratios from the crude (`f0r`), adjusted (`f1r`), race-specific (`f2r`), and sex-specific (`f3r`) models into a single data frame fr. It then converts relevant columns to numeric types and assigns descriptive labels (`age.grp`).

```
# Bind all processed HR data frames and set as data frame
fr <- rbind(f0r,f1r,f2r,f3r)
fr <- as.data.frame(fr)

# Convert to numeric columns
fr[,c(1,4,5)] <- sapply(fr[,c(1,4,5)], as.numeric)

# Assign 'age.grp' labels
fr$age.grp <- c('Started before 10', 'Started at 10-14',
                "Started at 15-17", "Started at 18-20",
                "Started after 20")

# Save the final plotting dataframe
saveRDS(fr, file = "data/fr.rds")
```

The following code uses the `ggplot2` package to visualize the comprehensive forest plot with the estimated Hazard Ratios and their 95% Confidence Intervals from all models (crude, adjusted, and subgroup-specific). This corresponds to Figure 2 in the published paper.

```
# Set factor levels for correct ordering in the plot
fr$age.grp <- factor(fr$age.grp,
                     levels = c('Started before 10',
                                'Started at 10-14',"Started at 15-17",
                                "Started at 18-20","Started after 20"))
fr$group <- factor(fr$group,
                     levels = rev(c("Crude", "Adjusted",
                                "White", "Black", "Hispanic",
                                "Male", "Female")))

levels(fr$group)[levels(fr$group) == "White"] <- "Non-Hispanic White"
levels(fr$group)[levels(fr$group) == "Black"] <- "Non-Hispanic Black"
fr$text_label <- sprintf("%.2f [%.2f;%.2f]", fr$mean, fr$lower, fr$upper)

# Create plot
ggplot(fr,
       aes(x = mean, y = group, colour = age.grp)) +
  geom_errorbar(aes(xmin = lower, xmax = upper),
```

```
              position = position_dodge(0.9), width = 0.25) +
geom_point(position = position_dodge(0.9), shape = 1) +
geom_text(aes(label = text_label, x = upper, hjust = -0.05),
          position = position_dodge(0.9), size = 3) +
scale_x_continuous(limits = c(0, 9)) +

labs(x = "Hazard Ratio", y = "",
     legend=TRUE, col = "Age group") +
theme_classic() +
theme(panel.grid.major.x = element_blank(),
      panel.border = element_blank(),
      legend.title=element_text(size=12),
      legend.text=element_text(size=12),
      plot.title = element_text(hjust = 0.5),
      axis.text.x = element_text(size = 10, face = "bold"),
      axis.text.y = element_text(size = 10, face = "bold"),
      legend.position=c(.8, .8)) +
geom_vline(xintercept = 1, linetype = "dashed", color = "grey")+
guides(color = guide_legend(override.aes = list(shape = 1, linetype = 1, size = 1)))
```

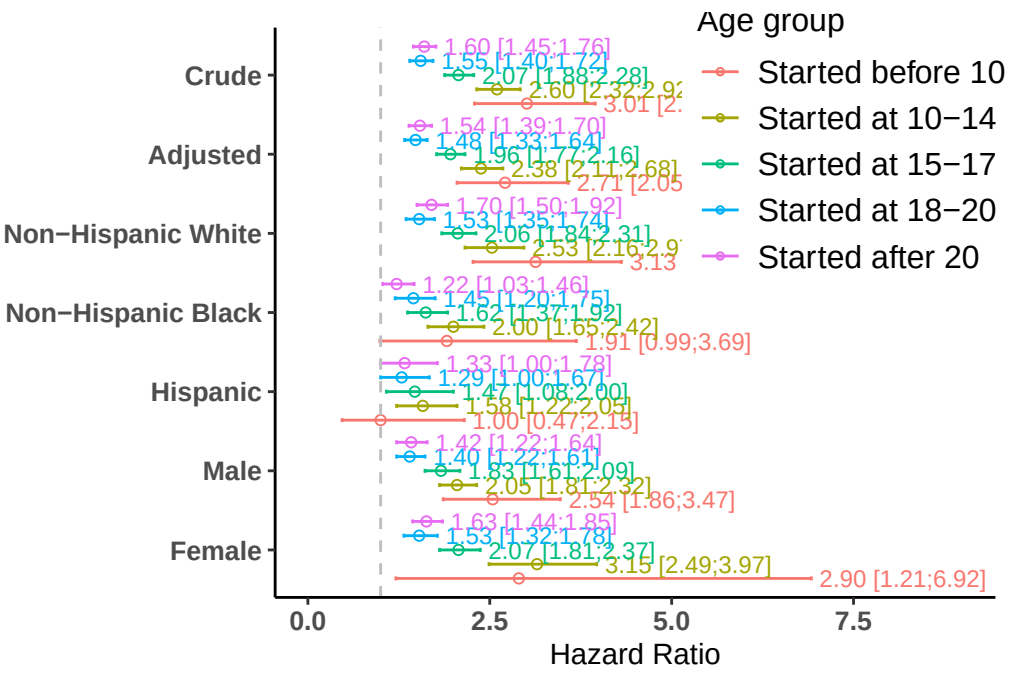


Figure 11.1: Forest plot summarizing Hazard Ratios (HRs) for all-cause mortality associated with age of smoking initiation. Results from the crude, main adjusted, and stratified models (by race/ethnicity and sex) are shown. This plot reproduces Figure 2 from the paper.

```
# Save
ggsave("images/forest.png")
```

11.4 Quantifying Additive Interaction (RERI)

- R Code Chunk 9: Relative Excess Risk Due to Interaction (RERI) Function

The following code creates the `reri.f` function which calculates the relative excess risk due to interaction (RERI). RERI is a common measure of additive interaction, providing insight into whether the combined effect of two factors is greater than the sum of their individual effects. The function computes the RERI estimate and its confidence interval based on the delta method, accounting for the variance and covariance of the estimated coefficients.

```r
reri.f <- function(model, coef, conf.level = 0.95){
  N. <- 1 - ((1 - conf.level)/2)
  z <- qnorm(N., mean = 0, sd = 1)
  theta1 <- as.numeric(model$coefficients[coef[1]])
  theta2 <- as.numeric(model$coefficients[coef[2]])
  theta3 <- as.numeric(model$coefficients[coef[3]])
  theta1.se <- summary(model)$coefficients[coef[1],2]
  theta2.se <- summary(model)$coefficients[coef[2],2]
  theta3.se <- summary(model)$coefficients[coef[3],2]

  cov.mat <- vcov(model)
  h1 <- exp(theta1 + theta2 + theta3) - exp(theta1)
  h2 <- exp(theta1 + theta2 + theta3) - exp(theta2)
  h3 <- exp(theta1 + theta2 + theta3)

  reri.var <- (h1^2 * theta1.se^2) +
       (h2^2 * theta2.se^2) +
       (h3^2 * theta3.se^2) +
       (2 * h1 * h2 * cov.mat[coef[1],coef[2]]) +
       (2 * h1 * h3 * cov.mat[coef[1],coef[3]]) +
       (2 * h2 * h3 * cov.mat[coef[2],coef[3]])
  reri.se <- sqrt(reri.var)

  reri.p <- exp(theta1 + theta2 + theta3) - exp(theta1) - exp(theta2) + 1
  reri.l <- reri.p - (z * reri.se)
  reri.u <- reri.p + (z * reri.se)
  reri <- data.frame(est = reri.p, lower = reri.l, upper = reri.u)
return(reri)
}
```

- R Code Chunk 10: Calculate and Report RERI for Sex Interaction

Corresponds to Appendix Table 5

The following code calculates the RERI for the interaction between smoking initiation age (`exposure.cat`) and `sex`, using the `fit3` model. The calculations are performed for each level of exposure.cat relative to its reference level. The results are presented in a table and saved to an Excel file (reriM2.xlsx).

```r
# Get the names of the coefficients from the sex-interaction model
coef.val <- names(fit3$coefficients)

# Define the active (non-reference) levels for exposure and the moderator (sex)
exp.active.lev <- paste0("exposure.cat", levels(w.design0$variables$exposure.cat)[-1])
mod.active.lev <- paste0("sex", levels(w.design0$variables$sex)[-1])
int.term <- paste(exp.active.lev, mod.active.lev, sep=":")
rr <- NULL

# Loop through each exposure level to calculate RERI
for (i in 1:length(exp.active.lev)){
  coef.val1 <- coef.val[c(which(coef.val==exp.active.lev[i]),
                          which(coef.val==mod.active.lev[1]),
                          which(coef.val==int.term[i]))]
  r1 <- reri.f(model=fit3, coef=coef.val1, conf.level = 0.95)
  rr <- rbind(rr,r1)
}

# Assign row names for clarity
row.names(rr) <- levels(w.design0$variables$exposure.cat)[-1]
```

```r
expss::xl_write_file(rr, filename = "data/reriM2.xlsx", rownames = TRUE)
kable(rr,
      booktabs = TRUE, digits = 2,
      caption="Appendix Table 5: Reporting RERI for the Sex Interaction.",
      col.names = c("Levels", "RERI", "Lower CI", "Upper CI")) %>%
  kable_styling(latex_options = "hold_position")
```

Table 11.2: Appendix Table 5: Reporting RERI for the Sex Interaction.

Levels	RERI	Lower CI	Upper CI
Started before 10	-0.26	-5.80	5.27
Started at 10-14	0.39	-6.24	7.02
Started at 15-17	-0.12	-3.70	3.47
Started at 18-20	-0.04	-2.52	2.43
Started after 20	0.01	-2.69	2.71

- R Code Chunk 11: Calculate and Report RERI for Race/Ethnicity Interaction

Corresponds to Appendix Table 4

The following code calculates the RERI for the interaction between smoking initiation age (`exposure.cat`) and `race`, using the `fit2` model. The calculations are performed for each exposure.cat level and for each non-reference racial/ethnic group. The results are presented in a table and saved to an Excel file (reriM3.xlsx).

```
# Get the names of the coefficients from the race-interaction model
coef.val <- names(fit2$coefficients)

# Define the active (non-reference) levels for exposure and the moderator (race)
exp.active.lev <- paste0("exposure.cat", levels(w.design0$variables$exposure.cat)[-1])
rmx <- NULL

# Outer loop: iterate through each non-reference race category
for (j in 1:length(levels(w.design0$variables$race)[-1])){
  mod.active.lev.j <- paste0("race", levels(w.design0$variables$race)[-1])[j]
  int.term <- paste(exp.active.lev, mod.active.lev.j, sep=":")

  # Inner loop: iterate through each non-reference exposure category
  rr <- NULL
  for (i in 1:length(exp.active.lev)){
    coef.val1 <- coef.val[c(which(coef.val==exp.active.lev[i]),
                            which(coef.val==mod.active.lev[1]),
                            which(coef.val==int.term[i]))]
    r1 <- reri.f(model=fit2, coef=coef.val1, conf.level = 0.95)
    rr <- rbind(rr,r1)
  }
  row.names(rr) <- paste(levels(w.design0$variables$exposure.cat)[-1],
                         levels(w.design0$variables$race)[-1][j])
  rmx <- rbind(rmx,rr)
}
```

```
expss::xl_write_file(rmx, filename = "data/reriM3.xlsx", rownames = TRUE)
kable(rmx,
      booktabs = TRUE, digits = 2,
      caption="Appendix Table 4: Reporting RERI for the Race/Ethnicity Interaction.",
      col.names = c("Levels", "RERI", "Lower CI", "Upper CI")) %>%
  kable_styling(latex_options = "hold_position")
```

Table 11.3: Appendix Table 4: Reporting RERI for the Race/Ethnicity Interaction.

Levels	RERI	Lower CI	Upper CI

Started before 10 Black	-1.45	-12.14	9.24
Started at 10-14 Black	-0.79	-6.58	5.00
Started at 15-17 Black	-0.60	-4.58	3.38
Started at 18-20 Black	-0.20	-2.67	2.27
Started after 20 Black	-0.53	-3.45	2.40
Started before 10 Hispanic	-2.13	-16.78	12.52
Started at 10-14 Hispanic	-1.10	-8.02	5.82
Started at 15-17 Hispanic	-0.71	-4.94	3.52
Started at 18-20 Hispanic	-0.32	-2.68	2.05
Started after 20 Hispanic	-0.45	-3.27	2.37
Started before 10 Others	0.48	-9.80	10.76
Started at 10-14 Others	-0.04	-5.88	5.81
Started at 15-17 Others	0.18	-5.17	5.53
Started at 18-20 Others	-0.37	-2.74	2.00
Started after 20 Others	-0.64	-3.78	2.50

This concludes the main statistical analysis conducted for the published paper.

11.5 Chapter Summary and Next Steps

In this chapter, we delved deeper into the analysis by investigating effect modification. We fitted interaction models and performed statistical tests to assess whether the relationship between smoking initiation and mortality was modified by race/ethnicity and sex. Our findings, particularly the significant interaction by sex, align with the conclusions of the original paper.

To ensure the robustness of these findings, we will next perform "Sensitivity Analyses" to test how our results hold up when adjusting for socioeconomic factors and when analyzing a different time period.

12 Sensitivity Analyses

This chapter presents several analyses conducted to further assess and understand the primary findings regarding association between early-smoking initiation and mortality. It includes two sensitivity analyses and one exploratory analysis of a potential mediator.

