Big Data for Epidemiology: Applied Data Analysis Using National Health Surveys

BIG DATA FOR EPIDEMIOLOGY: APPLIED DATA ANALYSIS USING NATIONAL HEALTH SURVEYS

TIFFANY B. KINDRATT

PEACE OSSOM-WILLIAMSON

Mavs Open Press

Arlington

CONTENTS

ABOUT THE PUBLISHER

MAVS OPEN PRESS

ABOUT MAVS OPEN PRESS

Creation of this resource was supported by Mavs Open Press, operated by the University of Texas at Arlington Libraries (UTA Libraries). Mavs Open Press offers no-cost services for UTA faculty, staff, and students who wish to openly publish their scholarship. The Libraries' program provides human and technological resources that empower our communities to publish new open access journals, to convert traditional print journals to open access publications, and to create or adapt open educational resources (OER). Our resources are openly licensed using Creative Commons licenses and are offered in various e-book formats free of charge, which can be downloaded from the Mavs Open Press OER catalog. Optional print copies of this text may be available through the UTA Bookstore or can be purchased directly from XanEdu, Mavs Open Press' exclusive print provider and distributor.

ABOUT OER

OER are free teaching and learning materials that are licensed to allow for revision and reuse. They can be fully self-contained textbooks, videos, quizzes, learning modules, and more. OER are distinct from public resources in that they permit others to use, copy, distribute, modify, or reuse the content. The legal permission to modify and customize OER to meet the specific learning objectives of a particular course make them a useful pedagogical tool.

ABOUT PRESSBOOKS

Pressbooks is an open source, web-based authoring tool based on WordPress, and it is the primary tool that Mavs Open Press uses to create and adapt course materials. Pressbooks should not be used with Internet Explorer. The following browsers are best to use with Pressbooks:

- Firefox
- Chrome
- Safari
- Edge

CONTACT US

Information about open education at UTA is available online. Contact us at oer@uta.edu for other inquires related to UTA Libraries publishing services.

ABOUT THIS PROJECT

OVERVIEW

National data sets provide an avenue for students to practice data analytic skills while also answering meaningful research questions. This open education resource was developed to train future public health professionals how to conduct secondary data analysis of national health surveys using SAS statistical software. SAS software was selected because it is one of the most commonly used software programs used among public health departments and academia. The book includes details on how to analyze public-use data from five common national health surveys, including the National Health Interview Survey (NHIS), Medical Expenditure Panel Survey (MEPS), Health Information National Trends Survey (HINTS), Behavior Risk Factor Surveillance System (BRFSS) and National Health and Nutrition and Examination Survey (NHANES). All datasets and corresponding syntax files are available from the Open ICPSR Data Repository (https://doi.org/10.3886/E172301V1). Future steps are to provide databases and syntax files for other analytic software, specifically STATA.

CREATION PROCESS

The creation of this textbook began in March 2020. Due to COVID-19 pandemic challenges on teaching and research in higher education, final production was delayed. Chapters 1-8 were piloted during Spring 2021 and Chapters 1-6 and 9-12 were piloted in Spring 2022 to Master of Public Health students enrolled in KINE 5386 Big Data for Epidemiology. Revisions were ongoing throughout the development process. Any corrections from the Spring 2021 pilot were made prior to the Spring 2022 pilot. Efforts are underway for additional reviews to be completed by a consultant, Peace Ossom-Williamson, MLA.

ABOUT THE AUTHOR

Tiffany B. Kindratt, PhD, MPH, is an assistant professor in the Public Health Program, Department of Kinesiology, College of Nursing and Health Innovation at the University of Texas at Arlington. She is Director of the Health Survey Research Laboratory and conducts research focused on predisposing (e.g. race/ethnicity, specifically Arab/Middle Eastern and North African) and enabling (e.g. patient-provider communication, patient experiences) factors that influence individuals' health behaviors, morbidity, mortality and use of health services across the life course using big data methodologies. She has an extensive background in epidemiologic and large database analysis, Arab/Middle Eastern and North African American health disparities, and training of medical learners. She has 13 years of experience analyzing large databases and complex surveys, including those included in this book. She currently has federal research funding from the National Institutes of Health (National Institute on Aging) and Health Resources and Services Administration and has over 50 manuscripts published in peer reviewed scientific journals.

ACKNOWLEDGMENTS

UTA CARES GRANT PROGRAM

Creation of this OER was funded by the UTA CARES Grant Program, which is sponsored by UTA Libraries. Under the auspices of UTA's Coalition for Alternative Resources in Education for Students (CARES), the grant program supports educators interested in practicing open education through the adoption of OER and, when no suitable open resource is available, through the creation of new OER or the adoption of library-licensed or other free content. Additionally, the program promotes innovation in teaching and learning through the exploration of open educational practices, such as collaborating with students to produce educational content of value to a wider community. Information about the grant program and funded projects is available online.

AUTHOR'S NOTE

The case studies developed in Chapter 6, National Health Interview Survey, and Chapter 9, Behavioral Risk Factor Surveillance System were supported by the National Institute On Aging of the National Institutes of Health under Award Numbers P30 AG059300 and R03AG070177. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health."

I would like to acknowledge, Dr. Erin Carlson, who is the Director of Graduate Public Health Programs at the University of Texas at Arlington. Dr. Carlson was integral in making my dream of creating this textbook in alignment with a Master of Public Health Big Data for Epidemiology course (KINE 5386) a reality. I would also like to acknowledge Dr. David Keller, Professor, Associate Dean, and Chair of Department of Kinesiology for his support of my research and teaching endeavors. I would like to acknowledge Dr. Bijal Balasubramanian, who served as my PhD advisor and mentor, who allowed me to create an independent study focused on analyzing national data at UTHealth, School of Public Health Regional Campus. Finally, I would like to acknowledge Dr. Florence J. Dallo, who served as my Master of Public Health advisor and mentor, who piqued my interest in analyzing national health data.

On a personal note, I would like to acknowledge my husband, Tony, for his love and support during the writing and publishing of this textbook. I would like to acknowledge my two sons, Cameron and Maxwell, one of whom was born during the writing of this textbook. Finally, I would like to acknowledge my parents, Diane and Gary Billmeier. This publication would not be possible without the love and support of them all.

LEAD AUTHOR

Tiffany B. Kindratt, PhD, MPH – Assistant Professor,

Public Health, Department of Kinesiology, College of Nursing and Health Innovation, University of Texas at Arlington.

CHAPTER 1.

INTRODUCTION

1.1 PURPOSE

The purpose of this textbook is to train future public health professionals, specifically Master of Public Health (MPH) students, how to conduct basic applied data analysis using secondary data collected from national health surveys. This textbook helps to eliminate gaps in knowledge, skills and analytical abilities that may prohibit MPH graduates from being successful in entry-level public health practice and research-focused positions. A recent study of local health departments demonstrated that entry-level public health professionals lacked the knowledge, skills and abilities for data collection, database management, data cleaning, quantitative data analysis/statistics, and data analysis using SAS statistical software.[1] Using publicly available data from national health surveys, this textbook will allow students to learn and practice data analytic skills with SAS statistical software to answer general surveillance and analytical research questions in preparation for their future public health practices.

1.2 HEALTH SERVICES RESEARCH FOCUS

The examples used in this textbook stem from previous studies and the current research laboratory focus of its primary author, Tiffany Kindratt, PhD, MPH. Established in Fall 2019, Dr. Kindratt's Health Survey Research (HSR) Lab is housed in the Public Health Program, Department of Kinesiology, College of Nursing and Health Innovation at the University of Texas at Arlington. The goal of the HSR lab is to conduct epidemiologic research studies focused on evaluating predisposing and enabling factors that influence individuals' health behaviors, morbidity, mortality and use of health services with big data methodologies. This includes the secondary analysis of large national health surveys that use complex samples, such as the Medical Expenditure Panel Survey (MEPS), National Health Interview Survey (NHIS), Health Information National Trends Survey (HINTS), American Community Survey (ACS), and others. Another goal of this lab is to collaborate with multidisciplinary teams and contribute to research studies designed to 1) train and mentor future public health and medical professionals and 2) implement community-based participatory research and quality improvement methodologies in community and clinical settings.

The HSR lab's research focus was developed to cover a wide range of epidemiology outcomes and contributing factors using Andersen's model of health services as its guiding framework.[2] Race, ethnicity, place of birth, and geographic context (urban or rural) disparities are evaluated to determine how individual predisposing factors contribute to health outcomes. Dr. Kindratt's research incorporates the examination of health disparities among Arab Americans, comprising of either born or tracing heritage to the Middle East or North Africa, who are largely underrepresented in health research because they are classified as non-Hispanic Whites by the United States (US) federal government.[3] The lab's research focus extends Andersen's model by incorporating patient experiences as contextual enabling factors of health services utilization and evaluating morbidity and mortality outcomes. Patient experiences that are examined include self-reports of qualities and modes of patient-provider communication, patient-provider gender and race concordance, care coordination, and provider satisfaction.

1.3 OUTLINE OF TEXTBOOK CHAPTERS

This textbook is separated into four sections, including: 1) introduction to national health surveys; 2) basic applied data analysis; 3) common national health surveys; and 4) dissemination and conclusions.

1.3.1 TEXTBOOK SECTION 1: INTRODUCTION TO NATIONAL HEALTH SURVEYS

The first section includes three chapters. Chapter 1 provides an overview of the textbook by outlining its purpose to train future public health professionals in the knowledge and skills to conduct applied secondary data analysis using national health surveys. Chapter 2 provides a general overview of the surveys used for the case studies presented in this textbook. The national surveys used for the case studies include the NHIS in Chapter 6, the MEPS in Chapter 7, the HINTS in Chapter 8, the Behavior Risk Factor Surveillance System (BRFSS) in Chapter 9, and the National Health and Nutrition Examination Survey (NHANES) in Chapter 10. Chapter 3 includes a literature review of previous studies that have used national health surveys to answer public health and health services related research questions that align with the case studies in each chapter.

1.3.2 TEXTBOOK SECTION 2: BASIC APPLIED DATA ANALYSIS

The second section includes two chapters. Chapter 4 reviews basic statistical functions commonly used for public health and health services research questions. It is expected that students who use this textbook will have some background knowledge of research methods and study design; however, this chapter includes some basics for students who do not have a strong foundation in research methodology. Chapter 4 includes basic terminology on types of data collected, descriptive, (frequencies, percentages, means, standard deviations) and analytical statistical procedures (chi square, logistic regression) used for analysis of national health surveys. Chapter 5 includes details on additional survey design features that need to be considered when analyzing complex surveys. These include using procedures like PROC SURVEYFREQ, including weights, primary sampling units, and stratum variables. SAS programming examples will be used with NHIS data in these chapters.

1.3.3 TEXTBOOK SECTION 3: COMMON NATIONAL HEALTH SURVEYS

The third section includes five chapters dedicated to common national health surveys used for secondary data analysis among public health and health services research professionals. Chapters provide:

1. A general overview of the survey and what it is used for;
2. An overview of the data files available;
3. Advantages of the survey;
4. Disadvantages of the survey;
5. Practical tips for conducting the analysis; and
6. Case study using a national health survey.

Each case study presents 1) a brief gap in the literature that the case study is attempting to address and 2) a research question. The case studies will outline the required steps to download, merge, create recoded (dummy) variables and analyze each dataset to answer research question. Sample SAS syntax will be provided.

Chapter 6 covers the NHIS. The objective of the NHIS survey case study is to explore whether Arab American/ Middle Eastern or North African (MENA) adults are more or less likely to receive an annual flu vaccine in comparison to other racial/ethnic groups, such as other non-Hispanic Whites. To answer this research question, 2018 NHIS person and sample adult files will be analyzed. Chapter 7 covers the MEPS. The objective of the MEPS survey case study is to explore whether adults who perceived their health care provider provided quality communication during their visits over the last 12 months are more or less likely to receive an annual flu vaccine in comparison those who did not receive quality patient-provider communication. To answer this research question, 2017 and 2018 MEPS household level in-person and self-administered questionnaire data will be analyzed. Chapter 8 covers the HINTS. The objective of the HINTS survey case study is to explore associations between electronic patient-provider communication and colon cancer screening uptake using HINTS 5 Cycle 3 data. Chapter 9 covers the BRFSS. The objective of the BRFSS survey case study is to explore how differences in caregiving experiences among urban and rural adults in Texas are moderated by race and ethnicity. To answer this research question, 2019 BRFSS state level data will be analyzed. Chapter 10 covers the NHANES. The objective of the NHANES survey case study is to estimate and compare sedentary behavior guideline adherence among US- and foreign-born adults by race and ethnicity using 2017-2020 pre-pandemic data.

1.3.4. TEXTBOOK SECTION 4: DISSEMINATION AND CONCLUSIONS

The fourth section includes two final chapters. Chapter 11 covers the dissemination of research studies using secondary data from national health surveys. It includes examples on how create poster presentations, oral presentations, abstracts, and full-length original research manuscripts. Chapter 12 provides a summary of what has been presented in the textbook and outlines potential recommendations for future editions.

1.4 SUMMARY

In summary, this textbook provides instruction on how to conduct basic applied data analysis using secondary data collected from national health surveys. The textbook has been developed based on a previous course, PH 2999:

Independent Study in Epidemiology. This individual study course was developed by Dr. Kindratt while receiving her PhD training at the University of Texas Health (UTHealth) School of Public Health Dallas Regional Campus. Dr. Kindratt developed for the University of Texas at Arlington's KINE 4352 Big Data for Epidemiology course. The content was originally created to meet the requirements of a breadth/ concentration in large database analysis because there was a lack of other courses which offered applied data analysis skills using secondary national health surveys to meet her professional goals and graduation requirements at that time. Learning objectives of the previous course were to:

1. Review existing research conducted using selected national health surveys;
2. Review sample designs and survey methods used when collecting national health survey data;
3. Develop SAS and STATA programs for merging and analyzing selected national health surveys; and
4. Create a teaching tool for each survey to summarize data analysis methods for future students.

The teaching tools developed for the course have been used as the model for each of the chapters in this textbook on specific national health surveys. The course included analysis of MEPS, BRFSS, and NHANES surveys. Examples of the teaching tools developed for PH 2999 are provided in the corresponding Open ICPSR data repository. The examples and content have been updated to reflect changes in survey designs, data collection modalities, and the research interests of the primary author. NHIS and HINTS case studies have been included to make this open textbook more comprehensive of what national surveys students will encounter in the workforce and may be used for students volunteering or working in UTA's HSR lab.