- **Adjustment for Socioeconomic Status (SES) Proxies:** This analysis investigates whether the observed associations are substantially influenced by socioeconomic factors. This is achieved by adjusting for additional variables, such as family poverty income ratio (PIR), which compares household income to the poverty threshold adjusted for household size and composition (NHANES variable INDFMPIR) , and the education level of the household head (NHANES variable DMDHREDU). These variables serve as proxies for family socioeconomic status.
- **Effect Modification by Race/Ethnicity (2011-2018):** This analysis examines how findings might vary when considering the "non-Hispanic Asian" racial classification, which was introduced in the 2011 survey cycle. It focuses on effect modification by race/ethnicity within the survey cycles spanning 2011 to 2018, which incorporate the Asian race/ethnicity category.
- **Exploratory Analysis of Smoking Duration:** This analysis investigates the relationship between the age of smoking initiation and the total duration of smoking, a potential mediator in the pathway to mortality.

The key methods and the paper figures/tables they reproduce are summarized below:

Analysis	Purpose	Key Method	Reproduces
Adjustment for SES Proxies	To test if the main findings are robust to confounding by socioeconomic status.	Added `pir` (income) and `HHedu` (education) as covariates to the adjusted Cox model.	**Appendix Figure 5**

Analysis	Purpose	Key Method	Reproduces
Effect Modification by Race (2011-2018)	To specifically investigate the association within the non-Hispanic Asian population.	Subset the data to the 2011-2018 cycles and included the `race2` variable (with the 'Asian' category) in the effect modification model.	**Appendix Table 3** and **Appendix Figure 4**
Smoking Duration	Explore if earlier initiation is linked to longer smoking duration.	Create boxplots of smoking duration stratified by initiation age.	**Appendix Figures 1-3**

```r
# Load required packages
library(readr)
library(car)
library(plyr)
library(dplyr)
library(DataExplorer)
library(tableone)
library(Publish)
library(survey)
options(survey.want.obsolete=TRUE)
library(survival)
require(knitr)
require(kableExtra)
require(stringr)
library(ggplot2)
library(expss)
library(readxl)
library(patchwork)
library(forcats)
library(nhanesA)
```

12.1 Sensitivity Analysis 1: Adjustment for SES Proxies

12.1.1 Data Reprocessing with SES Variables

Our primary analysis did not include socioeconomic status (SES) variables because they had a high number of missing values. To ensure these factors were not confounding our results, we performed this sensitivity analysis. This required reprocessing the raw NHANES data to create a new dataset that includes the following 2 variables as well:

Demographics:

- **"INDFMPIR": Ratio of family income to poverty (family SES)**
- **"DMDHREDU": Education level of the household head (family SES)**

- R Code Chunk 1: Data Reprocessing and Cleaning with SES Variables

The code steps for data downloading, cleaning, and merging are nearly identical to the process in main analysis. The key new data preparation steps for this sensitivity analysis are the cleaning of the two SES proxy variables, pir and HHedu.

```
# File Names
demo <- c("DEMO", "DEMO_B", "DEMO_C", "DEMO_D", "DEMO_E",
          "DEMO_F", "DEMO_G", "DEMO_H", "DEMO_I", "DEMO_J")
smoking <- c("SMQ", "SMQ_B", "SMQ_C", "SMQ_D", "SMQ_E",
             "SMQ_F", "SMQ_G", "SMQ_H", "SMQ_I", "SMQ_J")

# 10 Data Files
demo_data_files <- readRDS("data/demo_data_files.rds")
smoking_data_files <- readRDS("data/smoking_data_files.rds")

# Set Columns
demo_columns <- c( "SEQN", "RIDAGEYR", "RIAGENDR", "RIDRETH1",
                   "INDFMPIR", "DMDHREDU", "DMDHREDZ", # *** these are new
                   "DMDBORN", "DMDBORN2", "DMDBORN4", "SDDSRVYR",
                   "WTINT2YR", "WTMEC2YR", "SDMVPSU", "SDMVSTRA")
smoking_columns <- c("SEQN", "SMQ020", "SMD030", "SMQ040", "SMD055")


# DEMOGRAPHICS
demo_data_files_2 <- lapply(seq_along(demo_data_files), function(i)
  {
```

```
  current_cycle_data <- demo_data_files[[i]]
  original <- demo[i]

  # Select Columns
  subset_data <- current_cycle_data %>%
    dplyr::select(dplyr::any_of(demo_columns))
  # Translate
  translated_data <- nhanesTranslate(original,
                                     names(subset_data),
                                     data = subset_data)
  # Return
  return(translated_data)
})


# SMOKING
smoking_data_files_2 <- lapply(seq_along(smoking_data_files), function(i)
  {
  current_cycle_data <- smoking_data_files[[i]]
  original <- smoking[i]

  # Select Columns
  subset_data <- current_cycle_data %>%
    dplyr::select(dplyr::any_of(smoking_columns))
  # Translate
  translated_data <- nhanesTranslate(original,
                                     names(subset_data),
                                     data = subset_data)
  # Return
  return(translated_data)
})


# MERGE
data_all <- lapply(seq_along(demo_data_files_2), function(i) {
  demo_df <- demo_data_files_2[[i]]
  smoking_df <- smoking_data_files_2[[i]]

  # Merge by SEQN
  merged_df <- join_all(list(demo_df, smoking_df),
                        by = "SEQN",
```

```
                        type = 'full')
  return(merged_df)
})


# Rename
data_all2 <- data_all
for (i in seq_along(data_all2)) {
  df <- as.data.frame(data_all2[[i]])

  if ("DMDBORN2" %in% names(df)) {
    names(df)[names(df) == "DMDBORN2"] <- "DMDBORN"
  } else if ("DMDBORN4" %in% names(df)) {
    names(df)[names(df) == "DMDBORN4"] <- "DMDBORN"
  }
  data_all2[[i]] <- df
}
```

```
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  ## ID
  dat2$id <- dat2$SEQN

  ## Demographic
    ### Age (for eligibility)
  dat2$age <- dat2$RIDAGEYR

    ### Sex
  dat2$sex <- dat2$RIAGENDR

    ### Race/Ethnicity
  dat2$race <- dat2$RIDRETH1
  dat2$race <- car::recode(dat2$race, recodes = "
    'Non-Hispanic White'='White';
    'Non-Hispanic Black'='Black';
    c('Mexican American','Other Hispanic')='Hispanic';
    else='Others'")
  dat2$race <- factor(dat2$race,
```

```
                    levels = c("White", "Black", "Hispanic", "Others"))

    ### Country of birth
  dat2$born <- dat2$DMDBORN
  dat2$born <- car::recode(dat2$born, recodes = "
    c('Born in Mexico','Born Elsewhere', 'Others') = 'Other place';
    c('Born in 50 US States or Washington, DC', 'Born in 50 US states or Washington, DC')
    else = NA")
  dat2$born <- factor(dat2$born,
                    levels = c("Born in US", "Other place"))

  data_all2[[i]] <- dat2
}
```

Poverty Income Ratio: The poverty income ratio (INDFMPIR) is provided as a continuous variable. To make it easier to use in our models, the following code uses the `cut()` function to categorize the numeric ratio into five distinct, interpretable groups, from "Below Poverty Line" to "High Income."

```
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # Set Poverty Income Ratio
  dat2$pir <- dat2$INDFMPIR

  # Cut into five distinct groups
  dat2$pir <- cut(dat2$pir,
                  breaks = c(-Inf, 1, 1.99, 3.99, 4.99, Inf),
                  labels = c("Below Poverty Line",
                             "Near Poverty Line",
                             "Low to Middle Income",
                             "Middle to High Income",
                             "High Income"),
                  right = FALSE)

  # Return
  data_all2[[i]] <- dat2
}
```

Household Head Education Level: The coding for household head education also changes over the survey years, with the 2017-2018 cycle using a different variable name (DMDHREDZ)

and different text labels. The code below uses an if-else statement to handle this inconsistency. It applies one set of recoding rules for the first nine cycles and a different set for the final 2017-2018 cycle, ensuring the final HHedu variable is harmonized across all datasets.

```
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]
  if (i < 10) {

  # Set Education Level of the Household Head
  # Now, use case_when() for robust recoding
  dat2$HHedu <- case_when(
    dat2$DMDHREDU %in% c("Less Than 9th Grade", "9-11th Grade (Includes 12th grade with no
    dat2$DMDHREDU %in% c("High School Grad/GED or equivalent", "Some College or AA degree"
    dat2$DMDHREDU == "College Graduate or above" ~ "College graduate or above",
    TRUE ~ NA_character_ # This handles "Refused", "Don't know", and any other non-matches
    )

  # Convert to a factor with the correct levels
  dat2$HHedu <- factor(dat2$HHedu, levels = c('Less than high school',
                                              'High school or college',
                                              'College graduate or above'))
  # Return
  data_all2[[i]] <- dat2
  }

  # For the 10th dataset (2017-2018)
  else {
    dat2$HHedu <- as.character(dat2$DMDHREDZ)
    dat2$HHedu <- dplyr::recode(dat2$HHedu,
                          `Less than high school degree` = 'Less than high school',
                          `High school grad/GED or some college/AA degree` = 'High school o
                          `College graduate or above` = 'College graduate or above',
                          .default = NA_character_
                          #`Refused` = NA_character_,
                          #`Don't know` = NA_character_
                          )
    dat2$HHedu <- factor(dat2$HHedu, levels = c('Less than high school',
                                                'High school or college',
                                                'College graduate or above'))

    # Return
```

```
    data_all2[[i]] <- dat2
    }
}
```

```
# Remaining Variables

for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]
  ## Smoking
    ### Status
  dat2$smoking <- dat2$SMQ020
  dat2$smoking <- car::recode(dat2$smoking, " 'Yes' = 'Current smoker'; 'No' = 'Never smok
  dat2$smoking <- factor(dat2$smoking, levels = c("Never smoker", "Previous smoker", "Curr
  dat2$smoking[dat2$SMQ040 == "Not at all?" | dat2$SMQ040 == "Not at all"] <- "Previous sm

    ### Age smoking started
  dat2$smoking.age <- dat2$SMD030
  dat2$smoking.age[dat2$smoking.age == 777] <- NA
  dat2$smoking.age[dat2$smoking.age == 999] <- NA
  dat2$smoking.age[dat2$smoking == "Never smoker"] <- 99 # 99 means never smoker

    ### Whether smoking started age of 15 or before
  dat2$smoked.while.child <- car::recode(dat2$smoking.age, " 0 = 'No'; 1:15 = 'Yes'; else

  ## Survey features
    ### Weight
  dat2$survey.weight <- dat2$WTINT2YR #WTMEC2YR # MEC weights

    ### PSU
  dat2$psu <- as.factor(dat2$SDMVPSU)

    ### Strata
  dat2$strata <- as.factor(dat2$SDMVSTRA)

  ## Survey year
  dat2$year <- dat2$SDDSRVYR

  data_all2[[i]] <- dat2
}
```

NOTE: The following code chunk recodes the variable, `SMD055`, which stands for the age

cigarettes were last smoked regularly. This variable is not available for the cycle 2017-2018, thus it is set as `NA`. This part is essential for the plots in section 12.3 Appendix B: Exploratory Analysis of Smoking Duration below.

```
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]
  if (i < 10) {
    ### Age smoking quit
    dat2$smoking.quit.age <- dat2$SMD055
    dat2$smoking.quit.age[dat2$smoking.quit.age == 777] <- NA
    dat2$smoking.quit.age[dat2$smoking.quit.age == 999] <- NA
  }
  else {
    dat2$smoking.quit.age <- NA
  }
  data_all2[[i]] <- dat2
}
```

- R Code Chunk 2: Save New Files

After the same remaining data processing steps are taken, theses 10 new files are also saved with the selected variables saved in `vars`.

```
nhanes_all <- c("nhanes00", "nhanes01", "nhanes03", "nhanes05",
                "nhanes07", "nhanes09", "nhanes11",
                "nhanes13", "nhanes15", "nhanes17")

vars <- c("id", "age", "sex", "race", "born", "pir", "HHedu",
          "smoking.age", "smoking.quit.age",
          "smoked.while.child", "smoking",
          "survey.weight", "psu", "strata", "year")


cycle_years <- c("1999-2000", "2001-2002", "2003-2004", "2005-2006", "2007-2008",
                 "2009-2010", "2011-2012", "2013-2014", "2015-2016", "2017-2018")

for (i in seq_along(data_all2)) {
  dat2 <- data_all2[[i]]
```

```
  nhanes_i <- nhanes_all[i]
  assign(nhanes_i, dat2[, vars], envir = .GlobalEnv)

  analytic <- subset(get(nhanes_i), age >= 20)
  cat("Processing:", nhanes_i, "\n")
  print(dim(analytic))

  analytic_i <- paste0("analytic", substr(nhanes_i, 7, 8))
  assign(analytic_i, analytic, envir = .GlobalEnv)

  # Create 'data2' directory if it does not exist
  if (!dir.exists("data2")) {
    dir.create("data2")
  }

  # Save
  save(list = c(nhanes_i, analytic_i),
       file = file.path("data2", paste0(analytic_i, ".RData")))
}

# Bind all
dat.full <- rbind(nhanes00, nhanes01, nhanes03,
                  nhanes05, nhanes07, nhanes09,
                  nhanes11, nhanes13, nhanes15,
                  nhanes17)

# Corrected weights
dat.full$survey.weight.new <- dat.full$survey.weight/length(unique(dat.full$year))
dat.full$survey.weight <- NULL
names(dat.full)
dim(dat.full)
```

```
# Mortality Data
mort2000 <- readRDS(file = "data/Mortalitydata/mort2000.RData")
mort2001 <- readRDS(file = "data/Mortalitydata/mort2001.RData")
mort2003 <- readRDS(file = "data/Mortalitydata/mort2003.RData")
mort2005 <- readRDS(file = "data/Mortalitydata/mort2005.RData")
mort2007 <- readRDS(file = "data/Mortalitydata/mort2007.RData")
mort2009 <- readRDS(file = "data/Mortalitydata/mort2009.RData")
mort2011 <- readRDS(file = "data/Mortalitydata/mort2011.RData")
mort2013 <- readRDS(file = "data/Mortalitydata/mort2013.RData")
```

```r
mort2015 <- readRDS(file = "data/Mortalitydata/mort2015.RData")
mort2017 <- readRDS(file = "data/Mortalitydata/mort2017.RData")

# Merge
dat.mortality <- rbind(mort2000, mort2001, mort2003,
                       mort2005, mort2007, mort2009,
                       mort2011, mort2013, mort2015,
                       mort2017)
table(dat.mortality$mort_eligstat, useNA = "always")

# Merging analytic and mortality data
dat.full.with.mortality <- merge(dat.full, dat.mortality, by = "id", all.x = TRUE)
dim(dat.full.with.mortality)
```

```r
# Recoding
dat.full.with.mortality$exposure.cat <- car::recode(dat.full.with.mortality$smoking.age, "
                                                    c(0,99) = 'Never smoked';
                                                    1:9 = 'Started before 10';
                                                    10:14 = 'Started at 10-14';
                                                    15:17 = 'Started at 15-17';
                                                    18:20 = 'Started at 18-20';
                                                    21:80 = 'Started after 20';
                                                    else = NA ",
                                                    as.factor = TRUE)
dat.full.with.mortality$exposure.cat <- factor(dat.full.with.mortality$exposure.cat,
                                               levels = c("Never smoked",
                                                          'Started before 10',
                                                          'Started at 10-14',
                                                          "Started at 15-17",
                                                          "Started at 18-20",
                                                          "Started after 20"))

# Survival time - Person-Months of Follow-up from NHANES Interview date
# changed to "Number of Person Months of Follow-up from NHANES Mobile Examination Center (
dat.full.with.mortality$stime.since.interview <- dat.full.with.mortality$mort_permth_int #
summary(dat.full.with.mortality$stime.since.interview)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>     0.0    59.0   113.0   117.9   172.0   250.0   42252

# Age in month
dat.full.with.mortality$age.month <- dat.full.with.mortality$age * 12
```

```r
summary(dat.full.with.mortality$age.month)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>     0.0   120.0   288.0   373.5   624.0  1020.0

# Survival time - Person-Months of Follow-up from birth = birth to screening + screening t
dat.full.with.mortality$stime.since.birth <- with(dat.full.with.mortality, age.month + sti
# converted back to year so that KM plot is in years / HR results remain the same (added)
dat.full.with.mortality$stime.since.birth <- dat.full.with.mortality$stime.since.birth/12
dat.full.with.mortality$stime.since.birth[is.na(dat.full.with.mortality$stime.since.interv
summary(dat.full.with.mortality$stime.since.birth)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>   19.08   41.50   57.17   57.56   72.92  105.67   42252

# All-cause mortality status
dat.full.with.mortality$status_all <- dat.full.with.mortality$mort_stat
table(dat.full.with.mortality$status_all, useNA = "always")
#>
#>     0     1  <NA>
#> 49815  9249 42252

# Cause-specific mortality
dat.full.with.mortality$mort_diabetes <- NULL
dat.full.with.mortality$mort_hyperten <- NULL
dat.full.with.mortality$mort_permth_int <- NULL
dat.full.with.mortality$mort_permth_exm <- NULL
dat.full.with.mortality$mort_ucod_leading <- NULL
dat.full.with.mortality$mort_eligstat <- NULL

# Recode year variable
dat.full.with.mortality$year.cat <- dat.full.with.mortality$year

new_levels <- c(
  "1999-2000", "2001-2002", "2003-2004", "2005-2006", "2007-2008",
  "2009-2010", "2011-2012", "2013-2014", "2015-2016", "2017-2018"
)
levels(dat.full.with.mortality$year.cat) <- new_levels

# Display the unique levels of year.cat
levels(dat.full.with.mortality$year.cat)
#>  [1] "1999-2000" "2001-2002" "2003-2004" "2005-2006" "2007-2008" "2009-2010"
#>  [7] "2011-2012" "2013-2014" "2015-2016" "2017-2018"
```

```r
# Create the analytic subset
dat.full.with.mortality$smoked.while.child <- NULL

### Analytic dataset - age 20 - 79
dat.analytic <- subset(dat.full.with.mortality, age>=20 & age < 80)
dim(dat.full.with.mortality)
#> [1] 101316     21
dim(dat.analytic)
#> [1] 50824    21
dim(dat.full.with.mortality)[1] - dim(dat.analytic)[1] # added
#> [1] 50492

# Drop variables that are not being used
dat.analytic$stime.since.interview <- NULL
dat.analytic$born <- NULL
dat.analytic$age <- NULL
dat.analytic$age.month <- NULL

### Create a copy of the data for the smoking duration analysis later in the chapter
dat.analytic.duration.analysis <- dat.analytic

# Drop quit related variables
dat.analytic$smoking.quit.age <- NULL
```

- R Code Chunk 3: PIR and HHedu

The following code gives a summary table of the two variables: poverty income ratio (PIR) and education level of the household head.

```r
table(dat.analytic$pir, useNA = "always")
#>
#>    Below Poverty Line     Near Poverty Line  Low to Middle Income
#>                  9720                 11890                 12281
#> Middle to High Income           High Income                  <NA>
#>                  3808                  8378                  4747
table(dat.analytic$HHedu, useNA = "always")
#>
#>     Less than high school    High school or college College graduate or above
#>                     13009                     21704                     11048
#>                      <NA>
```

```
#>                       5063
```

- R Code Chunk 4: Accounting for Missing Data

Before analysis, we must handle missing data. The following code perform a complete-case analysis by first removing participants with missing exposure (`exposure.cat`) or outcome (`stime.since.birth`) information. Then, any remaining missing values in the covariates are handled, and the number of participants dropped at each stage is calculated.

```
# Sequentially remove participants with missing exposure or outcome
dat.analytic1 <- dat.analytic[complete.cases(dat.analytic$stime.since.birth),]
dat.analytic2 <- dat.analytic1[complete.cases(dat.analytic1$exposure.cat),]

# Create the final complete-case dataset by removing any other missing values
dat.complete <- na.omit(dat.analytic2)
dim(dat.complete)
#> [1] 41671    16

# Report on participants dropped
cat("Participants dropped due to missing exposure or outcome:",
    nrow(dat.analytic) - nrow(dat.analytic2), "\n")
#> Participants dropped due to missing exposure or outcome: 275
cat("Participants dropped due to missing covariates:",
    nrow(dat.analytic2) - nrow(dat.complete), "\n")
#> Participants dropped due to missing covariates: 8878
cat("Total participants dropped:",
    nrow(dat.full.with.mortality) - nrow(dat.complete), "\n")
#> Total participants dropped: 59645

# Overwrite dat.analytic2 to be the final complete dataset for analysis
dat.analytic2 <- dat.complete
dim(dat.analytic2)
#> [1] 41671    16

# Save
save(dat.full.with.mortality, dat.analytic2, file = "data/SensAnalysis.RData")
```

- R Code Chunk 5: Descriptive Statistics

The following code creates a descriptive summary table of the final analytic sample using the `CreateTableOne` function. The table shows the distribution of the exposure, race, and sex, stratified by mortality status.

```
tab1 <- CreateTableOne(vars = c("exposure.cat", "race", "sex"),
                       strata = "status_all",
                       data = dat.analytic2,
                       addOverall = TRUE,
                       test = TRUE)

tab1$CatTable
#>                        Stratified by status_all
#>                         Overall       0             1            p      test
#>   n                      41671         36426         5245
#>   exposure.cat (%)                                                  <0.001
#>      Never smoked       23570 (56.6)  21576 (59.2)  1994 (38.0)
#>      Started before 10    280 ( 0.7)    201 ( 0.6)    79 ( 1.5)
#>      Started at 10-14    3203 ( 7.7)   2560 ( 7.0)   643 (12.3)
#>      Started at 15-17    5949 (14.3)   4924 (13.5)  1025 (19.5)
#>      Started at 18-20    5194 (12.5)   4338 (11.9)   856 (16.3)
#>      Started after 20    3475 ( 8.3)   2827 ( 7.8)   648 (12.4)
#>   race (%)                                                          <0.001
#>      White              18234 (43.8)  15462 (42.4)  2772 (52.9)
#>      Black               8826 (21.2)   7571 (20.8)  1255 (23.9)
#>      Hispanic           10760 (25.8)   9735 (26.7)  1025 (19.5)
#>      Others              3851 ( 9.2)   3658 (10.0)   193 ( 3.7)
#>   sex = Female (%)      21429 (51.4)  19235 (52.8)  2194 (41.8)  <0.001
```

12.1.2 Analysis with SES Adjustment

- R Code Chunk 6: Specify the Survey Design

Once again, to ensure our results are representative, we must account for the complex sampling design of NHANES. The following code creates a survey design object using the `svydesign()` function, incorporating the survey weights, strata, and PSU variables

```
# Create a flag to identify participants in our final analytic sample
dat.full.with.mortality$miss <- 1
dat.full.with.mortality$miss[dat.full.with.mortality$id %in% dat.analytic2$id] <- 0
```

```r
table(dat.full.with.mortality$miss)
#>
#>     0     1
#> 41671 59645

# Create the survey design object
w.design <- svydesign(ids = ~psu, strata = ~strata,
                      weights = ~survey.weight.new,
                      data = dat.full.with.mortality, nest = T)

# Subset the design object to our analytic sample
w.design0 <- subset(w.design, miss == 0 & survey.weight.new > 0)
dim(w.design0)
#> [1] 41671    22
```

- R Code Chunk 7: Regression and Survival Analysis

The following section contains the code for core statistical models for the sensitivity analysis, including Kaplan-Meier curves and various Cox proportional hazards models.

Kaplan-Meier Survival Curve

The following code generates a survey-weighted Kaplan-Meier survival curve to visualize the probability of survival over time, stratified by the age of smoking initiation.

```r
dummy <- length(unique(as.factor(w.design0$variables$exposure.cat)))
formulax0 <- as.formula(Surv(stime.since.birth, status_all) ~ exposure.cat)
sA<-svykm(formulax0, design=w.design0)

# Plot
par(oma=rep(0,4))
par(mar=c(5,4,0,0) + 0.1)
plot(sA, pars=list(col=c(1:dummy)),
     xlab = "Time",
     ylab = "Proportion surviving")
legend("bottomleft", levels(as.factor(w.design0$variables$exposure.cat)),
       col = (1:dummy), lty = c(1,1))
```

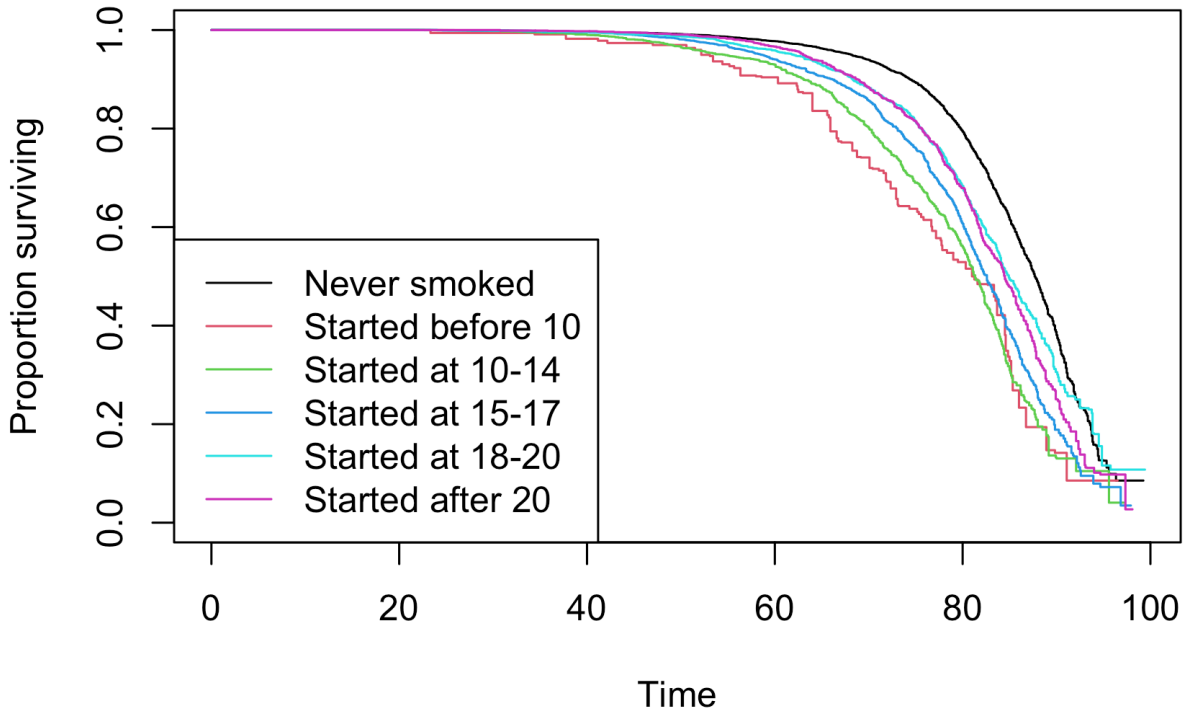


Figure 12.1: Sensitivity Analysis 1: Survey-Weighted Kaplan-Meier Curves

Log-Rank Test

The following code performs a weighted log-rank test to see if the differences between the curves are statistically significant.

```
# Weighted logrank test
lrt <- svylogrank(formulax0,design=subset(w.design0, survey.weight.new>0))
round(lrt[[2]],2)
#>  Chisq      p
#> 271.25   0.00
```

Cox Proportional Hazards Models

The following code fits several survey-weighted Cox models to estimate the Hazard Ratios (HRs) for all-cause mortality associated with smoking initiation age. We fit:

- 1. An unadjusted (crude) model.
- 2. An adjusted model, controlling for sex, race, survey year, and the SES proxy variables (pir and HHedu).