1.5 COVID-19 PANDEMIC CHANGES

The initial version of this textbook was written from June through December 2020 during the early waves of the COVID-19 pandemic. The methods described for the national surveys in this textbook represent "pre-pandemic" methodologies. Many surveillance systems and surveys had to modified due to safety concerns, stay-at-home orders, and data collection needs from 2020 onward.[4] Some chapters includes a brief section that discusses these changes for the respective survey.

1.6 REFERENCES

1. Ye J, Leep C, Robin N, Newman S. Perception of Workforce Skills Needed Among Public Health Professionals in Local Health Departments: Staff Versus Top Executives. *J Public Health Manag Pract*. 2015;21 Suppl 6:S151-158. doi:10.1097/PHH.0000000000000299

2. Andersen RM. National health surveys and the behavioral model of health services use. *Med Care*. 2008;46(7):647-653. doi:10.1097/MLR.0b013e31817a835d

3. Abuelezam NN, El-Sayed AM, Galea S. The Health of Arab Americans in the United States: An Updated Comprehensive Literature Review. *Front Public Health*. 2018;6:262. doi:10.3389/fpubh.2018.00262

4. Lau DT, Sosa P, Dasgupta N, He H. Surveillance, Surveys, and COVID-19. *Am J Public Health*. 2021;111(12):2085. doi:10.2105/AJPH.2021.306553

CHAPTER 2.

OVERVIEW OF NATIONAL HEALTH SURVEYS

2.1 INTRODUCTION

This chapter provides a general overview of the United States (US) national health surveys covered in this textbook. Summaries are provided describing the National Health Interview Survey (NHIS), Medical Expenditure Panel Survey (MEPS), Health Information National Trends Survey (HINTS), Behavioral Risk Factor Surveillance System (BRFSS), and National Health and Nutrition Examination Survey (NHANES). The corresponding chapters for each survey (Chapters 6-10) provide case studies using public-use data from each national health survey to answer research questions pertaining to how predisposing and enabling factors of individuals are associated with health behaviors and preventive services use.

2.2 HISTORY OF NATIONAL HEALTH SURVEYS

National health surveys have been used in the US since the 1920s. One of the first efforts to systematically collect health information from the US population came from the Committee on the Cost of Medical Care Studies (CMCS) who collected data from 1928 to 1933.[1] The CMCS received funding from private organizations to collect information on health care delivery and payments from 8,758 families in 17 states and Washington, DC. This monumental study documented the large disparity in health care costs, with 40% of the costs being incurred by only 10% of the families surveyed. A major limitation of this national data collection effort was that it did not include underrepresented minority groups, specifically Black or African American families.[1] From 1935-1936, the Public Health Service implemented the National Health Survey (NHS) to measure the incidence of illness and use of medical services. This survey was the first to use multistage area sampling across 21 states. The CMCS and NHS provided national health data until the early 1950s. In 1953, the Health Information Foundation in New York and National Opinion Research Center in Chicago collaborated to develop the first survey using a nation-wide probability sample, which laid the groundwork for national surveillance systems such as the NHIS to be conducted annually. More details of the history, design and context of national health surveys are provided elsewhere.[1]

2.3 NATIONAL HEALTH INTERVIEW SURVEY (NHIS)

Since 1957, the NHIS has been collected annually on a national scale by the National Center for Health Statistics (NCHS). The purpose of the NHIS is to monitor and explore trends in the health status and health care utilization among adults and children in the US.[2] Secondary analyses of NHIS data use cross-sectional study designs. Self-reported data are collected annually using a computer-assisted personal interviewing (CAPI) system during in-person interviews in households across the US. The NHIS has one of the largest annual sample sizes among national surveys. Data are collected from roughly 35,000 households and 87,500 individuals each year. From 2014-2018, the annual household response rates slightly decreased from 73.8% in 2014 to 64.2% in 2018.[2] Further details on the design, questionnaires, public-use data, and reports are available on the NHIS website.

2.4 MEDICAL EXPENDITURE PANEL SURVEY (MEPS)

Since 1996, the MEPS has been collected on a national scale by the Agency for Healthcare Research and Quality (AHRQ). The purpose of the MEPS is to gather information on health services used by adults and children in the US, including cost, frequency, and payment structures.[3] The MEPS uses a survey panel design that consists of five rounds of interviews over a two-year period. Therefore, the secondary analysis of MEPS data can use both longitudinal or cross-sectional study designs. Households recruited for each panel are selected based on a subsample of households who participated in the previous year's NHIS. Similar to the NHIS, data are collected in-person using a CAPI system. Self-administered paper questionnaires are also completed by participants. Medical providers are contacted by telephone to provide additional details on medical visit summaries, diagnostic codes and billing. The MEPS annual sample size is roughly 15,000 individuals. From 2014-2018, the annual combined (all five rounds completed) response rates slightly decreased from 48.5% in 2014 to 42.7% in 2018.[3] Further details on the design, questionnaires, public-use data, and reports are available on the MEPS website.

2.5 HEALTH INFORMATION NATIONAL TRENDS SURVEY (HINTS)

Since 2003, the HINTS has been collected on a national scale by the National Cancer Institute (NCI). The purpose of the HINTS is to evaluate how patterns of health information technology and health communication are related to health-related knowledge, attitudes and behaviors among the noninstitutionalized US adult civilian population.[4] After the first 3 iterations (HINTS 1 collected in 2003, HINTS 2 collected in 2005, HINT 3 collected in 2008), each iteration was separated into four cycles. HINTS 4 cycles were collected annually beginning in 2011 (HINTS 4, Cycle 1 in 2011; HINTS 4, Cycle 2 in 2012; HINTS 4, Cycle 3 in 2013; HINTS 4, Cycle 4 in 2014). HINTS 5 cycles were collected annually beginning in 2017 (HINTS 5, Cycle 1 in 2017; HINTS 5, Cycle 2 in 2018; HINTS 5, Cycle 3 in 2019; HINTS 5, Cycle 4 in 2020). Secondary analyses of HINTS data use cross-sectional study designs. Self-reported data have been collected using random digit dialing, mailings and web-based data collection options. The HINTS sample size is roughly 3,500-6,000 individuals for each iteration. Response rates are calculated for each data collection method. Total response rates are roughly 30% for each iteration.[4] Further details on the design, questionnaires, public-use data, and reports are available on the HINTS website.

2.6 BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM (BRFSS)

Since 1984, the BRFSS has been collected on the state level by the Centers for Disease Control and Prevention (CDC). The purpose of the BRFSS is collect data on health behaviors, physical activity, diet, hypertension and preventive safety measures (e.g. seat-belt use) among US adults.[5] In 1988, the system was expanded to include optional modules, including chronic disease, health care access, and preventive services uptake. Some optional modules include data collection among children. In 1993, the BRFSS was expanded to become an annual national surveillance system. Secondary analyses of BRFSS data use cross-sectional study designs. Self-reported data are collected annually using random-digit-dialing methods. Data are collected using a computer-assisted telephone interview (CATI) system. Prior to 2008, data were only collected from landline telephones. In 2008, the methodology was revised to conduct interviews using cell phones. The BRFSS is one of the largest health surveys collected worldwide with over 400,000 responses collected each year.[6] Response rates are calculated for landline, cell phone, and combined responses. In 2019, the overall response rate was 49.4%. In 2018, the landline response rate was 53.3% and the cell phone response rate was 43.4%.[5] Further details on the design, questionnaires, public-use data, and reports are available on the BRFSS website.

2.7 NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES)

Since 1960, national data on the health and diet of individuals in the US has been collected by the NCHS. Starting in 1999, NHANES data have been collected from adults and children on a consistent basis.[7] Topics have been expanded to include chronic diseases and other health indicators over the years. Secondary analyses of NHANES data use cross-sectional study designs. The NHANES differs from other national health surveys because it collects self-reported data using in-person household interviews while also collecting objective measurements by physical examinations and laboratory tests of participants at mobile examination centers.[7,8] The inclusion of both objective and subjective measurements allows for reliability comparisons. For example, participants self-report whether or not they have ever been diagnosed with diabetes during the household interviews and will have their glucose tested for verification at the mobile examination center. The sample size includes approximately 5,000 individuals every year and data are compiled across two-year data collection periods (e.g. 2015-2016, 2017-2018). Response rates are calculated for interviewed and examined samples. During 2017-2018, the interviewed response rate was 51.9% and the examination response rate was 48.8%.[8] Further details on the design, questionnaires, public-use data, and reports are available on the NHANES website.

2.8 RESTRICTED DATA

Although most data collected by national health surveys are made available to the public in a deidentified format, any data that may compromise the confidentiality of its participants are restricted and require special procedures and approvals to access. Among the national health surveys discussed in this textbook, the NHANES and NHIS have restricted data accessible by the NCHS and Federal Statistical Research Data Centers. The NHANES restricts data on geography (Census 2010 Block ID), genetics (e.g. BRCA1 associated protein), and the exact dates of participants' interviews and examinations.[9] The NHIS restricts data on geography.[10] Among individuals born in the US, data on participants' state of birth (variable: USBRTHPL) and whether they live in an urban or rural residence (variable: URB_RRL) are restricted. Among foreign-born participants, data on the year that participants came to the US (variable: USYR) and their country of birth (variable: COUNTRY) are restricted.[10] The primary author of this textbook (Kindratt) and colleagues have analyzed restricted NHIS data to determine chronic disease prevalence,[11] preventive cancer screenings and vaccinations among men[12] and women,[13] and smoking status[14] among Arab American immigrants. The NHIS collects data on country of birth and categorizes each country into 10 worldwide geographic regions (US, Mexico, Central America and Caribbean Islands, South American, Europe, Russia, Middle East, India subcontinent, Southeast Asia, and Asia). The Middle East region includes individuals who were born in 25 countries. However, some countries, such as Iran, are located in the Middle East but not part of the Arab League of Nations. Therefore, previous research by Kindratt and others used responses to the country of birth question to create a variable limited to individuals born in 15 countries that were part of the Arab League of Nations and geographically located in the Middle East region to ensure the findings were representative of the Arab ethnicity.[11-14] Using restricted data from the NHIS allowed the authors to disaggregate Arab American immigrants from other ethnicities and exclude non-Arab countries (e.g. Iran) from the grouping.

2.9 LINKED DATA

National health survey data can also be linked to each other and to other sources. The NHANES and NHIS can be linked to data from the National Death Index to determine mortality rates. For example, Borrell and colleagues linked NHANES III data collected from 1988-1994 with 2015 mortality data to examine associations between allostatic load and all-cause/ cardiovascular disease specific mortality among US adults.[15] Because the MEPS collects data from the previous year's NHIS sample, ID numbers from each survey can be linked to expand the variables for each survey.[16] Kindratt and colleagues have used linked NHIS and MEPS data to answer several research questions on Middle Eastern and North African cognitive health[17] and parents' perceptions of patient- and family-centered care practices among those whose children have developmental and chronic health conditions.[18] Additional studies using linked NHIS and MEPS data are underway.

2.10 OTHER SURVEYS

There are several other national health surveys that provide surveillance data for public health professionals and researchers to utilize and examine trends. Inclusion criteria, sample sizes, and health-related content differs across surveys. Although not a comprehensive list, a selection of other common national health surveys are listed below:

- American Community Survey (ACS)
- Health and Retirement Study (HRS)
- National Death Index (NDI)
- National Health and Aging Trends Survey (NHATS)
- National Longitudinal Study of Adolescent to Adult Health (Add Health)
- National Study of Caregiving (NSOC)
- National Survey of Family Growth (NSFG)
- National Vital Statistics System (NVSS)
- Youth Risk Factor Surveillance System (YRFSS)

2.11 SUMMARY

This chapter provided an overview of the national health surveys covered in this textbook> It also covers brief details of some expanded data analysis procedures and exposure to other health surveys to broaden students' knowledge of other data sources. More specific details on each survey are provided in Chapters 6-10 (Chapter 6 NHIS, Chapter 7 MEPS, Chapter 8 HINTS, Chapter 9 BRFSS, Chapter 10 NHANES).

2.12 REFERENCES

1. Andersen RM. National health surveys and the behavioral model of health services use. *Med Care*. 2008;46(7):647-653. doi:10.1097/MLR.0b013e31817a835d

2. National Center for Health Statistics. National Health Interview Survey, 2011-2015. Public-use data file and documentation. Published August 5, 2020. Accessed August 10, 2020. https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm

3. Agency for Healthcare Research and Quality (AHRQ). Medical Expenditure Panel Survey Home. Accessed September 2, 2020. https://meps.ahrq.gov/mepsweb/

4. National Cancer Institute. Health Information National Trends Survey (HINTS): Overview of the HINTS 5 Cycle 3 Survey and Data Analysis Recommendations, January 2020.

5. Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System: 2019 Summary Data Quality Report.*; 2020. Accessed June 19, 2022. https://www.cdc.gov/brfss/annual_data/2019/pdf/2019-sdqr-508.pdf

6. Centers for Disease Control and Prevention. About the Behavioral Risk Factor Surveillance System (BRFSS). Published February 9, 2019. Accessed December 10, 2020. https://www.cdc.gov/brfss/about/about_brfss.htm

7. NHANES – About the National Health and Nutrition Examination Survey. Published January 8, 2020. Accessed November 3, 2021. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm
8. Chen TC, Parker JD, Clark J, Shin HC, Rammon JR, Burt VL. National Health and Nutrition Examination Survey: Estimation Procedures, 2011-2014. *Vital Health Stat 2.* 2018;(177):1-26.
9. RDC – Restricted Data – NHANES. Published August 27, 2021. Accessed June 23, 2022. https://www.cdc.gov/rdc/b1datatype/Dt1222.htm
10. RDC – Restricted Data – NHIS. Published March 17, 2022. Accessed June 23, 2022. https://www.cdc.gov/rdc/b1datatype/Dt1225.htm
11. Dallo FJ, Kindratt TB. Disparities in Chronic Disease Prevalence Among Non-Hispanic Whites: Heterogeneity Among Foreign-Born Arab and European Americans. *J Racial Ethn Health Disparities*. 2016;3(4):590-598. doi:10.1007/s40615-015-0178-8
12. Dallo FJ, Kindratt TB. Disparities in preventive health behaviors among non-Hispanic White men: heterogeneity among foreign-born Arab and European Americans. *Am J Mens Health*. 2015;9(2):124-131. doi:10.1177/1557988314532285
13. Dallo FJ, Kindratt TB. Disparities in vaccinations and cancer screening among U.S.- and foreign-

born Arab and European American non-Hispanic White women. *Womens Health Issues*. 2015;25(1):56-62. doi:10.1016/j.whi.2014.10.002

14. Kindratt TB, Dallo FJ, Roddy J. Cigarette Smoking among US- and Foreign-Born European and Arab American Non-Hispanic White Men and Women. *J Racial Ethn Health Disparities*. 2018;5(6):1284-1292. doi:10.1007/s40615-018-0476-z

15. Borrell LN, Rodríguez-Álvarez E, Dallo FJ. Racial/ethnic inequities in the associations of allostatic load with all-cause and cardiovascular-specific mortality risk in U.S. adults. *PLoS One*. 2020;15(2):e0228336. doi:10.1371/journal.pone.0228336

16. NHIS – Medical Expenditure Panel Survey Linkage Files. Published May 10, 2019. Accessed June 23, 2022. https://www.cdc.gov/nchs/nhis/nhismep.htm

17. Kindratt TB, Dallo FJ, Zahodne LB, Ajrouch KJ. Cognitive Limitations Among Middle Eastern and North African Immigrants. *J Aging Health*. Published online May 23, 2022:8982643221103712. doi:10.1177/08982643221103712

18. Kindratt TB, Lark P, Ray M, Brannon GE. Disparities in Patient- and Family-Centered Care Among Children With Health Conditions. *J Patient Exp*. 2022;9:23743735221092496. doi:10.1177/23743735221092494

CHAPTER 3.