- 3. Two effect modification models to see if the association is different across racial/ethnic groups and by sex.

```
# Test One
fit0_ <- coxph(Surv(stime.since.birth, status_all) ~ exposure.cat, data = dat.complete)
fit0_


# Unadjusted model
fit0 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat,
                 design = w.design0)
fit0

# Adjusted model overall with SES proxies as covariates
fit1 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat + sex + race + year.ca
                 design = w.design0)
fit1

# Effect modification by race model + proxies
fit2 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat*race
                 + sex  + year.cat + pir +HHedu,
                 design = w.design0)
fit2

# Effect modification by sex model + proxies
fit3 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat*sex
                 + race + year.cat + pir + HHedu,
                 design = w.design0)
fit3
```

- R Code Chunk 8: Prepare Data for Forest Plot

The following code extracts the Hazard Ratios and 95% Confidence Intervals from each of the four models above: `fit0`, `fit1`, `fit2`, and `fit3`. It then combines them into a single data frame (`fr`) that will be used to create a summary forest plot.

```
# Extract results from crude model (fit0)
f0 <- publish(fit0)
f0r <- f0$regressionTable[2:6,c("HazardRatio","CI.95")]
f0r$group <- "Crude"
```

```r
# Extract results from adjusted model (fit1)
f1 <- publish(fit1)
f1r <- f1$regressionTable[2:6,c("HazardRatio","CI.95")]
f1r$group <- "Adjusted"

# Extract results from race interaction model (fit2)
f2 <- publish(fit2)
f2rW <- f2$regressionTable[(31:35)+8,c("HazardRatio","CI.95")]
f2rB <- f2$regressionTable[(36:40)+8,c("HazardRatio","CI.95")]
f2rH <- f2$regressionTable[(41:45)+8,c("HazardRatio","CI.95")]
f2rW$group <- "White"
f2rB$group <- "Black"
f2rH$group <- "Hispanic"
f2r <- rbind(f2rW, f2rB, f2rH)

# Extract results from sex interaction model (fit3)
f3 <- publish(fit3)
f3rM <- f3$regressionTable[29:33,c("HazardRatio","CI.95")]
f3rF <- f3$regressionTable[34:38,c("HazardRatio","CI.95")]
f3rM$group <- "Male"
f3rF$group <- "Female"
f3r <- rbind(f3rM, f3rF)

# Combine all results into one data frame
fr <- rbind(f0r,f1r,f2r,f3r)
fr <- as.data.frame(fr)
fr$age.grp <- c('Started before 10', 'Started at 10-14', "Started at 15-17",
                "Started at 18-20", "Started after 20")

# Clean up numeric columns
ci <- stringr::str_extract_all(fr$CI.95, '\\d+([.,]\\d+)?', simplify = TRUE)
fr$mean <- as.numeric(fr$HazardRatio)
fr$lower <- as.numeric(as.character(ci[,1]))
fr$upper <- as.numeric(as.character(ci[,2]))
#fr$CI.95 <- NULL
fr$HazardRatio <- NULL

# Save the plot data
frS <- fr
save(frS, file = "data/SensForest.RData")
```

12.1.3 Visualizing the SES-Adjusted Results

- R Code Chunk 9: Plot

The following code creates the forest plot using `ggplot2`. The first plot summarizes all the results from this sensitivity analysis. This plot corresponds to *Appendix Figure 5* in section C.2 Sensitivity analysis with proxy covariates, part of the supplementary materials published with the paper.

```
fr$age.grp <- factor(fr$age.grp,
                     levels = c('Started before 10', 'Started at 10-14',
                                "Started at 15-17", "Started at 18-20",
                                "Started after 20"))
fr$group <- factor(fr$group,
                   levels = rev(c("Crude", "Adjusted",
                                  "White", "Black", "Hispanic",
                                  "Male", "Female")))
levels(fr$group)[levels(fr$group) == "White"] <- "Non-Hispanic White"
levels(fr$group)[levels(fr$group) == "Black"] <- "Non-Hispanic Black"
fr$text_label <- sprintf("%.2f [%.2f;%.2f]", fr$mean, fr$lower, fr$upper)

# Plotting
ggplot(fr, aes(x = mean, y = group, colour = age.grp)) +
  geom_errorbar(aes(xmin = lower, xmax = upper), position = position_dodge(0.9), width = 0
  geom_point(position = position_dodge(0.9), shape = 1) +
  geom_text(aes(label = text_label, x = upper, hjust = -0.05), position = position_dodge(0
  scale_x_continuous(limits = c(0, 7.7)) +
  labs(x = "Hazard Ratio", y = "", legend=TRUE, col = "Age group") +
  theme_classic() +
  theme(panel.grid.major.x = element_blank(),
        panel.border = element_blank(),
        legend.title = element_text(size=12),
        legend.text = element_text(size=12),
        plot.title = element_text(hjust = 0.5),
        axis.text.x = element_text(size = 10, face = "bold"),
        axis.text.y = element_text(size = 10, face = "bold"),
        legend.position = c(.8, .8))+
  geom_vline(xintercept = 1, linetype = "dashed", color = "grey")+
  guides(color = guide_legend(override.aes = list(shape = 1, linetype = 1, size = 1)))
```

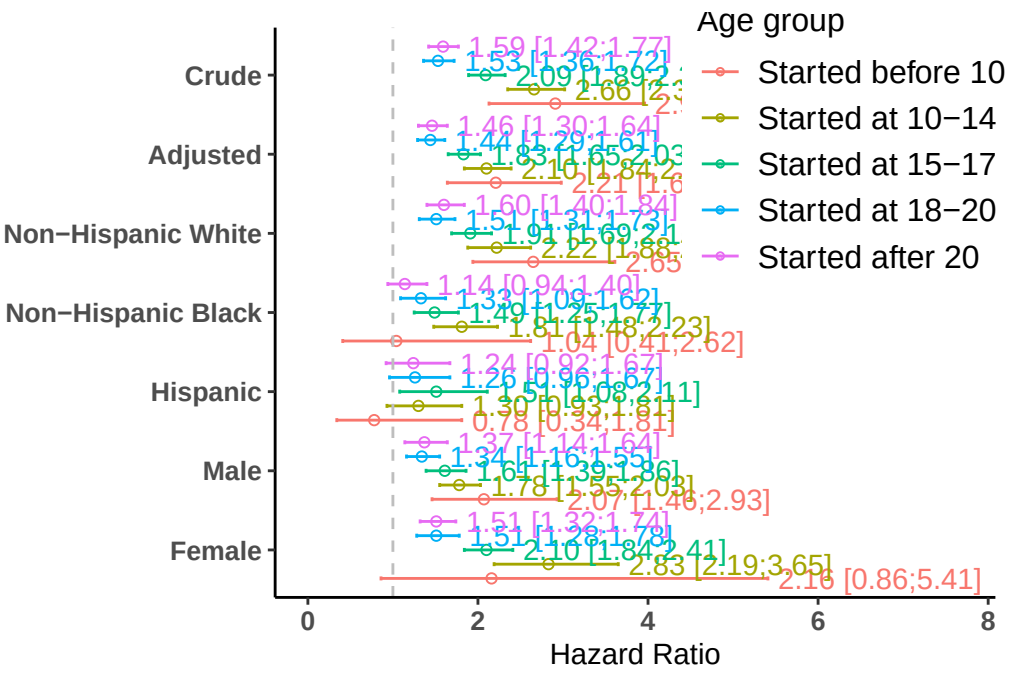


Figure 12.2: Appendix Figure 5: Forest plot of Hazard Ratios from the SES-adjusted sensitivity analysis.

```
# Save Plot
#ggsave("images/forestSens.png", width = 10, height = 8)
```

This second plot creates a direct visual comparison between the main analysis results and this sensitivity analysis.

```
# Load the data for both plots
load(file="data/SensForest.RData")
load(file="data/MainForest.RData")

# Combine data from fr (main) and frS (sensitivity), with an indicator variable
combined_data <- bind_rows(
  mutate(fr, dataset = 'Main Analysis'),
  mutate(frS, dataset = 'Sensitivity Analysis')
)
combined_data$group <- factor(combined_data$group, levels = unique(combined_data$group))

# Create the plot
ggplot(combined_data, aes(x = mean, y = group, colour = age.grp, shape = dataset)) +
```

```
geom_errorbar(aes(xmin = lower, xmax = upper), position = position_dodge(width = 0.7)) +
geom_point(position = position_dodge(width = 0.7), size = 2) +
scale_x_continuous(limits = c(0, 7)) +
labs(x = "Hazard Ratio", y = "", colour = "Age group", shape = "Analysis Type") +
theme_classic() +
theme(axis.text.x = element_text(size = 10, face = "bold"),
      axis.text.y = element_text(size = 10),
      legend.position = "right") +
geom_vline(xintercept = 1, linetype = "dashed", color = "grey")
```

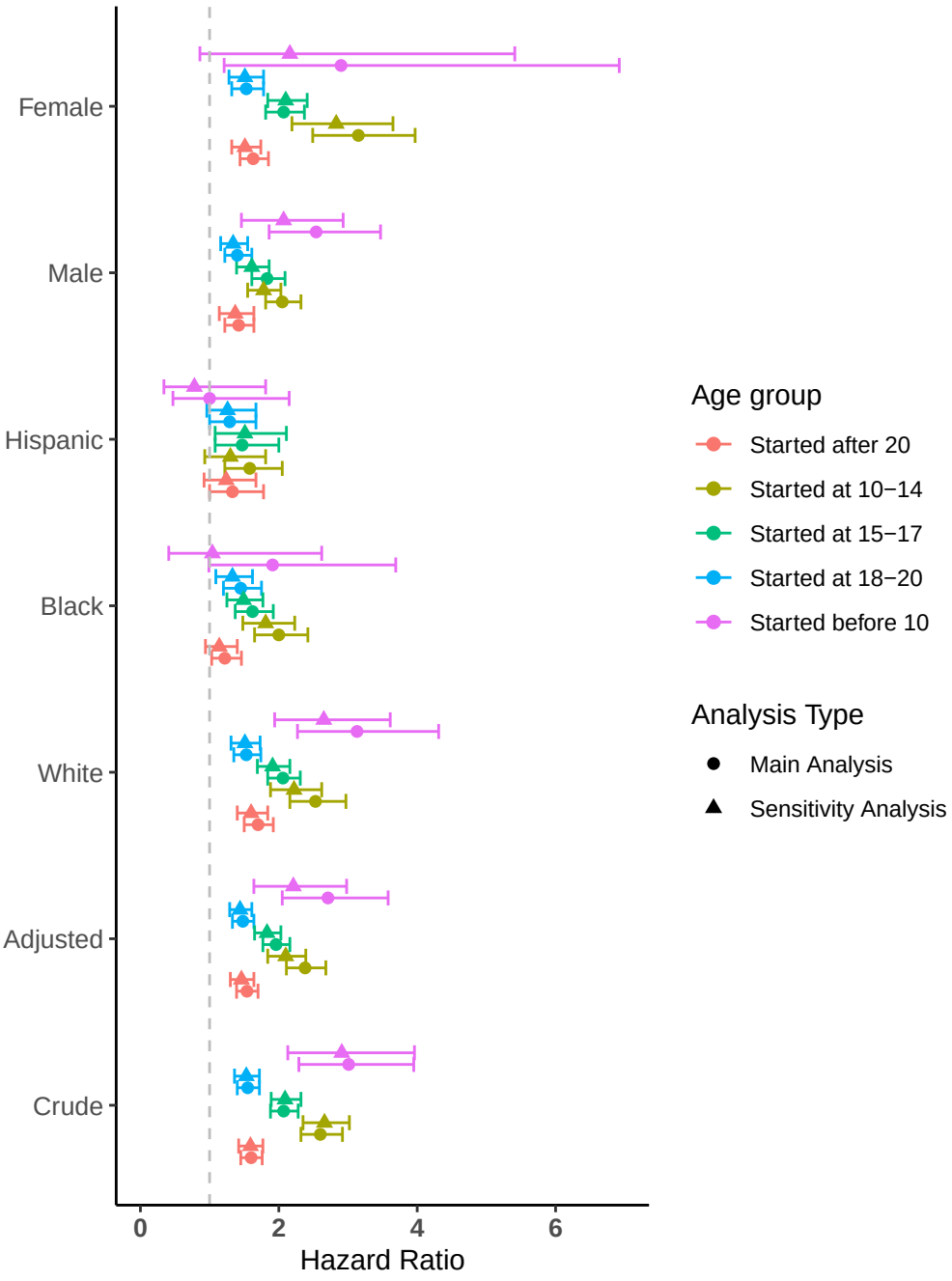


Figure 12.3: Comparison of results: Main vs. SES-adjusted Sensitivity Analysis.

```r
# Save Plot
#ggsave("images/combined_forest_plot.png", width = 10, height = 6)
```

12.2 Sensitivity Analysis 2: Effect Modification by Race/Ethnicity (2011-2018)

12.2.1 Data Reprocessing for the 2011-2018 Sub-period

The paper also notes that starting from the 2011—2012 cycle, NHANES introduced a new category labeled non-Hispanic Asian. To investigate the implications of this change, an additional analysis of effect modification by race/ethnicity was conducted specifically for the survey cycles spanning 2011-2018. The following sections re-run similar data merging and cleaning steps, and the different models as in the main analysis code sections, thus only the important changes are noted.

- R Code Chunk 1: Redo Data Preparation

The following code shows the respective data needed from the `nhanesA` library for the 4 cycles and the steps to prepare the specific dataset required for the 2011-2018 sub-period sensitivity analysis. This again, involves downloading raw data, merging them, and constructing necessary variables, mirroring the steps in the main data preparation but for this specific sub-period. The analysis focuses on the 4 cycles spanning 2011-2018.

The same steps are taken, thus only important changes are noted below.

- Load Analytic Data for 2011-2018 Cycles and Merge
- Load and Merge Mortality Data for Sub-period Analysis
- Variable Construction for Sub-period Data
- Construct Final Analytic Dataset for Sub-period Analysis
- Create Survey Design Object for Sub-period Analysis

```r
# 2011-2012, 2013-2014, 2015-2016, 2017-2018
demo <- c("DEMO_G", "DEMO_H", "DEMO_I", "DEMO_J")
smoking <- c("SMQ_G", "SMQ_H", "SMQ_I", "SMQ_J")

# DEMOGRAPHICS
demo_list_2 <- lapply(demo, nhanes)
demo_data_files <- demo_list_2
```

```
# SMOKING
smoking_list_2 <- lapply(smoking, nhanes)
smoking_data_files <- smoking_list_2
```

The following variables are selected from the data for 4 cycles spanning 2011-2018. The variable `RIDRETH3` includes the Race/Hispanic origin w/ NH Asian group. The remaining variables selected are the same as in the main analysis.

```
demo_columns_2 <- c("SEQN", "RIDAGEYR", "RIAGENDR",
                    "RIDRETH1", "RIDRETH3", "DMDBORN4",
                    "SDDSRVYR", "WTINT2YR", "WTMEC2YR",
                    "SDMVPSU", "SDMVSTRA")

smoking_columns_2 <- c("SEQN", "SMQ020", "SMD030", "SMQ040")
```

```
# DEMOGRAPHICS
demo_data_files_2 <- lapply(seq_along(demo_data_files), function(i)
  {
  current_cycle_data <- demo_data_files[[i]]
  original <- demo[i]

  # Select Columns
  subset_data <- current_cycle_data %>%
    dplyr::select(dplyr::any_of(demo_columns_2))
  # Translate
  translated_data <- nhanesTranslate(original,
                                     names(subset_data),
                                     data = subset_data)
  # Return
  return(translated_data)
})


# SMOKING
smoking_data_files_2 <- lapply(seq_along(smoking_data_files), function(i)
  {
  current_cycle_data <- smoking_data_files[[i]]
  original <- smoking[i]

  # Select Columns
  subset_data <- current_cycle_data %>%
```

```
    dplyr::select(dplyr::any_of(smoking_columns_2))
  # Translate
  translated_data <- nhanesTranslate(original,
                                     names(subset_data),
                                     data = subset_data)
  # Return
  return(translated_data)
})


# Merge
data_all <- lapply(seq_along(demo_data_files_2), function(i) {
  demo_df <- demo_data_files_2[[i]]
  smoking_df <- smoking_data_files_2[[i]]

  # Merge by SEQN
  merged_df <- join_all(list(demo_df, smoking_df),
                        by = "SEQN",
                        type = 'full')
  return(merged_df)
})
```

- R Code Chunk 2: Recode

The following sections of code apply the same recoding except we have another race/ethnicity this time.

ID:

```
data_all2 <- data_all

# ID
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # ID
  dat2$id <- dat2$SEQN
  data_all2[[i]] <- dat2
}
```

Demographic Variables:

- Below, the demographic variables like age, sex, and importantly, both race (from RIDRETH1) and race2 (from RIDRETH3 to include Asian) are recoded for the sub-period data.

```
# Demo
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # Age
  dat2$age <- dat2$RIDAGEYR

  # Sex
  dat2$sex <- dat2$RIAGENDR

  # Race/Ethnicity (RIDRETH1 and RIDRETH3)
  dat2$race <- dat2$RIDRETH1
  dat2$race <- car::recode(dat2$race, recodes = "
    'Non-Hispanic White'='White';
    'Non-Hispanic Black'='Black';
    c('Mexican American','Other Hispanic')='Hispanic';
    else='Others'")
  dat2$race <- factor(dat2$race,
                      levels = c("White", "Black", "Hispanic", "Others"))


  dat2$race2 <- dat2$RIDRETH3
  dat2$race2 <- car::recode(dat2$race2, recodes = "
            'Non-Hispanic White'='White';
            'Non-Hispanic Black'='Black';
            'Non-Hispanic Asian'='Asian';
            c('Mexican American','Other Hispanic')= 'Hispanic';
            else='Others' ")
  dat2$race2 <- factor(dat2$race2, levels = c("White", "Black",
                                              "Hispanic", "Asian",
                                              "Others"))



  # Country of birth / citizenship
```

```
  dat2$born <- dat2$DMDBORN4
  dat2$born <- car::recode(dat2$born, recodes = "
  'Others'='Other place';
  'Born in 50 US states or Washington, DC'= 'Born in US';
  else=NA")
  dat2$born <- factor(dat2$born,
                      levels = c("Born in US", "Other place"))

  data_all2[[i]] <- dat2
  }
```

Smoking:

- Below, the smoking-related variables, including smoking status, age started smoking and whether smoking started at age 15 or younger for the sub-period data are recoded.

```
# Smoking
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # Smoking Status
  dat2$smoking <- dat2$SMQ020
  dat2$smoking <- car::recode(dat2$smoking, "
    'Yes' = 'Current smoker';
    'No' = 'Never smoker';
    else = NA")
  dat2$smoking <- factor(dat2$smoking,
                         levels = c("Never smoker",
                                    "Previous smoker",
                                    "Current smoker"))

  # Ask about variable SMQ040 ***
  dat2$smoking[dat2$SMQ040 == "Not at all?" |
                 dat2$SMQ040 == "Not at all"] <- "Previous smoker"

  # Age Started Smoking
  dat2$smoking.age <- dat2$SMD030
  dat2$smoking.age[dat2$smoking.age %in% c(777, 999)] <- NA
  dat2$smoking.age[is.na(dat2$smoking.age) &
                     dat2$smoking == "Never smoker"] <- 0
```

```
  # Whether Smoking started age   15
  dat2$smoked.while.child <- car::recode(dat2$smoking.age,
  " 0 = 'No'; 6:15 = 'Yes'; else = NA ", as.factor = TRUE)

  data_all2[[i]] <- dat2
}
```

Survey Design:

```
# Survey Design
for (i in seq_along(data_all2)) {
  # Set Data
  dat2 <- data_all2[[i]]

  # Weight
  dat2$survey.weight <- dat2$WTINT2YR

  # PSU
  dat2$psu <- as.factor(dat2$SDMVPSU)

  # Strata
  dat2$strata <- as.factor(dat2$SDMVSTRA)

  # Survey year
  dat2$year <- dat2$SDDSRVYR

  data_all2[[i]] <- dat2
}
```

- R Code Chunk 3: Save Analytic Data for Sub-period Analysis (Per Cycle)

The following code processes the cleaned sub-period dataframes, applies the age filter (>20 years), selects the final set of variables, and saves them as .RData files for each cycle.

```
nhanes2_all <- c("nhanes2_11", "nhanes2_13", "nhanes2_15", "nhanes2_17")
vars_2 <- c("id", "age", "sex", "race", "race2", "born",
          "smoking.age", "smoked.while.child", "smoking",
          "survey.weight", "psu", "strata", "year")

for (i in seq_along(data_all2)) {
```

```
  dat2 <- data_all2[[i]]

  nhanes2_i <- nhanes2_all[i]
  assign(nhanes2_i, dat2[, vars_2], envir = .GlobalEnv)

  analytic <- subset(get(nhanes2_i), age >= 20)
  cat("Processing:", nhanes2_i, "\n")
  print(dim(analytic))

  analytic_i <- paste0("analytic2_", substr(nhanes2_i, 9, 10))
  analytic_i
  assign(analytic_i, analytic, envir = .GlobalEnv)

  # Create 'data' directory if it does not exist
  if (!dir.exists("data")) {
    dir.create("data")
  }
  # Save
  save(list = c(nhanes2_i, analytic_i),
       file = file.path("data", paste0(analytic_i, ".RData")))
  #print(analytic_i)
}
#> Processing: nhanes2_11
#> [1] 5560   13
#> Processing: nhanes2_13
#> [1] 5769   13
#> Processing: nhanes2_15
#> [1] 5719   13
#> Processing: nhanes2_17
#> [1] 5569   13
```

- R Code Chunk 4: Combine analytic files

The following code ensure the individual analytic .RData files for the sub-period are loaded and combines them into a single `dat.full.2` dataframe, recalculating weights for the combined period.

```
# Load the 4 new datasets
load(file="data/analytic2_11.RData")
load(file="data/analytic2_13.RData")
```

```r
load(file="data/analytic2_15.RData")
load(file="data/analytic2_17.RData")

# Bind
dat.full.2 <- rbind(nhanes2_11, nhanes2_13, nhanes2_15, nhanes2_17)

# Confirm
unique(dat.full.2$year)
#> [1] NHANES 2011-2012 public release NHANES 2013-2014 public release
#> [3] NHANES 2015-2016 public release NHANES 2017-2018 public release
#> 4 Levels: NHANES 2011-2012 public release ... NHANES 2017-2018 public release
length(unique(dat.full.2$year))
#> [1] 4

# Corrected weights
dat.full.2$survey.weight.new <- dat.full.2$survey.weight/length(unique(dat.full.2$year))
dat.full.2$survey.weight <- NULL
names(dat.full.2)
#>  [1] "id"                "age"               "sex"
#>  [4] "race"              "race2"             "born"
#>  [7] "smoking.age"       "smoked.while.child" "smoking"
#> [10] "psu"               "strata"            "year"
#> [13] "survey.weight.new"
dim(dat.full.2)
#> [1] 39156    13
```

- R Code Chunk 5: Load and Merge Mortality Data for Sub-period Analysis

The following code loads the individual mortality datasets for the 4 cycles and combines them. Subsequently, it merges this comprehensive mortality data with the full sub-period analytic dataset (`dat.full.2`) using the `id` variable.

```r
mort2000 <- readRDS(file = "data/Mortalitydata/mort2000.RData")
mort2001 <- readRDS(file = "data/Mortalitydata/mort2001.RData")
mort2003 <- readRDS(file = "data/Mortalitydata/mort2003.RData")
mort2005 <- readRDS(file = "data/Mortalitydata/mort2005.RData")
mort2007 <- readRDS(file = "data/Mortalitydata/mort2007.RData")
mort2009 <- readRDS(file = "data/Mortalitydata/mort2009.RData")
mort2011 <- readRDS(file = "data/Mortalitydata/mort2011.RData")
mort2013 <- readRDS(file = "data/Mortalitydata/mort2013.RData")
```

```
mort2015 <- readRDS(file = "data/Mortalitydata/mort2015.RData")
mort2017 <- readRDS(file = "data/Mortalitydata/mort2017.RData")

dat.mortality.2 <- rbind(mort2000, mort2001, mort2003,
                         mort2005, mort2007, mort2009,
                         mort2011, mort2013, mort2015,
                         mort2017)

dat.full.with.mortality.2 <- merge(dat.full.2, dat.mortality.2,
                                   by = "id", all.x = TRUE)
dim(dat.full.with.mortality.2)
#> [1] 39156    20
```

The following code applies recoding which is also similar to main analysis.

```
dat.full.with.mortality.2$exposure.cat <- car::recode(dat.full.with.mortality.2$smoking.ag
                                                      0 = 'Never smoked';
                                                      1:9 = 'Started before 10';
                                                      10:14 = 'Started at 10-14';
                                                      15:17 = 'Started at 15-17';
                                                      18:20 = 'Started at 18-20';
                                                      21:80 = 'Started after 20';
                                                      else = NA ",
                                                      as.factor = TRUE)
dat.full.with.mortality.2$exposure.cat <- factor(dat.full.with.mortality.2$exposure.cat,
                                                 levels = c("Never smoked",
                                                            'Started before 10',
                                                            'Started at 10-14',
                                                            "Started at 15-17",
                                                            "Started at 18-20",
                                                            "Started after 20"))


# Survival time - Person-Months of Follow-up from NHANES Interview date
# changed to "Number of Person Months of Follow-up from NHANES Mobile Examination Center (
dat.full.with.mortality.2$stime.since.interview <- dat.full.with.mortality.2$mort_permth_i
summary(dat.full.with.mortality.2$stime.since.interview)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>    1.00   34.00   58.00   58.67   82.00  113.00   15424

# Age in month
```

```r
dat.full.with.mortality.2$age.month <- dat.full.with.mortality.2$age * 12
summary(dat.full.with.mortality.2$age.month)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
#>     0.0   120.0   336.0   386.9   648.0   960.0

# Survival time - Person-Months of Follow-up from birth = birth to screening + screening t
dat.full.with.mortality.2$stime.since.birth <- with(dat.full.with.mortality.2, age.month +
# converted back to year so that KM plot is in years / HR results remain the same (added)
dat.full.with.mortality.2$stime.since.birth <- dat.full.with.mortality.2$stime.since.birth
dat.full.with.mortality.2$stime.since.birth[is.na(dat.full.with.mortality.2$stime.since.in
summary(dat.full.with.mortality.2$stime.since.birth)
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>   19.08   37.17   53.17   53.08   68.17   89.00   15424

# All-cause mortality status
dat.full.with.mortality.2$status_all <- dat.full.with.mortality.2$mort_stat
table(dat.full.with.mortality.2$status_all, useNA = "always")
#>
#>     0     1  <NA>
#> 22216  1516 15424

# Cause-specific mortality
dat.full.with.mortality.2$mort_diabetes <- NULL
dat.full.with.mortality.2$mort_hyperten <- NULL
dat.full.with.mortality.2$mort_permth_int <- NULL
dat.full.with.mortality.2$mort_permth_exm <- NULL
dat.full.with.mortality.2$mort_ucod_leading <- NULL
dat.full.with.mortality.2$mort_eligstat <- NULL
```