LITERATURE REVIEW

3.1 OVERVIEW

This chapter includes a literature review of previous studies that have used national health surveys to answer public health and health services research questions. The background literature provided lays the groundwork for the case studies used in Chapters 6-10 in this textbook. However, the literature reviews are not comprehensive. The reader is encouraged to conduct their own literature reviews in PubMed, Ovid MEDLINE, Google Scholar, and other library sources to gain a deeper understanding of the existing evidence for each topic.

3.2 NATIONAL HEALTH INTERVIEW SURVEY (NHIS) CASE STUDY

The objective of the NHIS survey case study is to determine associations between a combined measure of race, ethnicity, and nativity status and seasonal influenza vaccine uptake among foreign-born Arab Americans compared to other racial/ethnic groups. Data from the

2018 NHIS person and sample adult files will be used to fulfil this objective.

3.2.1 WHY EXAMINE DIFFERENCES IN INFLUENZA VACCINE UPTAKE AMONG FOREIGN-BORN ARAB AMERICANS COMPARED TO OTHER US- GROUPS?

During the 2018-2019 season, it was estimated that there were 38,000,000 cases of symptomatic illness, 18,000,000 medical visits, and 22,000 deaths in the US.[1] Seasonal influenza vaccination is recommended among all individuals ages 6 months and older to prevent morbidity and mortality from influenza and other health conditions.[2] Despite established benefits, disparities exist in vaccination coverage by race, ethnicity, and nativity status. Using 2010-2016 NHIS data, Lu and colleagues found that Hispanic and non-Hispanic Black adults were less likely to receive annual influenza vaccines compared to non-Hispanic Whites.[3] Using 2012 data, Lu and colleagues found foreign-born adults were less likely to receive an influenza vaccination than their US-born counterparts.[4] Results were similar among other studies.[5] Research on influenza vaccination coverage among Arab Americans is limited despite evidence showing that morbidity and mortality estimates for several health conditions are higher than other groups. For example, Dallo and colleagues evaluated administrative hospital data and found that Arab American women were more likely to have influenza or pneumonia than non-Hispanic White women in Michigan.[6] Furthermore, other research has demonstrated that Arab American males have higher mortality rates from influenza or pneumonia than other non-Hispanic White males.[7] In 2015, Dallo and Kindratt used NHIS data to determine the prevalence of not receiving influenza vaccinations among Arab American men and women compared to US- and foreign-born non-Hispanic White adults from Europe using NHIS person level and sample adult data.[8,9] A foreign-born Arab American ethnic group was created using restricted country of birth data collected from the NHIS. Results indicated that foreign-born Arab American men had 62% lower odds (OR=0.38; 95% CI=0.21-0.67) and foreign-born Arab American women had 66% lower odds (OR=0.34; 95% CI=0.21-0.58) of receiving an influenza vaccine compared to their US-born non-Hispanic White counterparts.[8,9] ***The NHIS case study will extend this previous research by using 2018 public-use person and sample adult data.***

3.3 MEDICAL EXPENDITURE PANEL SURVEY (MEPS) CASE STUDY

The objective of the MEPS case study is to determine associations between adults' perceptions of patient-provider communication quality and seasonal influenza vaccination uptake. Data from the 2015 and 2016 MEPS household level in-person and self-administered questionnaire data will be used to fulfil this objective.

3.3.1 WHY DETERMINE HOW ADULTS' PERCEPTIONS OF PATIENT-PROVIDER COMMUNICATION QUALITY ARE ASSOCIATED WITH INFLUENZA VACCINATION?

Efforts are needed to address barriers to influenza vaccination uptake among underrepresented racial, ethnic, and immigrant minority groups. Previous research suggests that effective communication between health care providers and patients during in-person and between visits may contribute to more adults receiving recommended preventive services, including cancer screenings and influenza vaccinations.[10-13] Kindratt and colleagues previous research using 2011-2015 MEPS data examined associations between adults' perceptions of specific qualities of patient-provider communication and their likelihood of receiving cancer screenings by racial and ethnic subgroups.[10] Results demonstrated that Hispanic and non-Hispanic Black adults who reported their health care providers gave them specific instructions had higher odds of receiving breast and colorectal cancer screenings. Non-Hispanic Asian women who reported their health care providers asked them to describe how they were going to follow the instructions given to them had higher odds of receiving cervical cancer screenings.[10] Research examining the role of patient-provider communication on influenza vaccine uptake using nationally representative samples is limited. Villani and Mortensen (2013) examined the influence of patient-provider communication qualities on preventive services uptake, including recommended cancer screenings and vaccinations, using 2009 MEPS data.[14] They did not find a statistically significant association between adults' (ages 50+ years) perceptions of patient-provider communication and influenza vaccine uptake. However, to my knowledge, no other studies have examined the role of patient-provider communication during face-to-face visits on influenza vaccine uptake using nationally representative MEPS data. ***The MEPS***

case study will extend this previous research by using 2015-2016 household data.

3.4 HEALTH INFORMATION NATIONAL TRENDS SURVEY (HINTS) CASE STUDY

The objective of the HINTS case study is to explore associations between e-mail communication and breast cancer screening uptake. Data from the HINTS 5, Cycle 3 data collected in 2019 will be used to fulfil this objective.

3.4.1 WHY DETERMINE HOW THE USE OF E-MAIL COMMUNICATION IS ASSOCIATED WITH BREAST CANCER SCREENING UPTAKE?

Advances in health information technology and the use of the internet as a mode of communication have allowed for greater interaction between health care providers and their patients between visits. In addition to traditional telephone communications, patients can communicate with their health care providers by e-mail, text messaging, patient portals, and mobile applications. Previous studies have examined patients' perceptions of the benefits of electronic patient-provider communication, specifically using e-mail communication. Patients identified some benefits to using e-mail communication, including convenient access at any time, increased level of comfort asking questions, and the ability to save and keep track of conversations.[16] Studies have shown that using e-mail patient-provider communication may lead to improved health outcomes. Research examining associations between e-mail patient-provider communication and adults' use of preventive services are limited. Using 2011-2015 NHIS data, Kindratt and colleagues demonstrated that adults who used e-mail to communicate with their health care providers had 1.51 times greater odds (95% CI=1.44-1.59) of receiving a seasonal influenza vaccine compared to those who do not use e-mail to communicate with their health care providers.[12] Using HINTS 4, Cycles 1-4 data, Kindratt and colleagues also looked at associations between e-mail patient-provider communication and cancer screenings using HINTS data. Results demonstrated that there was not a significant association between e-mail patient-provider communication and breast, cervical or colorectal cancer screenings.[11] No other studies have evaluated the influence of e-mail patient-provider communication practices on cancer screenings using national representative HINTS data. ***The HINTS case study will extend this previous research by using HINTS 5, Cycle 3 data.***

3.5 BEHAVIOR RISK FACTOR SURVEILLANCE SYSTEM (BRFSS) CASE STUDY

The objective of the BRFSS case study is to explore whether differences in Alzheimer's disease and related dementia (ADRD) caregiving experiences among urban and rural adults in Texas are moderated by race and ethnicity. The differences obtained among urban and rural adults will be evaluated as a whole, and stratified by racial and ethnic groups. Data from the 2019 BRFSS will be used to fulfil this objective.

3.5.1 WHY EXPLORE HOW DIFFERENCES IN ADRD CAREGIVING EXPERIENCES AMONG URBAN AND RURAL ADULTS IN TEXAS ARE MODERATED BY RACE AND ETHNICITY?

In 2020, the National Alliance for Caregiving and American Association of Retired Persons estimated that 21% of adults in the US are informal caregivers, which has increased by 9.5 million since 2015.[17] Over 11 million unpaid individuals, family or friends, are caregivers for persons living with ADRD.[17] While most older adults with ADRD are currently non-Hispanic White, the racial and ethnic diversity of older adults living with ADRD is increasing.[18]

Previous studies on ADRD caregiving experiences across geographic contexts highlight unmet resource needs and support the lack of dementia-specific[19] and respite services[20] in non-metro or rural areas. Urban/rural comparisons of ADRD caregiving experiences have been limited to descriptive analyses due to research studies only being conducted with small non-representative samples. Few studies have examined differences in caregiving experiences among racial and ethnic caregivers living in urban and rural areas.[21]

A recent study was conducted using data from the National Study of Caregiving (NSOC), which includes a sample of caregivers linked to the National Health and Aging Trends Survey (NHATS).[22] The aims of the study were to determine whether: 1) caregiver experiences and health differed across urban and rural areas and 2) the links between caregiving experiences and health were moderated by caregiver race/ethnicity. Results indicated non-metro ADRD caregivers were less racially/ethnically diverse (82.7% White), and more were spouses/partners (20.2%).[22] Among racial/ethnic minority ADRD caregivers, non-metro context was associated with having more chronic conditions, providing less care, and not co-residing with care recipients. Amid White ADRD caregivers, non-metro context was associated with not reporting caregiving was more than they could handle and finding financial assistance for caregiving. Non-metro minority ADRD caregivers had 3.09 times higher odds (95% CI=1.02-9.36) of reporting anxiety in comparison to metro minority ADRD caregivers.[22] While this study lays the groundwork for national research on ADRD caregiving by geographic context, large differences may exist by state. ***The BRFSS case study will extend this previous research by using BRFSS data from Texas.***

3.6 NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES) CASE STUDY

The NHANES case study will focus on movement behaviors among US adults. The objective of the NHANES case study is to evaluate adherence to 24-hour movement guidelines (sleep, sedentary behavior, and physical activity) among US adults and determine differences by race, ethnicity, and nativity status. Sedentary behavior will be used as the outcome of interest. Specifically, data from the 2017-2020 in-person interviews and examination data will be used to fulfil this objective.

3.6.1. WHY EVALUATE 24-HOUR MOVEMENT GUIDELINE ADHERENCE AMONG RACIAL AND ETHNIC GROUPS IN THE US?

The recently published 24-hour movement guidelines include recommendations for sedentary behavior, physical activity, and sleep among adults ages 18-64 years and 65 years and older.[23] The guidelines integrate recommendations for sleep, physical activity, and sedentary behavior with the acknowledgement that combination of these behaviors throughout the day is associated with health outcomes.[23] There are slight differences between recommendations for younger and older adults. For example, it is recommended that adults ages 18-64 years get 7 to 9 hours of good-quality sleep on a regular basis, with consistent bed and wake-up times. It is recommended that adults perform a variety of intensities and types of physical activity, including 1) moderate to vigorous aerobic physical activities that accumulate up to 150 minutes per week, 2) muscle strengthening activities using major muscle groups at least twice a week, and 30 several hours of light physical activities, including standing. Finally, it is recommended that adults limit sedentary behavior to 8 hours or less (~480 minutes), including no more than 3 hours of recreational screen time and breaking up long periods of sitting as often as possible.[24] Little is known about how adherence to these guidelines differs among US adults, particularly among different racial and ethnic groups. ***The NHANES case study will explore racial and ethnic differences in sedentary behavior among US- and foreign-born Hispanics, non-Hispanic Whites, non-Hispanic***

Blacks, and non-Hispanic Asians using 2017-2020 pre-pandemic data.

3.7 SUMMARY

In summary, this chapter provided a brief background to support the case studies used in Chapters 6-10. The topics of the case studies are broad and encompass the wide range of research being conducted using national health surveys by the primary author of this textbook. The reader is encouraged to conduct their own literature reviews using electronic databases to gain a deeper understanding of the content areas for each case study.