- R Code Chunk 6: Variable Construction and Final Analytic Dataset for Sub-period Analysis

The same variables are constructed as the main analysis, but note, only for the 4 cycles. Below, the year variable is recoded into categorical levels. Then, the final data set is saved.

```r
# Recode year variable
dat.full.with.mortality.2$year.cat <- dat.full.with.mortality.2$year

new_levels <- c("2011-2012", "2013-2014", "2015-2016", "2017-2018")
levels(dat.full.with.mortality.2$year.cat) <- new_levels
```

```r
# Display the unique levels of year.cat
levels(dat.full.with.mortality.2$year.cat)
#> [1] "2011-2012" "2013-2014" "2015-2016" "2017-2018"
```

```r
# Remove variable, not used
dat.full.with.mortality.2$smoked.while.child <- NULL

### Analytic dataset - age 20 - 79
dat.analytic.2 <- subset(dat.full.with.mortality.2, age>=20 & age < 80)
dim(dat.full.with.mortality.2)[1] - dim(dat.analytic.2)[1] # added
#> [1] 18057

# Drop variables that are not being used
dat.analytic.2$born <- NULL
dat.analytic.2$age <- NULL
dat.analytic.2$age.month <- NULL

# No missing exposure or outcome
dim(dat.analytic.2)
#> [1] 21099    16
dat.analytic1 <- dat.analytic.2[complete.cases(dat.analytic.2$stime.since.birth),]
dim(dat.analytic1)
#> [1] 21011    16
dat.analytic2 <- dat.analytic1[complete.cases(dat.analytic1$exposure.cat),]
dim(dat.analytic2)
#> [1] 20950    16
profile_missing(dat.analytic2)
#>                   feature num_missing pct_missing
#> 1                      id           0           0
#> 2                     sex           0           0
#> 3                    race           0           0
#> 4                   race2           0           0
#> 5              smoking.age           0           0
#> 6                  smoking           0           0
#> 7                      psu           0           0
#> 8                  strata           0           0
#> 9                    year           0           0
#> 10     survey.weight.new           0           0
#> 11             mort_stat           0           0
#> 12          exposure.cat           0           0
#> 13 stime.since.interview           0           0
```

```
#> 14     stime.since.birth           0           0
#> 15           status_all           0           0
#> 16             year.cat           0           0

# No missing values in covariates - complete case dataset
dat.complete.2 <- na.omit(dat.analytic2)
dim(dat.complete.2)
#> [1] 20950    16

# Participants dropped - overall
nrow(dat.full.with.mortality.2) - nrow(dat.complete.2)
#> [1] 18206

# Participants dropped - due to missing exposure or outcome
nrow(dat.analytic.2) - nrow(dat.analytic2)
#> [1] 149

# Participants dropped - due to missing covariates
nrow(dat.analytic2) - nrow(dat.complete.2)
#> [1] 0

# added
table(dat.analytic.2$exposure.cat, useNA = "always")
#>
#>      Never smoked Started before 10  Started at 10-14  Started at 15-17
#>             12437               126              1497              2769
#>  Started at 18-20  Started after 20              <NA>
#>              2440              1767                63
```

Save the final analytic data for sensitivity analysis used below.

```
save(dat.complete.2, file = "data/dat.complete.2.RData")
```

12.2.2 Analysis of the 2011-2018 Sub-period

- R Code Chunk 7: Analysis

The following code sets up the survey design object (`w.design.2.0`) specifically for the 2011-2018 sub-period. It uses the `psu`, `strata`, and `survey.weight.new` variables to

correctly account for the complex sampling methodology within this subset of the data. This `w.design.2.0` object is essential for running survey-weighted models in this sensitivity analysis.

The following code sets up the survey design object for the 2011-2018 sub-period as in the main analysis.

```
# Set up the design
dat.full.with.mortality.2$miss <- 1
dat.full.with.mortality.2$miss[dat.full.with.mortality.2$id %in% dat.analytic2$id] <- 0
table(dat.full.with.mortality.2$miss)
#>
#>     0     1
#> 20950 18206

w.design.2 <- svydesign(ids = ~psu, strata = ~strata, weights = ~survey.weight.new,
                        data = dat.full.with.mortality.2, nest = TRUE)

# Subset the design
w.design.2.0 <- subset(w.design.2, miss == 0 & survey.weight.new > 0)
```

- R Code Chunk 8: Descriptive Statistics for Sub-period (**Appendix Table 3**)

The following table reproduces *Appendix Table 3* from the paper. It summarizes the demographic characteristics of the cohort from the 2011-2018 NHANES cycles, stratified by smoking initiation categories, to provide context for this sensitivity analysis.

```
# Create the weighted table using the survey design object for the 2011-2018 data
tab_app_3_weighted <- svyCreateTableOne(vars = c("race2"),
                                        strata = "exposure.cat",
                                        data = w.design.2.0,
                                        addOverall = TRUE,
                                        test = TRUE)

# Print and format the table
tab_app_3_printed <- print(tab_app_3_weighted,
                           format = "p",
                           catDigits = 2,
                           showAllLevels = TRUE,
                           smd = TRUE)
```

```r
# Define the desired column order
new_order_t3 <- c("level", "Never smoked", "Started before 10",
                  "Started at 10-14", "Started at 15-17",
                  "Started at 18-20", "Started after 20",
                  "Overall", "p", "test", "SMD")

# Apply the new order to the table object
tab_app_3 <- tab_app_3_printed[, new_order_t3]
```

```r
# Display the formatted table using kable for a clean Quarto output
kable(tab_app_3, caption = "Appendix Table 3: Characteristics
      of the study sample for the 2011-2018 NHANES cycles,
      stratified by smoking initiation age.") %>%
  kable_styling(bootstrap_options = "striped", full_width = FALSE)
```

Table 12.2: Appendix Table 3: Characteristics of the study sample for th
cycles, stratified by smoking initiation age.

	level	Never smoked	Started before 10	Started at 10-14	Started at 15-17	Started at 18
n		128386920.10	1240174.76	16008335.09	32086914.62	26126847.18
race2 (%)	White	59.51	74.22	73.86	73.66	68.27
	Black	12.39	5.51	7.49	8.55	11.03
	Hispanic	17.72	11.12	12.94	11.46	12.77
	Asian	7.62	1.33	1.31	2.14	3.40
	Others	2.76	7.82	4.40	4.19	4.53

Sub-Period Specific Analyses This subsection presents the Kaplan-Meier curve, unadjusted and adjusted Cox models for the 2011-2018 sub-period data, providing initial insights before delving into effect modification.

- R Code Chunk 9 : Kaplan-Meier Curves and Log-Rank Test for Sub-period

The following code visualizes the survival probabilities for the 2011-2018 sub-period data using Kaplan-Meier curves, stratified by `exposure.cat`. It also performs a survey-weighted log-rank test to assess statistical differences between the survival curves in this specific subset of the data. The process is the same as the main analysis.

Kaplan-Meier Survival Curves

```r
dummy_sub <- length(unique(as.factor(w.design.2.0$variables$exposure.cat)))

# Define the survival formula for Kaplan-Meier plot and log-rank test
formulax0_sub <- as.formula(Surv(stime.since.birth, status_all) ~ exposure.cat)

# Calculate survey-weighted Kaplan-Meier curves for the sub-period
sA_sub<-svykm(formulax0_sub, design=w.design.2.0)

# Plot
par(oma=rep(0,4))
par(mar=c(5,4,0,0) + 0.1)
plot(sA_sub, pars=list(col=c(1:dummy_sub)),
     xlab = "Time",
     ylab = "Proportion surviving")
legend("bottomleft", levels(as.factor(w.design.2.0$variables$exposure.cat)),
       col = (1:dummy_sub), lty = c(1,1))
```

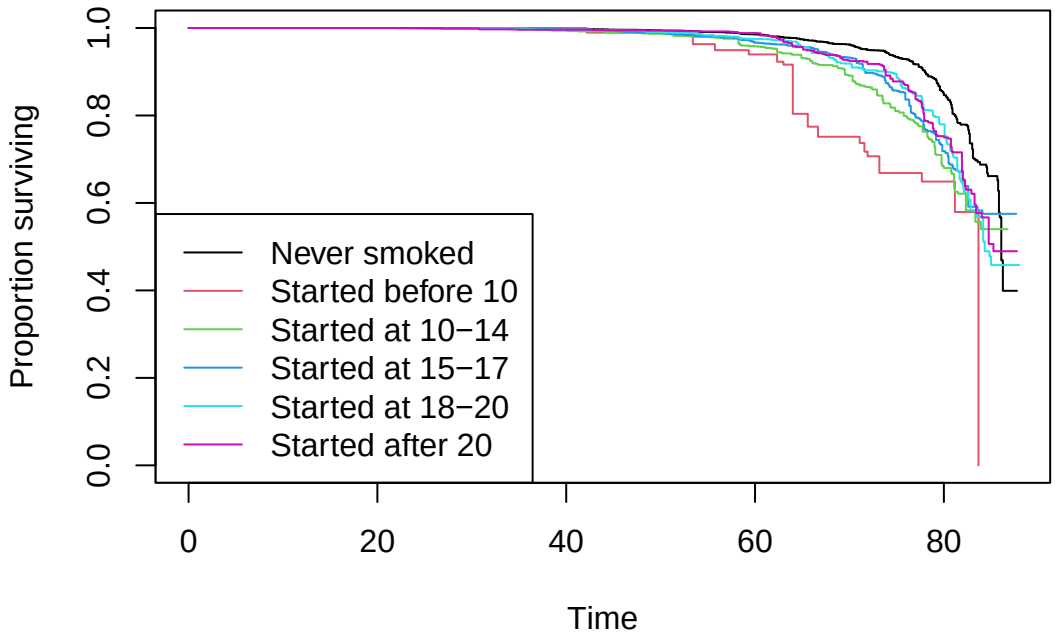


Figure 12.4: Sensitivity Analysis 2: Survey-Weighted Kaplan-Meier Curves

Log-Rank Test

```r
# Perform the survey-weighted log-rank test
lrt_sub <- svylogrank(formulax0_sub,design=subset(w.design.2.0, survey.weight.new>0))
round(lrt_sub[[2]],2)
#> Chisq     p
#> 40.71  0.00
```

Cox Proportional Hazards Models

- R Code Chunk 10 : Unadjusted Cox Proportional Hazards Models for Sub-period

The following code runs both unweighted (`coxph`) and survey-weighted (`svycoxph`) unadjusted Cox proportional hazards models on the sub-period data. These models serve as a baseline for comparison within this specific sensitivity analysis.

```r
# Unweighted Cox model for sub-period
fit0 <- coxph(Surv(stime.since.birth, status_all) ~ exposure.cat,
              data = dat.complete.2)

# Survey-weighted unadjusted Cox model for sub-period
fit0 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat,
                 design = w.design.2.0) # Using w.design.2.0
f0 <- publish(fit0)

# Process unadjusted HRs for sub-period (similar to main analysis f0r)
f0r <- f0$regressionTable[2:6,c("HazardRatio","CI.95")]
f0r$group <- "Crude (Sub-period)"
ci_f0_sub <- str_extract_all(f0r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f0r$CI.l <- as.numeric(as.character(ci_f0_sub[,1]))
f0r$CI.u <- as.numeric(as.character(ci_f0_sub[,2]))
names(f0r) <- c("mean","CI.95","group","lower","upper")
f0r
```

- R Code Chunk 11: Adjusted Cox Proportional Hazards Model for Sub-period

The following code fits the adjusted Cox proportional hazards model for the sub-period data, controlling for `sex`, `race2`, and `year.cat`.

```r
# Survey-weighted adjusted Cox model for sub-period
fit1 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat + sex + race2 + year.c
                 design = w.design.2.0)
f1 <- publish(fit1)

# Process adjusted HRs for sub-period (similar to main analysis f1r)
f1r <- f1$regressionTable[2:6,c("HazardRatio","CI.95")]
f1r$group <- "Adjusted (Sub-period)"
ci_f1_sub <- str_extract_all(f1r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f1r$CI.l <- as.numeric(as.character(ci_f1_sub[,1]))
f1r$CI.u <- as.numeric(as.character(ci_f1_sub[,2]))
names(f1r) <- c("mean","CI.95","group","lower","upper")
f1r
```

- R Code Chunk 12: Effect Modification by Race/Ethnicity in Sub-period (Model)

The following code fits the core model for the sub-period sensitivity analysis: a survey-weighted Cox proportional hazards model with an interaction term between `exposure.cat` and `race2`.

```r
# Model
fit2 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat*race2 + sex  + year.ca
                           design = w.design.2.0) # Using w.design.2.0
f2 <- publish(fit2)

# Process Race-Specific Hazard Ratios for Sub-period Plotting
f2rW <- f2$regressionTable[31:35,c("HazardRatio","CI.95")]
f2rB <- f2$regressionTable[36:40,c("HazardRatio","CI.95")]
f2rH <- f2$regressionTable[41:45,c("HazardRatio","CI.95")]
f2rA <- f2$regressionTable[46:50,c("HazardRatio","CI.95")]
# Assign group labels
f2rW$group <- "White (Sub-period)"
f2rB$group <- "Black (Sub-period)"
f2rH$group <- "Hispanic (Sub-period)"
f2rA$group <- "Asian (Sub-period)"
f2r <- rbind(f2rW, f2rB, f2rH, f2rA)
ci <- str_extract_all(f2r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f2r$CI.l <- as.numeric(as.character(ci[,1]))
f2r$CI.u <- as.numeric(as.character(ci[,2]))
names(f2r) <- c("mean","CI.95","group","lower","upper")
f2r
```

- R Code Chunk 13: Effect Modification by Sex in Sub-period (Model)

The following code fits a survey-weighted Cox proportional hazards model, including an interaction term between `exposure.cat` and `sex` for the sub-period data.

```
fit3 <- svycoxph(Surv(stime.since.birth, status_all) ~ exposure.cat*sex + race2 + year.cat
                 design = w.design.2.0)
f3 <- publish(fit3)

# Process Sex-Specific Hazard Ratios for Sub-period Plotting
f3rM <- f3$regressionTable[16:20,c("HazardRatio","CI.95")]
f3rF <- f3$regressionTable[21:25,c("HazardRatio","CI.95")]

f3rM$group <- "Male (Sub-period)"
f3rF$group <- "Female (Sub-period)"
f3r <- rbind(f3rM, f3rF)
ci <- str_extract_all(f3r[,2], '\\d+([.,]\\d+)?', simplify = TRUE)
f3r$CI.l <- as.numeric(as.character(ci[,1]))
f3r$CI.u <- as.numeric(as.character(ci[,2]))
names(f3r) <- c("mean","CI.95","group","lower","upper")
```

12.2.3 Visualizing the Sub-period Results

- R Code Chunk 14: Plot

The following code creates the plot using `ggplot2` and corresponds to *Appendix Figure 4* in section C.1, Sensitivity analysis with 'Asian' category, part of the supplementary materials published with the original paper.

```
fr <- rbind(f2r)
fr <- as.data.frame(fr)
fr[,c(1,4,5)] <- sapply(fr[,c(1,4,5)], as.numeric)
fr$age.grp <- c("Started before 10", "Started at 10-14", "Started at 15-17",
                "Started at 18-20", "Started after 20")
# Change levels
fr$group <- fct_recode(fr$group,
  "Non-Hispanic White" = "White (Sub-period)",
  "Non-Hispanic Black" = "Black (Sub-period)",
```

```
  "Hispanic"           = "Hispanic (Sub-period)",
  "Asian"              = "Asian (Sub-period)"
)

# Re-order
fr$age.grp <- factor(fr$age.grp,
                     levels = c("Started before 10",
                                "Started at 10-14",
                                "Started at 15-17",
                                "Started at 18-20",
                                "Started after 20"))

fr$group <- factor(fr$group, levels = c("Asian", "Hispanic",
                                        "Non-Hispanic Black", "Non-Hispanic White"))

# Plot
ggplot(fr, aes(x = mean, y = group, colour = age.grp)) +
  geom_vline(xintercept = 1, linetype = "dashed", color = "grey") +
  geom_errorbar(aes(xmax = lower, xmin = upper), position = "dodge") +
  geom_point(position = position_dodge(0.9)) +
  labs(x = "Hazard Ratio", y = "",
       legend=TRUE, col = "Age group") +
  theme_classic() +
  theme(panel.grid.major.x = element_blank(),
        panel.border = element_blank(),
        legend.title=element_text(size=12),
        legend.text=element_text(size=12),
        plot.title = element_text(hjust = 0),
        axis.text.x = element_text(size = 10, face = "bold"),
        axis.text.y = element_text(size = 10, face = "bold"),
        legend.position=c(.8, .8))
```

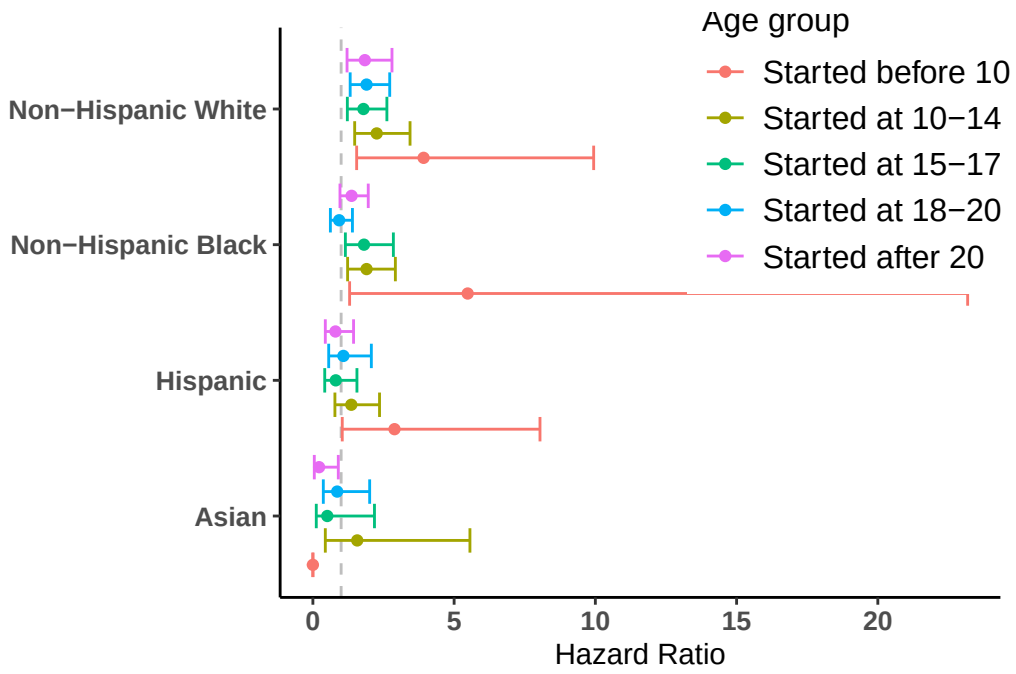


Figure 12.5: Appendix Figure 4: Forest plot of Hazard Ratios from the sensitivity analysis with 'Asian' category.

```
ggsave("images/forest2.png")
```

This concludes the sensitivity analysis.

12.3 Appendix B: Exploratory Analysis of Smoking Duration

As noted in the main paper, the duration of smoking is a potential mediator in the pathway from early smoking initiation to mortality. To investigate this secondary relationship, we create boxplots to visualize smoking duration across the different age-of-initiation categories. The following analysis uses the data we set aside earlier (`dat.analytic.duration.analysis`). This analysis reproduces Appendix Figures 1, 2, and 3 from the paper.

NOTE: This analysis is based on participants from 1999-2016, as the variable for age of smoking cessation (`SMD055`) was not collected in the 2017-2018 cycle.

12.3.1 Data Preparation for Duration Plots

- R Code Chunk 1: Prepare Data for Duration Analysis

The following code processes the data to calculate smoking duration by cleaning the age variables, computing the difference, correcting for data entry errors, and saving the final, plot-ready dataset.

```
# Clean the 'smoking.age' variable before calculation.
# Recode 0 and 99 (codes for never-smokers) to NA.
dat.analytic.duration.analysis$smoking.age[dat.analytic.duration.analysis$smoking.age == 0
dat.analytic.duration.analysis$smoking.age[dat.analytic.duration.analysis$smoking.age == 9

# Calculate smoking duration.
dat.analytic.duration.analysis$smoking.duration <- dat.analytic.duration.analysis$smoking.

# Correct any negative durations (from data entry errors) to 0.
dat.analytic.duration.analysis$smoking.duration[!is.na(dat.analytic.duration.analysis$smok

# Run the original summary checks to explore the new variable.
cat("Total number of participants with non-missing smoking duration data:\n")
#> Total number of participants with non-missing smoking duration data:
print(sum(!is.na(dat.analytic.duration.analysis$smoking.duration)))
#> [1] 8916

cat("\nSummary of smoking duration for all ever-smokers:\n")
#>
#> Summary of smoking duration for all ever-smokers:
print(summary(dat.analytic.duration.analysis$smoking.duration))
#>    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.    NA's
#>    0.00    8.00   18.00   19.99   30.00   68.00   41908

# Create the final data frame for plotting by filtering out non-smokers.
dat.analytic.plot <- dat.analytic.duration.analysis %>%
  filter(!exposure.cat %in% c("Never smoked", NA))

# Save the final plot-ready data.
saveRDS(dat.analytic.plot, file = "data/duration_plot_data.rds")
```

12.3.2 Visualizing Smoking Duration by Initiation Age

- R Code Chunk 2: Smoking Duration Plots

Now, the following code uses the `dat.analytic.plot` data version to calculate the median smoking duration for each exposure category, which will be used to draw a trend line on the plots.

```
# Pre-calculate medians for the trend line
median_data <- dat.analytic.plot %>%
  group_by(exposure.cat) %>%
  summarize(median = median(smoking.duration, na.rm = TRUE)) %>%
  arrange(factor(exposure.cat, levels = c('Started before 10',
                                          'Started at 10-14',
                                          "Started at 15-17",
                                          "Started at 18-20",
                                          "Started after 20")))
```

The following code chunks generate the three plots corresponding to **Appendix Figures 1, 2, and 3** in the paper's supplementary material.

Appendix Figure 1: Overall Smoking Duration

```
ggplot(dat.analytic.plot, aes(x = exposure.cat, y = smoking.duration, group = exposure.cat
  geom_boxplot(fill = "white") +
  geom_line(data = median_data, aes(x = exposure.cat, y = median, group = 1), color = "gre
  geom_smooth(method = "loess", se = TRUE, color = "blue", size = 1, span = 0.5, level = 0
  theme_minimal() +
  theme(
    legend.position = "none",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  labs(x = "", y = "Smoking Duration (Years)",
       title = "Boxplot of Smoking Duration by Exposure Category")
```

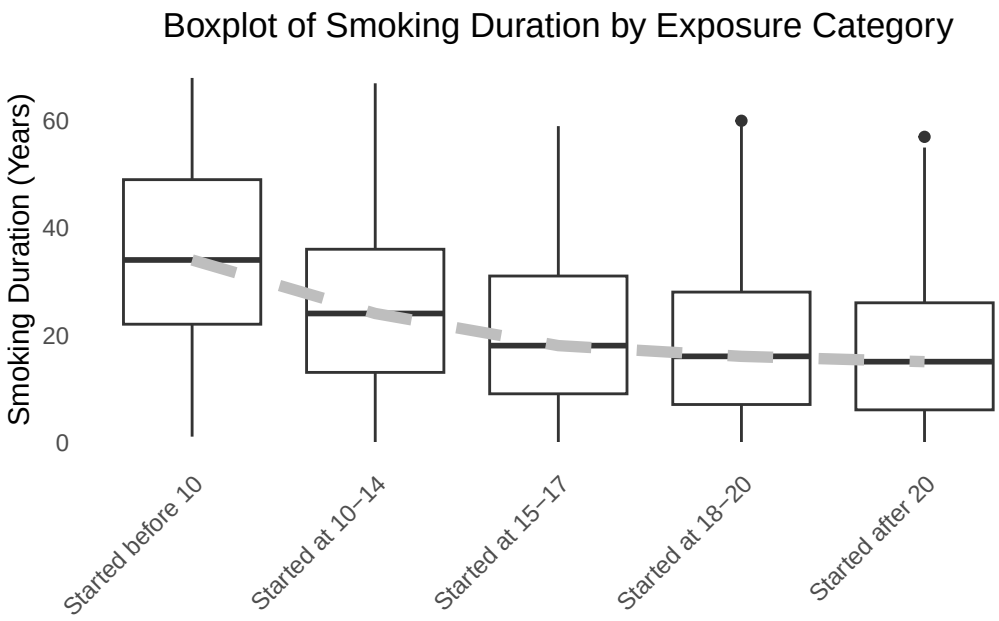
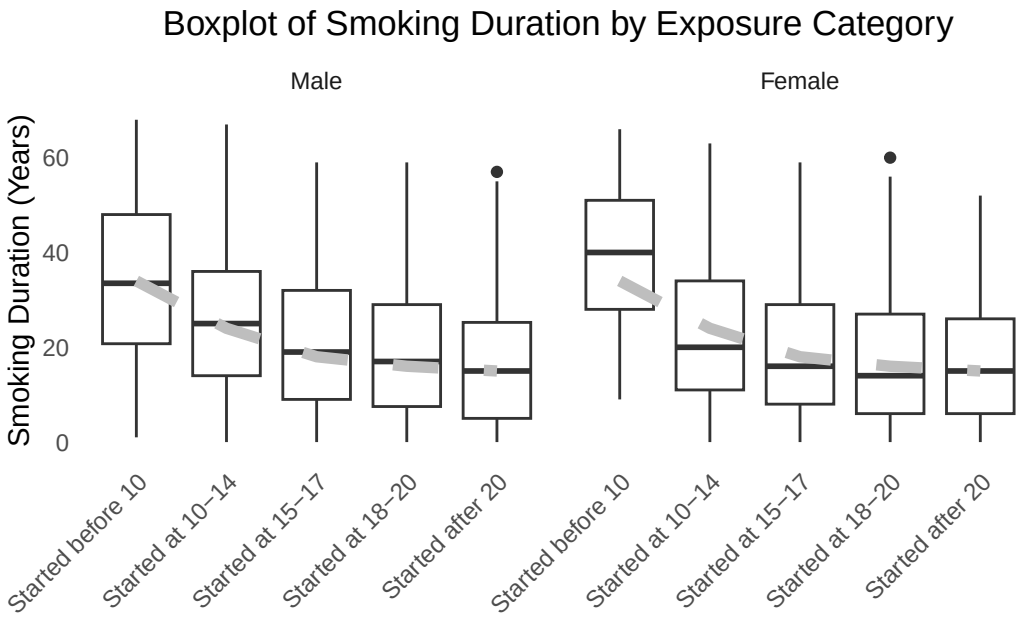


Figure 12.6: Smoking duration by age of initiation category among ever-smokers.