3.8 REFERENCES

1. Centers for Disease Control and Prevention (CDC). Estimated Influenza Illnesses, Medical visits, Hospitalizations, and Deaths in the United States — 2019–2020 Influenza Season | CDC. Published October 6, 2020. Accessed December 15, 2020. https://www.cdc.gov/flu/about/burden/2019-2020.html
2. Grohskopf LA, Alyanak E, Broder KR, et al. Prevention and Control of Seasonal Influenza with Vaccines: Recommendations of the Advisory Committee on Immunization Practices – United States, 2020-21 Influenza Season. *MMWR Recomm Rep*. 2020;69(8):1-24. doi:10.15585/mmwr.rr6908a1
3. Lu PJ, Hung MC, O'Halloran AC, et al. Seasonal Influenza Vaccination Coverage Trends Among Adult Populations, U.S., 2010-2016. *Am J Prev*

Med. 2019;57(4):458-469. doi:10.1016/j.amepre.2019.04.007

4. Lu PJ, Rodriguez-Lainz A, O'Halloran A, Greby S, Williams WW. Adult vaccination disparities among foreign-born populations in the U.S., 2012. *Am J Prev Med*. 2014;47(6):722-733. doi:10.1016/j.amepre.2014.08.009
5. Vlahov D, Bond KT, Jones KC, Ompad DC. Factors associated with differential uptake of seasonal influenza immunizations among underserved communities during the 2009-2010 influenza season. *J Community Health*. 2012;37(2):282-287. doi:10.1007/s10900-011-9443-x
6. Dallo FJ, Ruterbusch JJ, Kirma JD, Schwartz K, Fakhouri M. A Health Profile of Arab Americans in Michigan: A Novel Approach to Using a Hospital Administrative Database. *J Immigr Minor Health*. 2016;18(6):1449-1454. doi:10.1007/s10903-015-0296-8
7. Dallo FJ, Schwartz K, Ruterbusch JJ, Booza J, Williams DR. Mortality rates among Arab Americans in Michigan. *J Immigr Minor Health*. 2012;14(2):236-241. doi:10.1007/s10903-011-9441-1
8. Dallo FJ, Kindratt TB. Disparities in preventive health behaviors among non-Hispanic White men: heterogeneity among foreign-born Arab and European Americans. *Am J Mens Health*. 2015;9(2):124-131. doi:10.1177/1557988314532285

9. Dallo FJ, Kindratt TB. Disparities in vaccinations and cancer screening among U.S.- and foreign-born Arab and European American non-Hispanic White women. *Womens Health Issues*. 2015;25(1):56-62. doi:10.1016/j.whi.2014.10.002

10. Kindratt TB, Dallo FJ, Allicock M, Atem F, Balasubramanian BA. The influence of patient-provider communication on cancer screenings differs among racial and ethnic groups. *Prev Med Rep*. 2020;18:101086. doi:10.1016/j.pmedr.2020.101086

11. Kindratt TB, Atem F, Dallo FJ, Allicock M, Balasubramanian BA. The Influence of Patient–Provider Communication on Cancer Screening. *Journal of Patient Experience*. Published online May 11, 2020:2374373520924993. doi:10.1177/2374373520924993

12. Kindratt TB, Allicock M, Atem F, Dallo FJ, Balasubramanian BA. Email Patient-Provider Communication and Cancer Screenings Among US Adults: Cross-sectional Study. *JMIR Cancer*. 2021;7(3):e23790. doi:10.2196/23790

13. Kindratt T, Callender L, Cobbaert M, Wondrack J, Bandiera F, Salvo D. Health information technology use and influenza vaccine uptake among US adults. *Int J Med Inform*. 2019;129:37-42. doi:10.1016/j.ijmedinf.2019.05.025

14. Villani J, Mortensen K. Patient-provider communication and timely receipt of preventive services. *Prev Med*. 2013;57(5):658-663.

doi:10.1016/j.ypmed.2013.08.034

15. Emily S. Lau MD, Sharonne N. Hayes MD, Annabelle Santos Volgman MD, et al. Does Patient-Physician Gender Concordance Influence Patient Perceptions or Outcomes? *Journal of the American College of Cardiology*. Published online March 2, 2021. Accessed November 15, 2021. https://www.jacc.org/doi/10.1016/j.jacc.2020.12.031

16. Ye J, Rust G, Fry-Johnson Y, Strothers H. E-mail in patient-provider communication: a systematic review. *Patient Educ Couns*. 2010;80(2):266-273. doi:10.1016/j.pec.2009.09.038

17. Jr SM. Caregiving in the US 2020 | The National Alliance for Caregiving. Published May 11, 2020. Accessed June 25, 2021. https://www.caregiving.org/caregiving-in-the-us-2020/

18. 2022 Alzheimer's disease facts and figures. *Alzheimers Dement*. 2022;18(4):700-789. doi:10.1002/alz.12638

19. Gibson A, Holmes SD, Fields NL, Richardson VE. Providing Care for Persons with Dementia in Rural Communities: Informal Caregivers' Perceptions of Supports and Services. *J Gerontol Soc Work*. 2019;62(6):630-648. doi:10.1080/01634372.2019.1636332

20. Kosloski K, Schaefer JP, Allwardt D, Montgomery RJV, Karner TX. The role of cultural factors on clients' attitudes toward caregiving, perceptions of service delivery, and service utilization. *Home*

Health Care Serv Q. 2002;21(3-4):65-88. doi:10.1300/J027v21n03_04

21. Dilworth-Anderson P, Moon H, Aranda MP. Dementia Caregiving Research: Expanding and Reframing the Lens of Diversity, Inclusivity, and Intersectionality. Bowers BJ, ed. *The Gerontologist*. 2020;60(5):797-805. doi:10.1093/geront/gnaa050

22. Kindratt T, Sylvers D, Yoshikawa A, Anuarbe ML, Webster N, Bouldin E. ADRD Caregiving Experiences and Health by Race, Ethnicity and Care Recipient Geographic Context. *Innov Aging*. 2021;5(Suppl 1):990. doi:10.1093/geroni/igab046.3557

23. Ross R, Tremblay M. Introduction to the Canadian 24-Hour Movement Guidelines for Adults aged 18-64 years and Adults aged 65 years or older: an integration of physical activity, sedentary behaviour, and sleep. *Appl Physiol Nutr Metab*. 2020;45(10 (Suppl. 2)):v-xi. doi:10.1139/apnm-2020-0843

24. Ross R, Chaput JP, Giangregorio LM, et al. Canadian 24-Hour Movement Guidelines for Adults aged 18-64 years and Adults aged 65 years or older: an integration of physical activity, sedentary behaviour, and sleep. *Appl Physiol Nutr Metab*. 2020;45(10 (Suppl. 2)):S57-S102. doi:10.1139/apnm-2020-0467

CHAPTER 4.

BASIC DATA ANALYSIS

4.1. OVERVIEW

This chapter covers the steps for preparing secondary data for statistical analysis and how to run common statistical tests. The reader will be introduced to some basic data analysis procedures using SAS 9.4. It is important to follow the steps for preparing secondary data for statistical analysis to ensure accuracy, particularly when using numerous years or combining multiple data files within each year. The examples in this chapter will use data from the 2018 National Health Interview Survey (NHIS) Sample Adult file. The examples will demonstrate ways to answer the following two research questions among adults ages 18 and older in the United States (US):

- Research Question 4.1: What are the associations between region (geographic location where participant lives) and health information technology (HIT) usage?
- Research Question 4.2: What is the association

between sex and HIT usage?

The NHIS includes five questions to measure individuals' use of different aspects of HIT usage during the past 12 months, including whether or not adults use computers to fill prescriptions, schedule appointments, communicate with others through chat groups, look up health information online, and communicate with health care providers by e-mail.[1] The primary author of this textbook (Kindratt) and colleagues have used NHIS data to explore how HIT usage influences vaccination and cancer screening uptake.[2,3] The examples in this chapter will focus on differences by sociodemographic factors, such as place of residence and sex. SAS 9.4 procedures will be used to demonstrate how to meet the following research objectives using common statistical tests.

- Objective 4.1: To determine the association between region and looking up health information on the internet
- Objective 4.2: To determine the association between region and filling prescriptions on the internet
- Objective 4.3: To determine the association between region and scheduling appointments on the internet
- Objective 4.4: To determine the association between region and communicating with health care provider by e-mail
- Objective 4.5: To determine the association between sex and the number of HIT uses (scale from 0 to 4)

- Objective 4.6: To determine the associations between sex and each HIT use (looking up health information online, filling prescriptions, scheduling appointments, and communicating with a health care provider by e-mail) before and after controlling for other contributing factors

It is important to note that the examples provided in this chapter are not weighted and do not include procedures for adjusting the results based on the complex survey design. Therefore, estimates you obtain as the results are not representative of the true findings. The complex sample design features that should be used when analyzing national health data will be described in Chapter 5.

4.2 FOUR-STEP PROCESS FOR SECONDARY DATA ANALYSIS

When preparing secondary data for statistical analysis, it is important to follow these steps to ensure accuracy and completeness of your data. This 4-step process was developed based on recommendations from Elliot and colleagues for preparing and managing primary data in Microsoft Excel and Database Creation and Coding sessions by Kindratt for training medical and physician assistant students.[4-6] The four steps include: 1) data selection; 2) data collection; 3) data verification; and 4) data storage.

4.2.1 DATA SELECTION

The first step is data selection, which is the process of determining the appropriate data type and source, as well as suitable instruments to collect data. Once a research question has been solidified, the next step in secondary data analysis is to determine the appropriate data source. For this example, the research question focuses on differences in where adults live in the US and their HIT use. Since the NHIS is one of the largest national surveys that collects this information and provides publicly available data, it is a good selection to answer this research question.

4.2.2 DATA COLLECTION

The second step is data collection, which is the process of gathering and measuring information on variables of interest in an established and systematic fashion that enables one to answer stated research questions and test hypotheses. For secondary data analysis, data collection procedures include downloading the necessary data file and supporting documentation from websites and collecting data from only the variables needed to answer the research question and objectives. The 2018 NHIS Sample Adult file and Sample SAS statements can be downloaded from the NHIS data release website.

To collect the appropriate variables for this analysis, complete the following:

- Go to your "C:\" drive and create a folder named "NHIS"
- In the "NHIS" folder, create a folder named "18"
- Download the 2018 Sample Adult ASCII data file (.dat) from the 2018 NHIS data release website
- Unzip the file and save it in the folder "C:\NHIS\18"

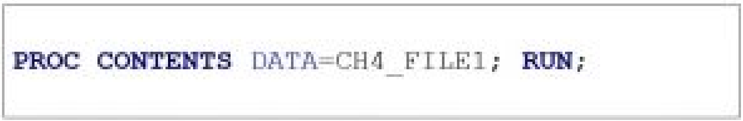
- Download “File 4.1 Sample SAS Program to Create NHIS 2018 Sample Adult File” from the Open ICPSR data repository
- Open the file and select “Run” from the top menu bar

Your analytic dataset should include the following 9 variables:

1. SRVY_YR
2. FPX
3. AGE_P
4. SEX
5. REGION
6. HIT1A
7. HIT2A
8. HIT3A
9. HIT4A

To verify the variables in the analytic dataset, run the PROC CONTENTS statement in Box 4.1.

Box 4.1. SAS procedure (PROC CONTENTS) for verifying the variables included in the 2018 NHIS Sample Adult example analytic dataset



```
PROC CONTENTS DATA=CH4_FILE1; RUN;
```

If you are unable to create the permanent analytic dataset using the preceding steps, click on “File 4.2

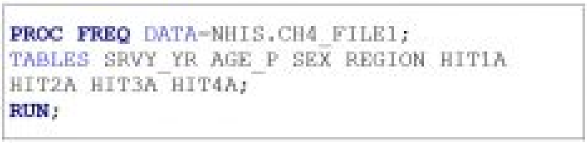
Permanent SAS Analytic Database" to download and save the database in the folder "C:\NHIS\18."

4.2.3 DATA VERIFICATION

The third step is data verification, which involves verifying that the results from the analytic dataset you created match those provided by the original dataset. Most national health surveys, including the NHIS, provide at least unweighted frequency counts for you to match your findings with those published on the website. Other national health surveys, such as the Medical Expenditure Panel Survey, provide more detailed information like weighted frequencies and percentages. The frequency verification should be completed prior to making any changes to the variables you collected for your analysis. Your results will not match if you verify the frequencies after applying any limitations to the data. For example, if your sample only includes adults ages 45 and older, I recommend that you verify the results for adults ages 18 and older prior to removing individuals ages 18-44 years to ensure the accuracy of the analytic dataset.

In this example, you can view the unweighted frequencies for variables you collected by going to the 2018 NHIS data release website and clicking on "Variable frequencies" in the "Sample Adult File" section. You should be able to search for each variable by clicking "Ctrl" and "F" and typing the name of the variable into the search box. You will need to run frequencies for each variable of interest in SAS to verify results. PROC FREQ is the procedure for displaying frequencies in SAS. You can enter the SAS syntax from Box 4.2 to determine the frequencies for all variables in the analytic dataset. Variable FPX is excluded because it represents the ID number of the participant.

Box 4.2. SAS procedure (PROC FREQ) for verifying frequencies in the 2018 NHIS Sample Adult example analytic dataset



```
PROC FREQ DATA=NHIS.CH4_FILE1;
TABLES SRVY_YR AGE_P SEX REGION HIT1A
HIT2A HIT3A HIT4A;
RUN;
```

The unweighted frequencies for the variables collected in this example are presented in Table 4.1. The variable for age (AGE_P) is excluded from the table due to multiple response options (age 18-85 years and older).

Table 4.1. Frequencies for data collected in the 2018 NHIS Sample Adult example analytic dataset

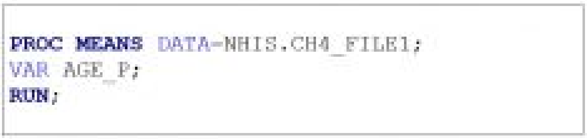
Variable Description (VARIABLE NAME)	Frequency
Survey Year (SRVY_YR)	
2018	25,417
Sex (SEX)	
1=Male	11,550
2=Female	13,867
Region (REGION)	
1=Northeast	4,143
2=Midwest	5,949
3=South	9,312
4=West	6,013
Looked up health information on internet, past 12 months (HIT1A)	
1=Yes	13,677
2=No	11,431
7=Refused	11
8=Not ascertained	273
9=Don’t know	25
Filled a prescription on internet, past 12 months (HIT2A)	
1=Yes	2,892
2=No	22,240
7=Refused	6
8=Not ascertained	274
9=Don’t know	5

Table 4.1 (continued). Frequencies for data collected in the 2018 NHIS Sample Adult example analytic dataset

Variable Description (VARIABLE NAME)	Frequency
Scheduled medical appointment on internet, past 12 months (HIT3A)	
1=Yes	3,962
2=No	21,163
7=Refused	7
8=Not ascertained	274
9=Don’t know	11
Communicated with health care provider by email, past 12 months (HIT4A)	
1=Yes	4,176
2=No	20,948
7=Refused	7
8=Not ascertained	274
9=Don’t know	12

For continuous or discrete variables with multiple responses (variable: AGE_P), summary statistics can be used to verify the measure of central tendency (mean), measure of dispersion (spread), range, and total number of responses. PROC MEANS is the procedure used for displaying the mean and standard deviation. You can enter the SAS syntax from Box 4.3 to determine the mean and standard deviation for AGE_P in the example analytic dataset.

Box 4.3. SAS procedure (PROC MEANS) for verifying continuous/discrete variable distributions in the 2018 NHIS Sample Adult example analytic dataset



```
PROC MEANS DATA=NHIS.CH4_FILE1;
VAR AGE_P;
RUN;
```

4.2.4 DATA STORAGE

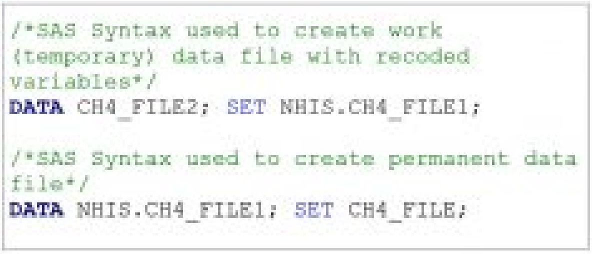
The fourth step is data storage. When using secondary data from publicly available sources, it is important to save the original datasets that you downloaded from the website. This helps ensure that you always have your data in case the website address changes or there is a change in policy that prohibits you from accessing the data at no cost. For restricted data, you must follow the rules for data storage set forth by the agency that owns the restricted data. You may not be able to keep the data over a certain period of time and you may be asked to destroy any outputs with results after the publication of your research.

Each time you make changes to your analytic dataset in SAS, you can save the data using the following two ways:

1. “work” file
2. “permanent” file

The work file is temporary and will only be saved for the current analysis. A limitation of creating a work file is that you will have to re-run the code again each time you use the data if you made any changes to the file (i.e. create recoded variables). However, some benefits to creating a work file are that it will not take up as much space on your computer and the procedures often run faster. Another benefit to saving a permanent data file is that the cleaned, recoded and organized analytic dataset is saved and available for running analyses again after your study is completed. Box 4.4 provides SAS coding for creating a new work file from a permanent data file and a permanent data file from a work file. The permanent file includes a libname in front of the temporary file name (e.g. NHIS.ch4_file1 (permanent) vs. ch4_file (work)).

Box 4.4. SAS programming statements to create work and permanent analytic datasets



```
/*SAS Syntax used to create work
(temporary) data file with recoded
variables*/
DATA CH4_FILE2; SET NHIS.CH4_FILE1;

/*SAS Syntax used to create permanent data
file*/
DATA NHIS.CH4_FILE1; SET CH4_FILE;
```

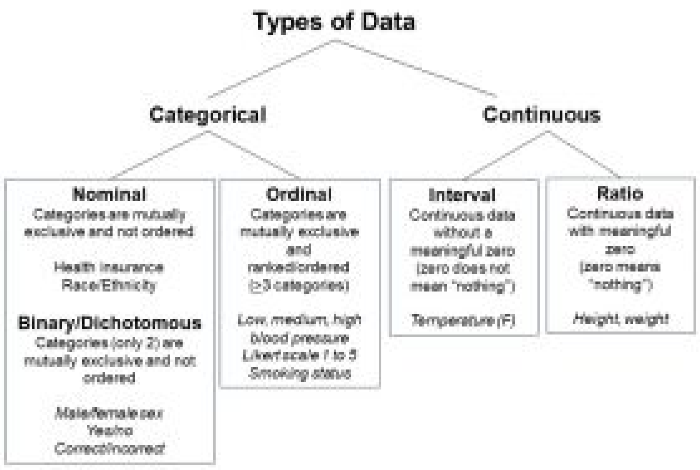
4.3 COMMON STATISTICAL TESTS

Prior to choosing the type of statistical test you must know the type of data collected and the variables you will use to answer your research questions and fulfil your objectives. You must know whether the data are continuous or categorical and specify the independent, dependent, and other contributing variables that will be included in the analysis.

4.3.1 TYPES OF DATA

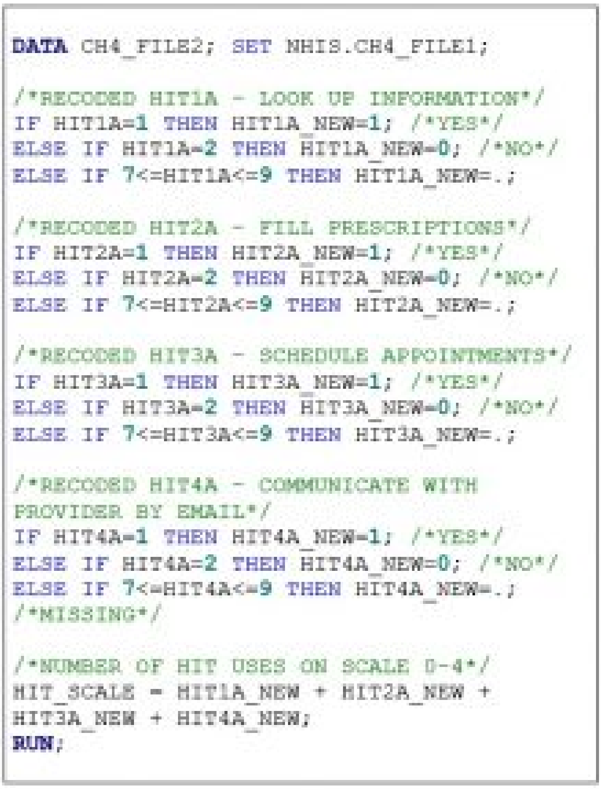
Secondary data sources can include continuous and categorical variables. Continuous variables are numeric responses and can be ratio (with a meaningful zero, such as height) or interval (without a meaningful zero, such as temperature). Continuous responses may be subjective (e.g. self-reported), objective (e.g. clinically measured) or both measurements. Categorical variables include nominal and ordinal responses. Nominal responses include categories that do not have a more or less than relationship. A nominal variable for the example used in this chapter is "REGION." The US region where an individual resides includes the following values: 1=Northeast; 2=Midwest; 3=South; or 4=West. An individual who lives in the West region is not any better than someone who lives in the South region. Binary, also referred to as dichotomous, responses represent variables with only two options. A binary variable for the example used in this chapter is "SEX." The sex of each individual is represented as 1=male and 2=female. There is no more or less than relationship between males and females. Each measure of HIT use can also be represented as a binary variable once the responses for 7=Refused, 8=Not Ascertained, and 9=Don't Know are removed/made missing. Each recoded variable for HIT use with 1=Yes and 0=No responses will then be binary/dichotomous. Ordinal or ranked variables represent data that have a more or less than relationship. Variables must include at least three categories. An ordinal variable for the example in this chapter can be created by adding up all of the HIT uses of participants to determine the total number of HIT uses (0=Does not use any health information technology to 4=Use the internet to look up information, fill prescriptions, schedule appointments, and communicate with a health care provider by email). Other common examples of ordinal variables are Likert scales of agreement (1=strongly disagree to 5=strongly agree) and smoking status (0=never, 1=former, 2=current). An overview of the types of data used for research using national health surveys is provided in Figure 4.1.

Figure 4.1: Flowchart of types of data


Types of Data
Categorical
Continuous
Nominal
Categories are mutually exclusive and not ordered
Health insurance
Race/Ethnicity
Binary/Dichotomous
Categories (only 2) are mutually exclusive and not ordered
Male/female sex
Yes/no
Correct/incorrect
Ordinal
Categories are mutually exclusive and ranked/ordered (≥3 categories)
Low, medium, high blood pressure
Likert scale 1 to 5
Smoking status
Interval
Continuous data without a meaningful zero (zero does not mean "nothing")
Temperature (F)
Ratio
Continuous data with meaningful zero (zero means "nothing")
Height, weight

The type of variable that you need to answer your research question may be different than what is available using the publicly available data files. It is common to recode variables to limit the responses based on how you want the data to be used for answering your research questions. IF/THEN programming statements can be used for recoding prior to statistical analysis. SAS syntax used to create new binary variables and an ordinal HIT usage variable is provided in Box 4.5.

Box 4.5. SAS programming statements (IF/THEN) for recoding secondary data



```
DATA CH4_FILE2; SET NHIS.CH4_FILE1;

/*RECODED HIT1A - LOOK UP INFORMATION*/
IF HIT1A=1 THEN HIT1A_NEW=1; /*YES*/
ELSE IF HIT1A=2 THEN HIT1A_NEW=0; /*NO*/
ELSE IF 7<=HIT1A<=9 THEN HIT1A_NEW=.;

/*RECODED HIT2A - FILL PRESCRIPTIONS*/
IF HIT2A=1 THEN HIT2A_NEW=1; /*YES*/
ELSE IF HIT2A=2 THEN HIT2A_NEW=0; /*NO*/
ELSE IF 7<=HIT2A<=9 THEN HIT2A_NEW=.;

/*RECODED HIT3A - SCHEDULE APPOINTMENTS*/
IF HIT3A=1 THEN HIT3A_NEW=1; /*YES*/
ELSE IF HIT3A=2 THEN HIT3A_NEW=0; /*NO*/
ELSE IF 7<=HIT3A<=9 THEN HIT3A_NEW=.;

/*RECODED HIT4A - COMMUNICATE WITH
PROVIDER BY EMAIL*/
IF HIT4A=1 THEN HIT4A_NEW=1; /*YES*/
ELSE IF HIT4A=2 THEN HIT4A_NEW=0; /*NO*/
ELSE IF 7<=HIT4A<=9 THEN HIT4A_NEW=.;
/*MISSING*/

/*NUMBER OF HIT USES ON SCALE 0-4*/
HIT_SCALE = HIT1A_NEW + HIT2A_NEW +
HIT3A_NEW + HIT4A_NEW;
RUN;
```

In Box 4.5, each HIT use variable has been recoded to a binary variable with "_NEW" added to the end of the original variable name (e.g. HIT4A_NEW for yes/no responses to communicating with a health care provider by e-mail in the past 12 months). A new ordinal variable named HIT_SCALE was created to represent the total number of HIT uses.

4.3.2 TYPES OF VARIABLES

Once the type of data is determined, the independent, dependent and other contributing variables need to be identified. The independent variable in a statistical model is the variable used determine the influence or impact on the outcome.[7] The independent variable is also known as the "predictor" or "exposure" variable. In the 2018 NHIS Sample Adult example, the independent variable for research question 4.1 is region and the independent variable for research question 4.2 is sex. The dependent variable in a statistical model is the variable used as the outcome, for which differences or variations in the dependent variable are being studied.[7] The dependent variable is also known as the "outcome" variable. In the 2018 NHIS Sample Adult example, the dependent variables are each individual HIT usage (looking up health information on the internet, filling prescriptions, scheduling appointments, and communicating with a health care provider by e-mail) and the HIT usage scale (0-4). Other contributing factors, such as confounders and covariates, must also be determined. Confounders are defined as any variables that are causally associated with the dependent variable, not causally or causally associated with the independent variable, but are not intermediate variables in the casual pathway between the independent and dependent variables.[8] Covariates are defined as variables that are potentially related to the dependent variable.[9] This term is often used to represent any contributing or explanatory factor that may bias the results. Covariates can be adjusted for during statistical analysis to reduce bias. In the 2018 NHIS Sample Adult example, a covariate is age. Other covariates often controlled for in statistical analysis include risk factors, social determinants of health, and health behaviors. Covariates may differ based on the independent and dependent variables of interest in your study.

4.3.3 SAS PROCEDURES FOR COMMON STATISTICAL TESTS

Common statistical tests used for categorical data analysis with national health surveys include chi square, Wilcoxon rank sum (also known as Mann Whitney U) tests, and logistic regression tests.

4.3.3.a Chi Square

The Pearson chi square test is used to compare categorical independent variables and categorical dependent variables that are binary/dichotomous or nominal.[7] Results pertaining to the first four research objectives mentioned in section 4.1 can be calculated by using chi square tests. As a reminder, the first four research objectives are:

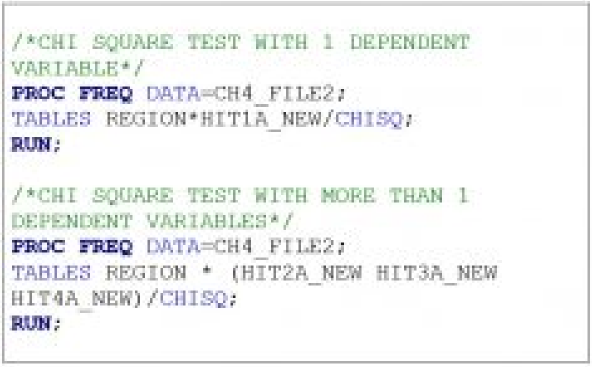
- Objective 4.1: To determine the association between region (independent variable) and looking up health information on the internet (dependent variable)
- Objective 4.2: To determine the association between region (independent variable) and filling prescriptions on the internet (dependent variable)
- Objective 4.3: To determine the association between region (independent variable) and scheduling appointments on the internet (dependent variable)

- Objective 4.4: To determine the association between region (independent variable) and communicating with a health care provider by e-mail (dependent variable)

The independent (predictor or exposure) and dependent (outcome) variables are identified in each objective. PROC FREQ is the procedure used to display the crosstabulation of variables and calculation of the chi square test result. To calculate the chi square test result, you must list a "*" between the independent and dependent variable and add "/CHISQ" at the end of the statement that begins with "tables." You can calculate the chi square test for each dependent variable separately or you can run then together.

You can enter the SAS syntax from Box 4.6 to determine the chi square test results.

Box 4.6. SAS procedure (PROC FREQ with /CHISQ) for running chi square test using the 2018 NHIS Sample Adult example analytic dataset



```
/*CHI SQUARE TEST WITH 1 DEPENDENT
VARIABLE*/
PROC FREQ DATA=CH4_FILE2;
TABLES REGION*HIT1A_NEW/CHISQ;
RUN;

/*CHI SQUARE TEST WITH MORE THAN 1
DEPENDENT VARIABLES*/
PROC FREQ DATA=CH4_FILE2;
TABLES REGION * (HIT2A_NEW HIT3A_NEW
HIT4A_NEW)/CHISQ;
RUN;
```

4.3.3.b Wilcoxon Rank Sum/Mann-Whitney U Test

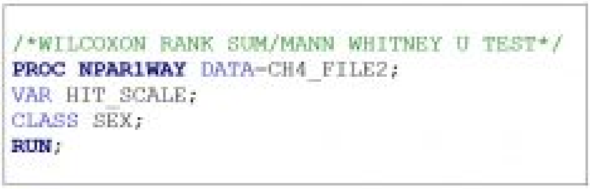
The Wilcoxon Rank Sum (also known as the Mann-

Whitney U) test is used to compare differences in the medians of ordinal dependent variables among binary/ dichotomous independent variables.[7] The median is the center of observations when listed in rank order. The independent variable must only have two categories. For two or more categories, the Kruskal-Wallis (H) test should be used instead.[7] In SAS 9.4, the coding statements for both tests are the same. The Wilcoxon Rank Sum test can be used to obtain results for research objective 4.5 in the 2018 NHIS Sample Adult example. As a reminder:

- Objective 4.5: To determine association between sex (independent variable) and the number of HIT usages (dependent variable)

The independent (predictor or exposure) and dependent (outcome) variables are identified in the objective. PROC NPAR1WAY is the procedure used to calculate the Wilcoxon Rank Sum test result. The SAS procedure for PROC NPAR1WAY is similar to PROC MEANS. Instead of "tables" in the second line, you will use the statement "VAR" before the dependent variable (HIT_SCALE). The independent variable (SEX) is listed next to a "CLASS" statement on the next line. You can enter the SAS syntax from Box 4.7 to determine the Wilcoxon Rank Sum test result.