Appendix Figure 2: Stratified by Sex

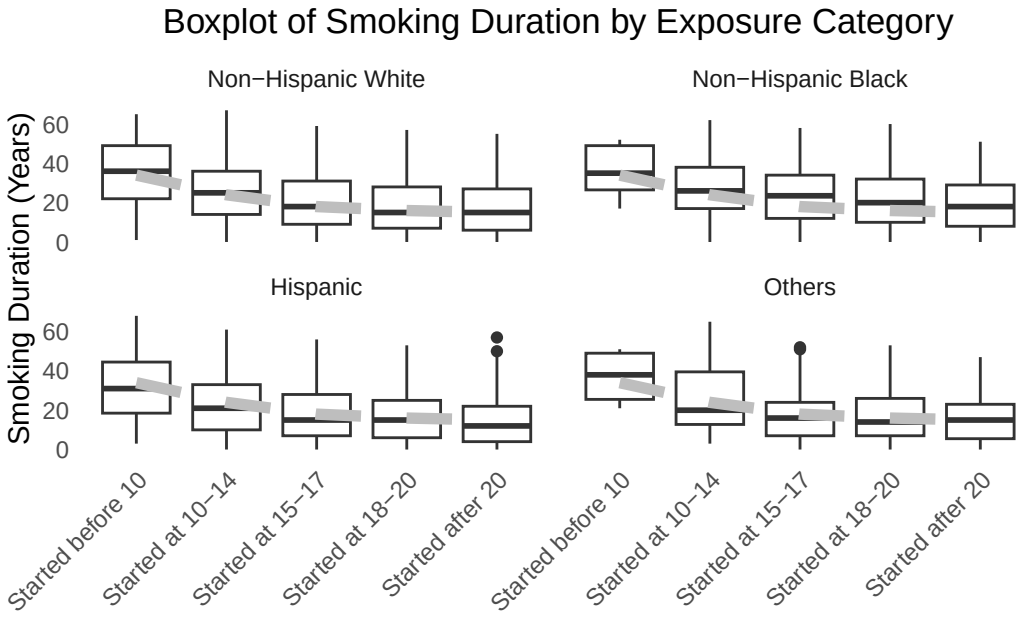
```
ggplot(dat.analytic.plot, aes(x = exposure.cat, y = smoking.duration, group = exposure.cat
  geom_boxplot(fill = "white") +
  geom_line(data = median_data, aes(x = exposure.cat, y = median, group = 1), color = "gre
  geom_smooth(method = "loess", se = TRUE, color = "blue", size = 1, span = 0.5, level = 0
  facet_wrap(~sex) +
  theme_minimal() +
  theme(
    legend.position = "none",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  labs(x = "", y = "Smoking Duration (Years)",
       title = "Boxplot of Smoking Duration by Exposure Category")
```



Appendix Figure 3: Stratified by Race/Ethnicity

```
levels(dat.analytic.plot$race)[levels(dat.analytic.plot$race) == "White"] <- "Non-Hispanic
levels(dat.analytic.plot$race)[levels(dat.analytic.plot$race) == "Black"] <- "Non-Hispanic

ggplot(dat.analytic.plot, aes(x = exposure.cat, y = smoking.duration, group = exposure.cat
  geom_boxplot(fill = "white") +
  geom_line(data = median_data, aes(x = exposure.cat, y = median, group = 1), color = "gre
  geom_smooth(method = "loess", se = TRUE, color = "blue", size = 1, span = 0.5, level = 0
  facet_wrap(~race) +
  theme_minimal() +
  theme(
    legend.position = "none",
    panel.grid.major = element_blank(),
    panel.grid.minor = element_blank(),
    plot.title = element_text(hjust = 0.5),
    axis.text.x = element_text(angle = 45, hjust = 1)
  ) +
  labs(x = "", y = "Smoking Duration (Years)",
       title = "Boxplot of Smoking Duration by Exposure Category")
```



12.3.3 Interpretation of Results

The boxplots consistently show a clear inverse relationship. The median smoking duration (indicated by the grey dashed line) is highest for those who started smoking before age 10 and steadily decreases for groups that started later in life. This trend holds true when the data is stratified by both sex and race/ethnicity. This visual evidence supports the hypothesis that **earlier smoking initiation leads to a longer cumulative exposure to tobacco**, which is a likely mediator in the pathway to increased mortality risk.

12.4 Chapter Summary and Next Steps

We have now completed two important sensitivity analyses to test the robustness of our primary findings. First, we adjusted for socioeconomic status proxies, and second, we re-ran the effect modification analysis on the 2011-2018 data to include the non-Hispanic Asian category. The results from both analyses were largely consistent with the main findings, strengthening our confidence in the conclusions.

We also completed the exploratory analysis which suggested a potential mechanism for our findings, showing that earlier smoking initiation is strongly associated with a longer lifetime smoking duration.

With the analytical portion complete, the next chapter, "Discussion of Results," will synthesize the findings from all analyses and interpret their meaning.

Part IV

Discussion

13 Discussion of Results

The findings from this project consistently demonstrate a strong, dose-response relationship between the age of smoking initiation and all-cause mortality. Across all analyses, a clear pattern emerged: the earlier an individual starts smoking, the higher their risk of premature death. The main adjusted model found that individuals who started smoking before age 10 had a hazard ratio of **2.71** (95% CI: 2.05, 3.58) compared to never smokers, quantifying the profound danger of very early initiation.

This chapter synthesizes the evidence from the four key analyses performed in this walkthrough to build a comprehensive picture of these findings.

13.1 Synthesizing the Evidence

13.1.1 The Main Finding: A Clear Dose-Response Relationship

The **Main Survival and Effect Modification Analysis** formed the core of this investigation. Using data from all 10 NHANES cycles (1999–2018), it established that the risk of mortality incrementally increased with earlier ages of smoking initiation. Furthermore, this analysis revealed that the association was not uniform across all demographic groups. While the interaction by race/ethnicity was not statistically significant, the interaction by **sex was significant ($p = 0.001$)**, with females showing slightly higher hazard ratios than males across most initiation categories.

13.1.2 Supporting Evidence: The Role of Smoking Duration

The **Exploratory Analysis of Smoking Duration** (1999–2016) provided important context for our main finding. It revealed a clear trend where an earlier age of initiation was associated with a longer total duration of smoking. This suggests a potential mediating pathway: starting to smoke earlier leads to a longer cumulative exposure to tobacco, which in turn increases the risk of mortality.

13.1.3 Confirming Robustness: The Sensitivity Analyses

Two sensitivity analyses were conducted to challenge our primary findings and ensure their robustness:

1. **Adjusting for SES Proxies**: This analysis confirmed that the association between smoking initiation and mortality held even after adjusting for family income and education level. This strengthens our conclusion that the observed effect is not simply a result of socioeconomic confounding.
2. **Including the Non-Hispanic Asian Category (2011–2018)**: By focusing on a more recent subset of data, we were able to conduct a more nuanced effect modification analysis. While the smaller sample size resulted in wider confidence intervals, the overall trends remained consistent with the main analysis.

13.2 Public Health Implications

The consistency of these findings across multiple analyses underscores a critical public health message: **preventing smoking initiation among youth is one of the most effective strategies to reduce premature mortality.** The particularly high risk for those who start before age 10 highlights the need for interventions targeted at children and adolescents. The significant differences observed by sex also suggest that prevention and cessation programs may need to be tailored to address gender-specific factors.

13.3 Chapter Summary and Next Steps

This chapter synthesized the key findings from the main analysis and the subsequent sensitivity and exploratory analyses. We confirmed a consistent and strong dose-response relationship between earlier smoking initiation and higher all-cause mortality, with this effect being significantly modified by sex.

While the results are compelling, it is crucial to acknowledge the study's constraints. The next chapter, **"Limitations and Future Directions,"** will discuss the methodological limitations of the analysis and suggest avenues for future research.

14 Limitations and Future Directions

While this analysis provides robust, nationally representative findings, it's essential to acknowledge its limitations and consider avenues for future research. This chapter discusses the primary challenges encountered and proposes next steps to build upon this work.

14.1 Methodological Limitations

14.1.1 Data Harmonization Across Cycles

A significant challenge in this project was combining data across ten different NHANES cycles (1999–2018). This required considerable effort to harmonize variables, as names and coding schemes frequently changed over time. For instance:

- **Inconsistent Variable Names**: The variable for household head education was named `DMDHREDU` in early cycles but changed to `DMDHREDZ` in the 2017–2018 cycle, requiring conditional logic to process correctly.
- **Evolving Definitions**: The definition of race/ethnicity evolved, with a distinct variable for the non-Hispanic Asian population (`RIDRETH3`) only becoming available from 2011 onwards. Our main analysis had to group this population into an "Others" category to maintain consistency, which may mask effects specific to this group. This was the primary motivation for the second sensitivity analysis.

14.1.2 Unmeasured Confounding and Missing Data

While we adjusted for key demographic variables, the main analysis does not account for all potential confounders, such as detailed family medical history, genetic predispositions, or granular socioeconomic status (SES).

- Our sensitivity analysis that included SES proxies (`pir` and `HHedu`) provided some reassurance, but it came at the cost of a significantly reduced sample size due to missing data. This highlights the classic trade-off between confounding control and statistical power.

14.1.3 Reliance on Self-Reported Data

The primary exposure—age of smoking initiation—is based on self-report and may be subject to recall bias, where participants may not accurately remember when they started smoking regularly.

14.2 Future Directions

The limitations of this study highlight several exciting avenues for future research:

14.2.1 Investigating Cause-Specific Mortality

This analysis focused on all-cause mortality. A valuable next step would be to examine cause-specific mortality (e.g., from cardiovascular disease, cancer, or respiratory illness). This could reveal more specific pathways through which early smoking initiation impacts long-term health.

14.2.2 Modeling Time-Dependent Smoking Behavior

As noted in the original paper, smoking behavior is dynamic. People's smoking habits change over their lifetime. Future research could employ more advanced statistical models (e.g., marginal structural models) to incorporate time-dependent variables like:

- **Smoking intensity** (cigarettes per day)
- **Periods of cessation and relapse**
- **Use of other tobacco products**

This would provide a more nuanced understanding of smoking's cumulative impact, though it would require more detailed longitudinal data than is currently available in the public-use NHANES files.

14.3 Chapter Summary and Next Steps

In this chapter, we have critically examined the limitations of our analysis, from the practical challenges of data harmonization to the methodological considerations of unmeasured confounding. We also proposed several exciting avenues for future research that could build upon this work, such as exploring cause-specific mortality and dynamic smoking behaviors.

This concludes the main body of the walkthrough. The final sections of this book include the **Appendices,** which provide a glossary of key terms, and the **References.**

15 Appendices

This section contains supplementary materials, including a glossary of key terms and additional exploratory analyses referenced in the main text.

15.1 Appendix A: Glossary of Key Terms

- **Hazard Ratio (HR)**: A measure of effect in survival analysis. It represents the ratio of the hazard rates corresponding to the conditions described by two levels of an explanatory variable. An HR of 2, for example, means that at any given time, the group of interest is twice as likely to experience the event (e.g., death) as the comparison group. An HR of 1 indicates no difference in risk.
- **Kaplan-Meier Curve**: A non-parametric statistic used to estimate the survival function from lifetime data. The curve is a series of horizontal steps of declining magnitude which, when taken together, approximates the true survival function for that population. It allows for a visual comparison of survival probabilities between different groups over time.
- **Primary Sampling Units (PSUs)**: The first stage of sampling in a multi-stage survey design like NHANES. PSUs are typically large geographic areas (like counties). The survey first samples a set of PSUs, and then samples smaller units within them. Accounting for PSUs is critical for accurate variance estimation.
- **Relative Excess Risk due to Interaction (RERI)**: A measure used to assess additive interaction between two exposures. RERI > 0 indicates a positive interaction (synergism), meaning the combined effect of the two exposures is greater than the sum of their individual effects. RERI < 0 indicates a negative interaction (antagonism).

15.2 Chapter Summary

This appendix provided supplementary materials to support the analyses. It begins with a glossary of key statistical and epidemiological terms used throughout the book to aid the reader's understanding.

References