Box 4.7. SAS procedure (PROC NPAR1WAY) for the Wilcoxon Rank Sum test in the 2018 NHIS Sample Adult example analytic dataset



```
/*WILCOXON RANK SUM/MANN WHITNEY U TEST*/
PROC NPAR1WAY DATA=CH4_FILE2;
VAR HIT_SCALE;
CLASS SEX;
RUN;
```

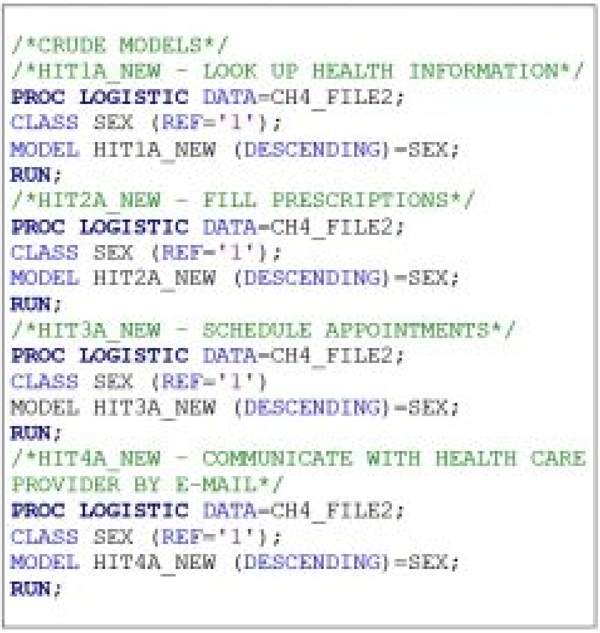
4.3.3.c Logistic Regression

Regression modeling is used to explore relations between independent (predictor) variables and dependent (outcome) variables before and after adjusting for other contributing factors that may bias the results. Logistic regression models are used when the dependent (outcome) variable is binary/dichotomous.[7] Regression models that do not adjust for other contributing factors are called “crude” or “unadjusted” models. Regression models that do adjust for other contributing factors are called “multivariable” or “adjusted” models. Crude and multivariable logistic regression models can be used to obtain results for research objective 4.6 in the 2018 NHIS Sample Adult example. As a reminder:

- Objective 4.6: To determine the associations between sex (independent variable) and each HIT usage — looking up health information online, filling prescriptions, scheduling appointments, and communicating with a health care provider by e-mail — (dependent variables) before and after controlling for other contributing factors

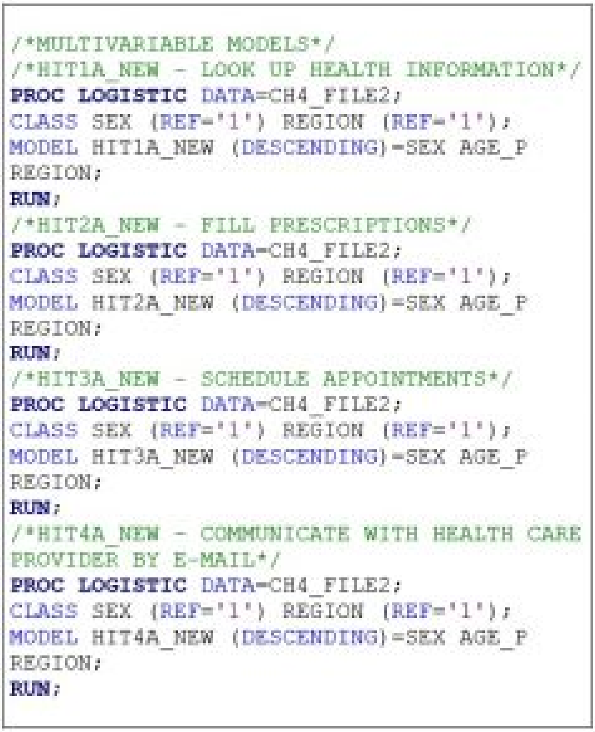
The independent (predictor or exposure) and dependent (outcome) variables are identified in the objective. The contributing factors that will be controlled for to reduce bias are region and age. PROC LOGISTIC is the procedure used to calculate the results. Enter the SAS syntax from Box 4.8 and Box 4.9 to determine the crude and multivariable results, respectively. The CLASS statement identifies the reference or comparison group for the categorical independent variable and other covariates. For the variable SEX, comparisons are made between females and males. In the analytic dataset, males are represented by "1" and females are represented by "2." Therefore, the reference category is "1" so that our results will represent the odds of females using HIT in comparison to males. The MODEL statement is set up as "dependent variable = independent variable (+ covariates for multivariable analysis)." The DESCENDING statement after the dependent variable identifies the higher value as the outcome of interest. For each HIT use, "1=yes" should be the outcome that is modeled. Therefore, the results will indicate the odds of using HIT instead of the odds of not using HIT.

Box 4.8. SAS procedure (PROC LOGISTIC) for crude logistic regression results in the 2018 NHIS Sample Adult example



```
/*CRUDE MODELS*/
/*HIT1A_NEW - LOOK UP HEALTH INFORMATION*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1');
MODEL HIT1A_NEW (DESCENDING)=SEX;
RUN;
/*HIT2A_NEW - FILL PRESCRIPTIONS*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1');
MODEL HIT2A_NEW (DESCENDING)=SEX;
RUN;
/*HIT3A_NEW - SCHEDULE APPOINTMENTS*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1')
MODEL HIT3A_NEW (DESCENDING)=SEX;
RUN;
/*HIT4A_NEW - COMMUNICATE WITH HEALTH CARE
PROVIDER BY E-MAIL*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1');
MODEL HIT4A_NEW (DESCENDING)=SEX;
RUN;
```

Box 4.9. SAS Procedure (PROC LOGISTIC) for multivariable logistic regression results in the 2018 NHIS Sample Adult example



```
/*MULTIVARIABLE MODELS*/
/*HIT1A_NEW - LOOK UP HEALTH INFORMATION*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1') REGION (REF='1');
MODEL HIT1A_NEW (DESCENDING)=SEX AGE_P
REGION;
RUN;
/*HIT2A_NEW - FILL PRESCRIPTIONS*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1') REGION (REF='1');
MODEL HIT2A_NEW (DESCENDING)=SEX AGE_P
REGION;
RUN;
/*HIT3A_NEW - SCHEDULE APPOINTMENTS*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1') REGION (REF='1');
MODEL HIT3A_NEW (DESCENDING)=SEX AGE_P
REGION;
RUN;
/*HIT4A_NEW - COMMUNICATE WITH HEALTH CARE
PROVIDER BY E-MAIL*/
PROC LOGISTIC DATA=CH4_FILE2;
CLASS SEX (REF='1') REGION (REF='1');
MODEL HIT4A_NEW (DESCENDING)=SEX AGE_P
REGION;
RUN;
```

4.4. SUMMARY

This chapter provided an overview of the steps for preparing secondary data for statistical analysis and how to run common statistical tests in SAS 9.4 using primarily categorical data. The examples used in this chapter are from the 2018 NHIS Sample Adult public-use data files. The examples in this chapter focused on determining associations between demographic factors (region, sex) and HIT usage (looking up health information online, filling prescriptions, scheduling appointments, and communicating with health care providers by e-mail) among US adults. SAS 9.4 analysis procedures were demonstrated for running descriptive statistics, recoding variables, and conducting comparative statistical analyses with chi square tests, Wilcoxon Rank Sum tests, and logistic regression models. As previously mentioned, the examples provided in this chapter are not weighted and do not include procedures for adjusting the results based on the complex survey design. The use of complex sample design features for national health data will be described in Chapter 5.

4.5 REFERENCES

1. National Center for Health Statistics. National Health Interview Survey, 2011-2015. Public-use data file and documentation. Published August 5, 2020. Accessed August 10, 2020. https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm
2. Kindratt T, Callender L, Cobbaert M, Wondrack J, Bandiera F, Salvo D. Health information technology use and influenza vaccine uptake among US adults. *Int J Med Inform*. 2019;129:37-42. doi:10.1016/j.ijmedinf.2019.05.025
3. Kindratt TB, Allicock M, Atem F, Dallo FJ, Balasubramanian BA. Email Patient-Provider Communication and Cancer Screenings Among US Adults: Cross-sectional Study. *JMIR Cancer*. 2021;7(3):e23790. doi:10.2196/23790
4. Elliott AC, Hynan LS, Reisch JS, Smith JP.

Preparing data for analysis using Microsoft Excel. *J Investig Med*. 2006;54(6):334-341. doi:10.2310/6650.2006.05038

5. Kindratt TB. Research Extension Experience in Directed Studies: Solidifying Evidence-Based Medicine Competencies Through Research Participation. *J Physician Assist Educ*. 2020;31(1):36-41. doi:10.1097/JPA.0000000000000291
6. Dehaven MJ, Gimpel NE, Dallo FJ, Billmeier TM. Reaching the underserved through community-based participatory research and service learning: description and evaluation of a unique medical student training program. *J Public Health Manag Pract*. 2011;17(4):363-368. doi:10.1097/PHH.0b013e3182214707
7. Jacobsen KH. *Introduction to Health Research Methods: A Practical Guide*. 2nd ed. Jones & Bartlett Learning; 2017.
8. Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*. 3rd ed. Jones & Bartlett Learning; 2014.
9. Field A. *Discovering Statistics Using IBM SPSS Statistics*. 4th ed. SAGE; 2013.

CHAPTER 5.

COMPLEX SURVEY DESIGN FEATURES

5.1 OVERVIEW

In Chapter 4, you learned the steps for preparing secondary data for statistical analysis and how to run common statistical tests using examples from the 2018 National Health Interview Survey (NHIS) Sample Adult public-use data. SAS 9.4 programming statements were provided to run summary statistics (frequencies, means and standard deviations), chi square tests, Wilcoxon Rank Sum/Mann Whitney U tests, and crude and multivariable logistic regression models. In this chapter, you will learn how to run statistical tests using special procedures designed to account for the complex sample designs used for national health surveys. In order to produce representative estimates using national health surveys, variables for the cluster, stratum, and weighting variables must be included using SAS SURVEY procedures. The examples in this chapter will use 2018 NHIS Sample Adult data to answer the following research questions.

- Research Question 5.1: What are associations between region and health information technology (HIT) usage among adults ages 18 and older in the United States (US)?

- Research Question 5.2: What are associations between sex and HIT usage among adults ages 18 and older in the US?

The research objectives are:

- Objective 5.1: To determine the association between region and looking up health information on the internet
- Objective 5.2: To determine the association between region and filling prescriptions on the internet
- Objective 5.3: To determine the association between region and scheduling appointments on the internet
- Objective 5.4: To determine the association between region and communicating with health care provider by e-mail
- Objective 5.5: To determine the associations between sex and each HIT usage (looking up health information online, filling prescriptions, scheduling appointments, and communicating with a health care provider by e-mail) before and after controlling for other contributing factors

SAS 9.4 SURVEY procedures will be used to demonstrate how to meet the research objectives. It is important to note that Objective 4.5 from Chapter 4 has been removed from this chapter because there are no corresponding survey procedures for the Wilcoxon Rank Sum/Mann Whitney U test at the time of this writing.

5.2 COMPLEX SURVEY DESIGN FEATURES

The secondary analysis of national health surveys involves using advanced survey-based statistical procedures to account for the sophisticated sampling designs. Stratification, clustering, and weighting techniques must be used with Taylor Series Linearization methods, which are outlined in the data analytic recommendations for each survey. Table 5.1 provides some examples of the complex design variables for the national health surveys included in this textbook.

TABLE 5.1. SELECTED CLUSTERING, STRATIFICATION, AND WEIGHTING VARIABLES USED FOR ANALYZING COMPLEX NATIONAL HEALTH SURVEYS

Survey	Stratification	Clustering	Weighting
National Health Interview Survey (NHIS)	PSTRAT	PPSU	WTFA_SA
Medical Expenditure Panel Survey (MEPS)	VARSTR	VARPSU	PERWTF18F
Health Information National Trends Survey (HINTS)	VAR_STRATUM	VAR_CLUSTER	TG_all_FINWT0
Behavior Risk Factor Surveillance Survey (BRFSS)	_STSTR	_PSU	_LLCP_WGT
National Health and Nutrition Examination Survey (NHANES)	SDMVSTRA	SDMVPSU	WTINT2YR

5.2.1 STRATIFICATION

National health surveys use stratification methods to draw a sample from the larger sampling frame of the population. The large sampling frame is divided into mutually exclusive strata and the sample is then selected from each stratum.[1] Lewis (2017) states the following three reasons for using stratification in complex national health surveys: 1) to improve representation of smaller subgroups in the population; 2) to allow for multiple modes of data collection; and 3) to improve precision of statistical estimates.[1] First, stratification allows for oversampling of subgroups by race/ethnicity and among states with smaller populations. For example, the NHIS oversampled underrepresented minority populations (non-Hispanic Black, Hispanic, and non-Hispanic Asian) from 2006-2015.[2] In 2016, the NHIS sample design was modified to collect larger samples in smaller states. Sample sizes were increased in the 10 least populous states and Washington D.C.[2] Second, stratification allows for the use of random-digit-dialing, in-person, direct mailing, and online data collection methods to be conducted in a systematic manner. The Health Information National Trends Survey (HINTS) has used several data collection methods, including random digit dialing, direct mailing, and the recently adopted web-based data collection procedures.[3] Finally, the results will be more precise with homogeneous strata collected in relation to the outcome variable.[1]

5.2.2 CLUSTERING

National health surveys use clustering methods to gain large samples while reducing data collection costs. Clusters are identified in national health survey documentation as primary, secondary, and tertiary sampling units.[1] Clustering allows for national health surveys to collect data on multiple individuals from each household, multiple patients attending one clinic, or multiple students from one school.[1] At the state level, primary sampling unit clusters are created to select samples from large metropolitan statistical areas and counties. Within each cluster, census housing blocks are selected as secondary sampling units. Within each census block, households are selected as tertiary sampling units. For the NHIS, self-reported interview data are collected from all individuals in each household then a designated sample adult and sample child are selected from households to answer additional questions on health conditions and behaviors.[4] Complex national health survey analysis procedures usually require using clustering to account for the primary sampling units.

5.2.3 WEIGHTING

Sample weights are used to account for the underrepresentation or overrepresentation of each individual in the sample. When combining multiple years of national health surveys, the weight needs to be divided by the total years combined. Specific details and formulas on how sampling weights are calculated are provided in the analytic documentation for each complex national health survey.[5]

5.3 SAS SURVEY PROCEDURES

To complete the following SAS SURVEY examples, data from the 2018 NHIS Sample Adult file and Sample SAS statements can be downloaded from the NHIS data release website.

Complete the following:

- Go to your "C:\" drive and create a folder named "NHIS"
- In the "NHIS" folder, create a folder named "18"
- Download the 2018 Sample Adult ASCII data file (.dat) from the 2018 NHIS data release website

- Unzip the file and save it in the folder “C:\NHIS\18”
- Download “File 5.1 Sample SAS Program to Create NHIS 2018 Sample Adult File” from the Open ICPSR data repository
- Open the file and select “Run” from the top menu bar

Your analytic dataset should include the following 12 variables:

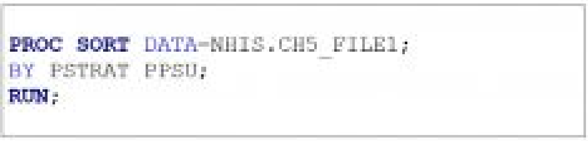
1. SRVY_YR
2. FPX
3. AGE_P
4. SEX
5. REGION
6. HIT1A
7. HIT2A
8. HIT3A
9. HIT4A
10. PPSU
11. PSTRATUM
12. WTFA_SA

The analytic dataset should include 3 more variables (PPSU, PSTRATUM, WTFA_SA) than those collected in Chapter 4. To determine the variables in the analytic dataset, run a PROC CONTENTS statement. An example of a PROC CONTENTS statement is provided in Box 4.1

in Chapter 4. If you are unable to create the permanent analytic dataset using the preceding steps, you can go to the Open ICPSR data repository, download "File 5.3 Permanent SAS Analytic Database," and save the database in the folder "C:\NHIS\18."

Prior to conducting survey procedures, complex survey data must be sorted by stratum and primary sampling unit variables. You can enter the SAS syntax from Box 5.1 to sort the data.

BOX 5.1. SAS PROCEDURE FOR SORTING ANALYTIC DATASET BY STRATUM AND PRIMARY SAMPLING UNIT

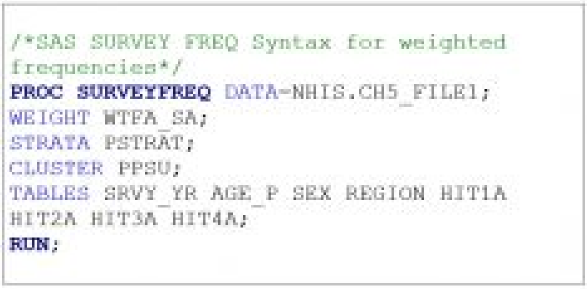


```
PROC SORT DATA=NHIS.CH5_FILE1;
BY PSTRAT PPSU;
RUN;
```

5.3.1 SURVEY PROCEDURES FOR FREQUENCIES AND PERCENTAGES

PROC SURVEYFREQ is the procedure for displaying weighted frequencies and percentages in SAS for complex surveys. Primary sampling unit, stratum, and weighting variables must be included in the programming statements. Enter the SAS syntax from Box 5.2 to determine the weighted frequencies for all variables in the analytic dataset. Variable FPX is excluded because it represents the ID number of the participant.

Box 5.2. SAS SURVEY procedure (PROC SURVEYFREQ) for determining frequencies in the 2018 NHIS Sample Adult analytic dataset



```
/*SAS SURVEY FREQ Syntax for weighted
frequencies*/
PROC SURVEYFREQ DATA=NHIS.CH5_FILE1;
WEIGHT WTFA_SA;
STRATA PSTRAT;
CLUSTER PPSU;
TABLES SRVY_YR AGE_P SEX REGION HIT1A
HIT2A HIT3A HIT4A;
RUN;
```

The weighted frequencies for the variables collected in this example are presented in Table 5.2. Although the NHIS does not publish the weighted frequencies in the dataset documentation on their website, other surveys, such as the Medical Expenditure Panel Survey (MEPS) and HINTS include the weighted results for data analysts to verify their outputs are correct prior to conducting additional analytic procedures.

Table 5.2. Weighted frequencies for data collected in the 2018 NHIS Sample Adult Example analytic dataset

Variable Description (VARIABLE NAME)	Weighted Frequency
Survey Year (SRVY_YR)	
2018	249,455,533
Sex (SEX)	
1=Male	120,441,598
2=Female	129,013,935
Region (REGION)	
1=Northeast	43,261,774
2=Midwest	54,817,888
3=South	92,043,276
4=West	59,332,595
Looked up health information on internet, past 12 months (HIT1A)	
1=Yes	136,281,936
2=No	110,137,528
7=Refused	87,014
8=Not ascertained	2,655,609
9=Don’t know	293,446
Filled a prescription on internet, past 12 months (HIT2A)	
1=Yes	28,308,262
2=No	218,302,863
7=Refused	59,192
8=Not ascertained	2,666,760
9=Don’t know	118,456

Table 5.2 (continued). Weighted frequencies for data collected in

the 2018 NHIS Sample Adult Example analytic dataset

Variable Description (VARIABLE NAME)	Weighted Frequency
Scheduled medical appointment on internet, past 12 months (HIT3A)	
1=Yes	41,617,782
2=No	204,932,293
7=Refused	67,681
8=Not ascertained	2,666,760
9=Don't know	171,017
Communicated with health care provider by email, past 12 months (HIT4A)	
1=Yes	41,094,984
2=No	205,459,751
7=Refused	67,681
8=Not ascertained	2,666,760
9=Don't know	166,357

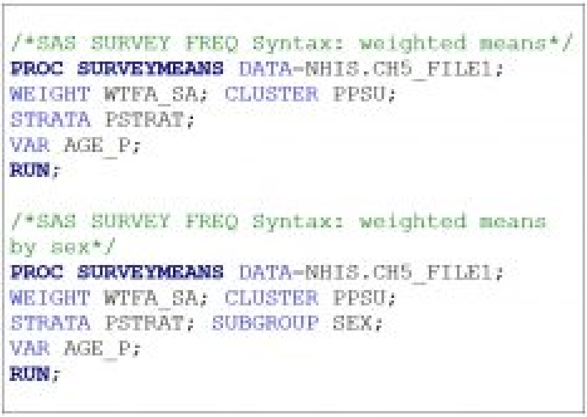
The variable for age (AGE_P) is excluded from the table due to multiple response options (age 18-85 years and older).

5.3.2 SURVEY PROCEDURES FOR MEANS

For continuous or discrete variables with multiple responses (variable: AGE_P), summary statistics can be used to verify the measure of central tendency (mean) and total responses. PROC SURVEYMEANS is the procedure used for displaying the mean and can be separated by subgroups such as region and sex. Enter the SAS syntax from Box 5.3 to determine the means for AGE_P collectively and separated by sex in the analytic dataset.

Box 5.3. SAS SURVEY procedure (PROC SURVEYMEANS) for

verifying continuous/discrete variable distributions in the 2018 NHIS Sample Adult analytic dataset



```
/*SAS SURVEY FREQ Syntax: weighted means*/
PROC SURVEYMEANS DATA=NHIS.CH5_FILE1;
WEIGHT WTFA_SA; CLUSTER PPSU;
STRATA PSTRAT;
VAR AGE_P;
RUN;

/*SAS SURVEY FREQ Syntax: weighted means
by sex*/
PROC SURVEYMEANS DATA=NHIS.CH5_FILE1;
WEIGHT WTFA_SA; CLUSTER PPSU;
STRATA PSTRAT; SUBGROUP SEX;
VAR AGE_P;
RUN;
```

5.3.3 SURVEY PROCEDURES FOR CHI SQUARE TESTS

Results pertaining to the first four research objectives mentioned in section 5.1 can be calculated by using SAS SURVEY procedures for chi square tests. IF/THEN procedures used to remove “refused,” “not ascertained,” and “don’t know” responses for the dependent variables are described in Section 4.3.1. SAS coding statements are provided in Box 4.5.

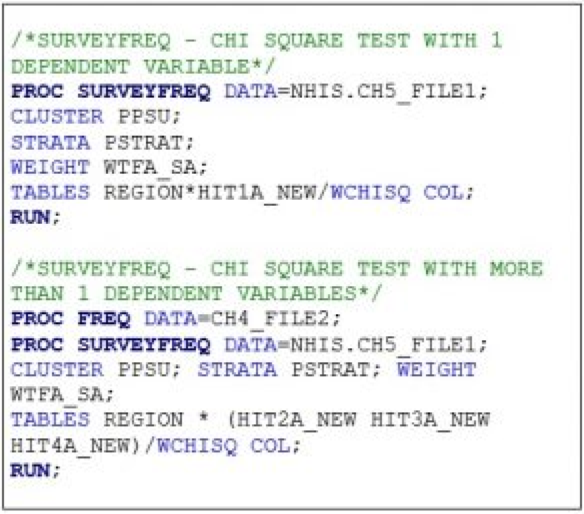
As a reminder, the first four research objectives are:

- Objective 5.1: To determine the association between region (independent variable) and looking up health information on the internet (dependent variable)
- Objective 5.2: To determine the association between region (independent variable) and filling prescriptions on the internet (dependent variable)

- Objective 5.3: To determine the association between region (independent and scheduling appointments on the internet
- Objective 5.4: To determine the association between region and communicating with health care provider by e-mail

The independent (predictor or exposure) and dependent (outcome) variables are identified in each objective. PROC SURVEYFREQ is the procedure used to display the crosstabulation of variables and calculation of the chi square test result. You must add "/WCHISQ" at the end of the statement that begins with "tables." You can calculate the chi square test for each dependent variable separately or you can run then together. You can enter the SAS syntax from Box 5.3 to determine the chi square test results.

Box 5.4. SAS SURVEY Procedure (PROC SURVEYFREQ with /WCHISQ) for running chi Square tests using 2018 NHIS Sample Adult analytic dataset



```
/*SURVEYFREQ - CHI SQUARE TEST WITH 1
DEPENDENT VARIABLE*/
PROC SURVEYFREQ DATA=NHIS.CH5_FILE1;
CLUSTER PPSU;
STRATA PSTRAT;
WEIGHT WTFA_SA;
TABLES REGION*HIT1A_NEW/WCHISQ COL;
RUN;

/*SURVEYFREQ - CHI SQUARE TEST WITH MORE
THAN 1 DEPENDENT VARIABLES*/
PROC FREQ DATA=CH4_FILE2;
PROC SURVEYFREQ DATA=NHIS.CH5_FILE1;
CLUSTER PPSU; STRATA PSTRAT; WEIGHT
WTFA_SA;
TABLES REGION * (HIT2A_NEW HIT3A_NEW
HIT4A_NEW)/WCHISQ COL;
RUN;
```

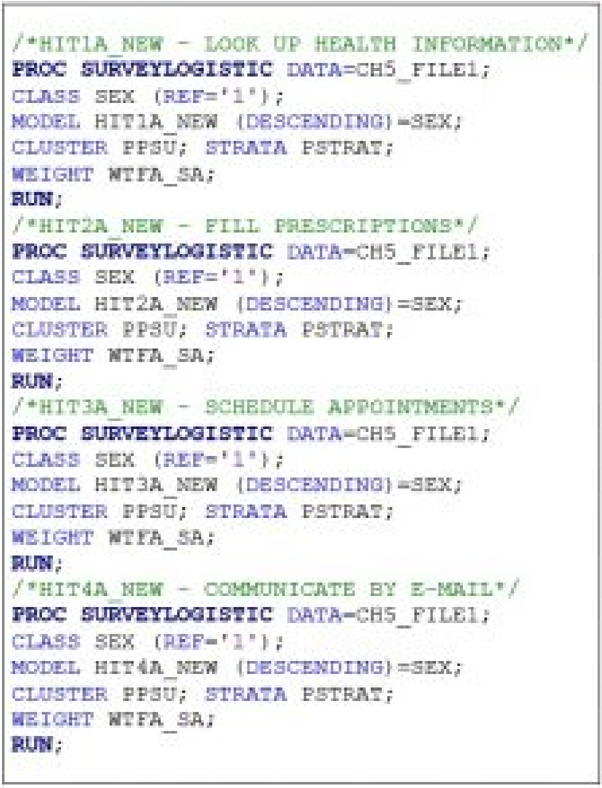
5.3.4 SURVEY PROCEDURES FOR LOGISTIC REGRESSION

As stated in Chapter 4, crude and adjusted logistic regression models are used when the dependent (outcome) variable is binary/dichotomous.[6] Crude and multivariable logistic regression models can be used to obtain results for research objective 5.5 in this chapter's 2018 NHIS Sample Adult example. As a reminder:

- Objective 5.5: To determine the associations between sex (independent variable) and each HIT usage — looking up health information online, filling prescriptions, scheduling appointments, and communicating with a health care provider by e-mail — (dependent variables) before and after controlling for other contributing factors

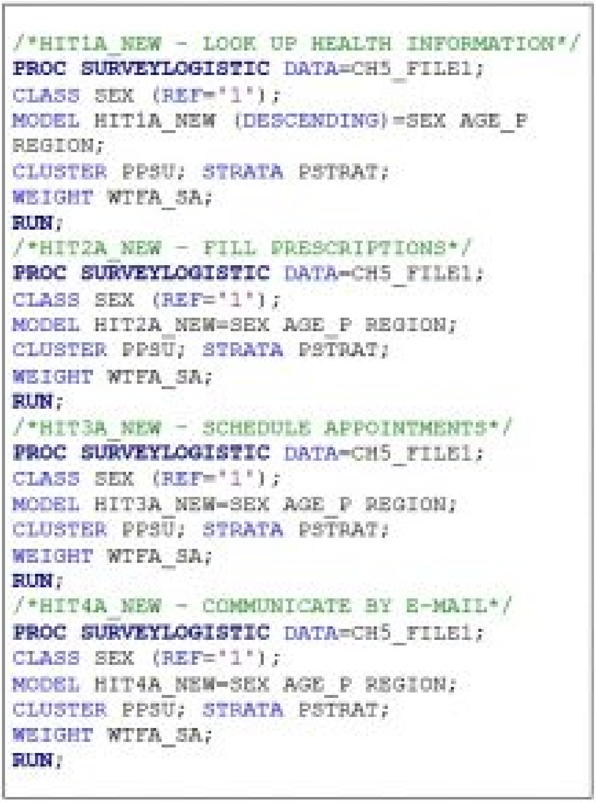
The independent (predictor or exposure) and dependent (outcome) variables are identified in the objective. In this example, the contributing factors that will be controlled for to reduce bias are region and age. PROC SURVEYLOGISTIC is the procedure used to calculate the results after adjusting for the stratum, cluster and weight variables. You can enter the SAS syntax provided to determine the crude (Box 5.5) and multivariable (Box 5.6) results.

Box 5.5. SAS SURVEY procedure (PROC SURVEYLOGISTIC) for crude logistic regression results in the 2018 NHIS Sample Adult analytic dataset



```
/*HIT1A_NEW - LOOK UP HEALTH INFORMATION*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT1A_NEW (DESCENDING)=SEX;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
/*HIT2A_NEW - FILL PRESCRIPTIONS*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT2A_NEW (DESCENDING)=SEX;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
/*HIT3A_NEW - SCHEDULE APPOINTMENTS*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT3A_NEW (DESCENDING)=SEX;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
/*HIT4A_NEW - COMMUNICATE BY E-MAIL*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT4A_NEW (DESCENDING)=SEX;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
```

Box 5.6. SAS SURVEY procedure (PROC SURVEYLOGISTIC) for multivariable logistic regression results in the 2018 NHIS Sample Adult analytic dataset



```
/*HIT1A_NEW - LOOK UP HEALTH INFORMATION*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT1A_NEW (DESCENDING)=SEX AGE_P
REGION;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
/*HIT2A_NEW - FILL PRESCRIPTIONS*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT2A_NEW=SEX AGE_P REGION;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
/*HIT3A_NEW - SCHEDULE APPOINTMENTS*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT3A_NEW=SEX AGE_P REGION;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
/*HIT4A_NEW - COMMUNICATE BY E-MAIL*/
PROC SURVEYLOGISTIC DATA=CH5_FILE1;
CLASS SEX (REF='1');
MODEL HIT4A_NEW=SEX AGE_P REGION;
CLUSTER PPSU; STRATA PSTRAT;
WEIGHT WTFA_SA;
RUN;
```

The CLASS statement identifies the reference or comparison group for the categorical independent variable and other covariates. For the variable SEX, females will be compared to males. In the analytic dataset, males are represented by “1” and females are represented by “2.” Therefore, the reference category is “1” so that our results will represent the odds of females using HIT in comparison to males. The MODEL statement is set up as “dependent variable = independent variable (+ covariates for multivariable analysis)." The DESCENDING statement after the dependent variable identifies the higher value as the outcome of interest. For each HIT usage, "1=yes" will be the outcome that is modeled. Therefore, our results will indicate the odds of using health information technology instead of the odds of not using health information technology.

5.4 SUMMARY

This chapter provided an overview of complex design features of national health surveys and how to run SAS SURVEY procedures to determine national estimates using primarily categorical data. The examples used in this chapter come from the 2018 NHIS Sample Adult public-use data files and focused on determining associations between demographic factors (region, sex) and HIT uses (look up health information, fill prescriptions, schedule appointments, and communicate with health care providers by e-mail) among US adults. SAS SURVEY procedures were demonstrated for running frequencies, means, chi square tests, and logistic regression models that account for the stratum, cluster, and weight variables.

5.5. REFERENCES

1. Lewis TH. *Complex Survey Data Analysis with SAS*. CRC Press, Taylor & Francis Group; 2017.
2. Blewett LA, Dahlen HM, Spencer D, Rivera Drew JA, Lukanen E. Changes to the Design of the National Health Interview Survey to Support Enhanced Monitoring of Health Reform Impacts at the State Level. *Am J Public Health*.

2016;106(11):1961-1966. doi:10.2105/AJPH.2016.303430

3. Finney Rutten LJ, Blake KD, Skolnick VG, Davis T, Moser RP, Hesse BW. Data Resource Profile: The National Cancer Institute's Health Information National Trends Survey (HINTS). *Int J Epidemiol*. 2020;49(1):17-17j. doi:10.1093/ije/dyz083
4. National Center for Health Statistics. National Health Interview Survey, 2011-2015. Public-use data file and documentation. Published August 5, 2020. Accessed August 10, 2020. https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm
5. Parsons V, Moriarity C, Jonas K. *Design and Estimation for the National Health Interview Survey, 2006–2015.*; 2014:53. https://www.cdc.gov/nchs/data/series/sr_02/sr02_165.pdf
6. Jacobsen KH. *Introduction to Health Research Methods: A Practical Guide*. 2nd ed. Jones & Bartlett Learning; 2017.

CHAPTER 6.

NATIONAL HEALTH INTERVIEW SURVEY

6.1 INTRODUCTION

Chapter 6 covers the National Health Interview Survey (NHIS). The NHIS has been collected by the National Center for Health Statistics (NCHS) since 1957 to monitor and explore trends in the health status and health care utilization among the US population.[1] This chapter includes details on: how data are collected; how data are made publicly available as machine-actionable data files; what variables must be included to address design features of the complex sample; the strengths and limitations of the survey; and practical tips for conducting statistical analysis; and how to answer research questions using a case study. The practical tips provided for analysis of NHIS data are based on the primary author's previous experiences analyzing NHIS data from 2000-2018 to answer questions related examining the associations between predisposing and enabling factors that contribute to health behaviors, morbidity, mortality and health services use. The NHIS case study will explore whether Arab American adults are more or less likely to receive an annual flu vaccine in comparison to other racial/ethnic groups, such as other non-Hispanic Whites. The bulk of this chapter will comprise of section *6.6: NHIS Case Study* in order to give the reader hands-on practice downloading and cleaning large databases and conducting basic categorical data analysis using PROC SURVEYFREQ and PROC SURVEYLOGISTIC. The syntax provided was created for use with SAS 9.4.

6.2 DATA COLLECTION

The NHIS uses a cross-sectional study design to collect face-to-face household interviews from the US civilian, non-institutionalized population to produce national health estimates. The sample design excludes individuals living in correctional facilities (e.g. prisons), long-term care institutions (e.g. nursing homes), military personnel, and US nationals living overseas.[1] The sample design is modified every ten years after the decennial census.[1] Prior to 2016, the NHIS used a multi-stage area probability design. The sample was drawn from over 400 primary sampling units (PSU) (counties, small groups of counties, or metropolitan statistical areas) covering all 50 states and Washington, D.C. Within each PSU, addresses were sampled and underrepresented minority groups (non-Hispanic Black, Hispanic, and non-Hispanic Asian) were oversampled. In 2016, the NHIS sample design was modified to provide more robust estimates for state-level analysis. The multi-stage process was removed and over 300 clusters located within the boundaries of each state were used as sampling units.[1] Instead of oversampling racial and ethnic minority groups, the new design oversamples certain populations within each state to increase sample sizes among less populous states.[2] Sample sizes have been increased in the 10 least populous states and Washington D.C. and decreased in the 40 most populous states.[2] In 2019, a redesigned version of the NHIS was implemented to reduce response burden for participants, improve coverage of health topics, streamline overlapping content with other national health surveys, develop a long-term plan for periodic and ongoing topics, and incorporate new methodologies.[2,3] Data for 2019 were not available as of this writing. Further details of the NHIS sampling design and data collection methods are reported on the NHIS website.[4]

6.3 DATA FILES

The NHIS is comprised of two main components: 1) core questions that are consistently collected on an annual basis and 2) an assortment of supplements sponsored by other agencies outside of NCHS, such as National Heart, Lung, and Blood Institute, Center for Tobacco Products, and National Center for Environmental Health. Core questionnaire sections include the family core (which includes the household composition and person level file), sample adult core and sample child core. In 2018, supplemental questions focused on food security, health care access and utilization among families, asthma, occupational health, cancer screening, functioning and disability (including cognitive disability), immunizations, e-cigarettes and other tobacco use, heart disease and stroke among adults, and asthma and immunizations among children.[1]

6.3.1 HOUSEHOLD FILE

One adult member of each household provides responses to the household composition section of the NHIS questionnaire. The participant provides relationship information and demographics of all members of the household to determine the number of families in each housing unit.[1] Each household has a unique identifier, represented by variable HHX, which is used for merging data files within and across years. The household data file includes questions assessing the type of living quarters (e.g. house, apartment, mobile home), number of families, number of persons, and the US region (Northeast, Midwest, South, West) where the household is located.[1]

6.3.2 FAMILY FILE

One adult member of each family within each household is identified to answer questions about adult and child members of the family. Any adults within the household not designated as the "family respondent" can respond for themselves.[1] Each family is defined by groups of two or more related persons living in the same household or unrelated individuals living together, such as unmarried couples.[1] Each family has a unique identifier, represented by variable FMX, which is used for merging data files within and across years. The family data file includes topics such as telephone use (landline, cell phones), family type and structure, family member disabilities, family level income, health insurance, and government assistant programs (e.g. Women's, Infants and Children [WIC]).

6.3.3 PERSON FILE

Person level data are collected as part of the family core component for each household. Person level data are collected from all adults and children in the household. Each person has a unique identifier (variable name: PX years 2000-2003; FPX years 2004 onward) which is used for merging data files within and across years. The person level data file includes topics such as demographics, socioeconomics, health status, limitations of activity, health care access and utilization, health insurance and English language proficiency. Race and ethnicity are collected in accordance with the 1997 Office of Management and Budget standards.[5]

6.3.4 SAMPLE ADULT

One adult per family is randomly selected as a sample adult. There is an increased likelihood of being selected if the adult is non-Hispanic Black, non-Hispanic Asian, Hispanic, or ages 65 years or older. The sample adult responds to questions about their own health for this section. If the participant is not mentally or physically able to do so, questions about their health will be asked to a proxy.[1] Individuals represented as the sample adult in each household are identified in the person level file with a "flag" (variable: ASTATFLG). The sample adult file includes individual-level topics such as demographics, socioeconomics, adult health conditions, adult health status and activity limitations, health behaviors, health care access and utilization, and additional selected items. Additional selected items cover a wide range of health and social determinants of health topics such as frequency of computer use, satisfaction with health care received, perceived neighborhood social cohesion, sexual orientation and HIV testing.

6.3.5 SAMPLE CHILD

One child per family is randomly selected as a sample child. Information about the sample child is collected from an adult in the family that has the most knowledge about the child's health.[1] Individuals represented as the sample child in the household are identified in the person level file with a "flag" (variable: CSTATFLG). The sample child file includes topics such as child health conditions, limitations of activity, health status, child health care access and utilization, child mental health, and child influenza vaccination.

6.4 STRENGTHS AND LIMITATIONS

A strength of the NHIS is the ability to evaluate health characteristics by many different sociodemographic characteristics.[1] Limitations of the NHIS include an inability to calculate reliable statewide estimates. Statewide estimates can only be calculated using data from restricted data centers. All data collected are cross-sectional and self-reported. Although this is seen as a limitation, in the context of preventive health services research, previous studies have demonstrated that self-reported receipt of mammograms for early detection of breast cancer and other preventive health services are consistent with reports from medical providers and electronic medical records.

6.5 DESIGN FEATURES

Data analysts must use special procedures to account for the complex sample design used by the NHIS. Analytic procedures must include variables to adjust for the clustering, stratification, and weighting of each data file. Each time the NHIS sample design has been modified, the clustering and stratification variables were renamed. An overview of the NHIS clustering and stratification variables from 2000 through 2018 are provided in Table 6.1.

TABLE 6.1. OVERVIEW OF COMPLEX SAMPLE DESIGN VARIABLES ACROSS NHIS DATA COLLECTION YEARS

Survey Years	Clustering	Stratification
2000-2005	PSU	STRATUM
2006-2015	PSU_P	STRAT_P
2016-2018	PPSU	PSTRAT

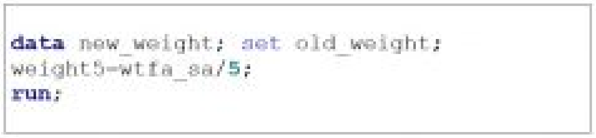
Interim and annual weights are provided for each file. The final annual weights are used to provide population estimates. An overview of final annual weight variables for each file type are provided in Table 6.2.

Table 6.2. Overview of NHIS final annual weight variables for each file type

Household	Family	Person	Sample Adult	Sample Child
WTFA_HH	WTFA_FAM	WTFA	WTFA_SA	WTFA_SC

When combining multiple years of NHIS data, you must divide the total annual weight by the total number of years in the merge prior to conducting their statistical analysis. For example, if combining data from sample adults from 2014-2018, WTFA_SA should be divided by 5. An example of how to do this in SAS 9.4 is provided in Box 6.1.

Box 6.1. SAS program to create new weight variable for five years of NHIS data



```
data new_weight; set old_weight;
weight5=wtfa_sa/5;
run;
```

6.6 NHIS CASE STUDY

Previous studies have demonstrated that Arab Americans are less likely than US-born Whites to receive preventive services, including influenza vaccines. Previous research using NHIS data has demonstrated that Arab American women were 66% less likely (OR=0.34; 95% CI=0.21-0.58) and Arab American men were 62% less likely (OR=0.38; 95% CI=0.21-0.67) to receive a flu vaccine when compared to their US-born non-Hispanic White counterparts.[6,7] These studies used restricted 2000-2011 NHIS data and created a variable for foreign-born Arab Americans. The Arab American ethnicity group was limited to foreign-born adults who were born in 15 countries that belong to the Arab League of Nations geographically located in the Middle East. In this case study, we will determine whether the results are similar using public-use 2018 NHIS data. Arab Americans will be defined as any adult who identified as self-reporting a White race, non-Hispanic or Latino/a ethnicity, and was born in a country in the Middle East region based on previous studies by the primary author and others.[8-13]

6.6.1 SPECIFIC AIMS

- Aim 6.1: Compare sociodemographic and health related characteristics of Arab Americans compared to US-born Whites and foreign-born non-Hispanic Whites from Europe (including Russia and Former USSR).
- Aim 6.2: Determine associations between region of birth and flu vaccine uptake among Arab Americans compared to US-born Whites and foreign-born non-Hispanic Whites from Europe.

6.6.2 METHODS

Complete the following steps to download, clean, recode, and analyze NHIS data to answer the specific aims.

Step 1: Download Person and Sample Adult datasets and SAS programming files

The association between region of birth and flu vaccine uptake can be examined using data from the NHIS person and sample adult level files.

- Go to the 2018 NHIS data release website
- Click (+) next to "Data Files"
- Under Person File, click on "ASCII data" and save to computer. A zip file will be downloaded which contains the person file. Open the zip file and save the data file to a permanent location on your computer. It is recommended that you create a folder on the 'C Drive' labeled NHIS and separated by each year (e.g. "18" for "2018') so that the location is consistent with the examples in this

textbook.

- Under Person File, click on “Sample SAS Statements” and save the SAS programming statements in the same folder as the data files on your computer.
- Under Sample Adult File, click on “ASCII data” and save to your computer. A zip file will be downloaded which contains the sample adult file. Open the zip file and save the data file to a permanent location on your computer. Similar to the Sample Adult file, I recommend creating a folder on the ‘C Drive’ labeled NHIS and separated by each year (e.g. “18” for “2018’) so that the location is consistent with the examples in this textbook.
- Under Sample Adult File, click on “Sample SAS Statements” and save the SAS programming statements in the same folder as the data file.

Step 2: Run SAS programming statements to create library and input person and sample adult files

Sample SAS programs to create the libraries and input the 2018 person and sample adult files are in provided in Box 6.2 and Box 6.3, respectively. To create these programming statements, complete the following steps:

Person Level SAS Programming File

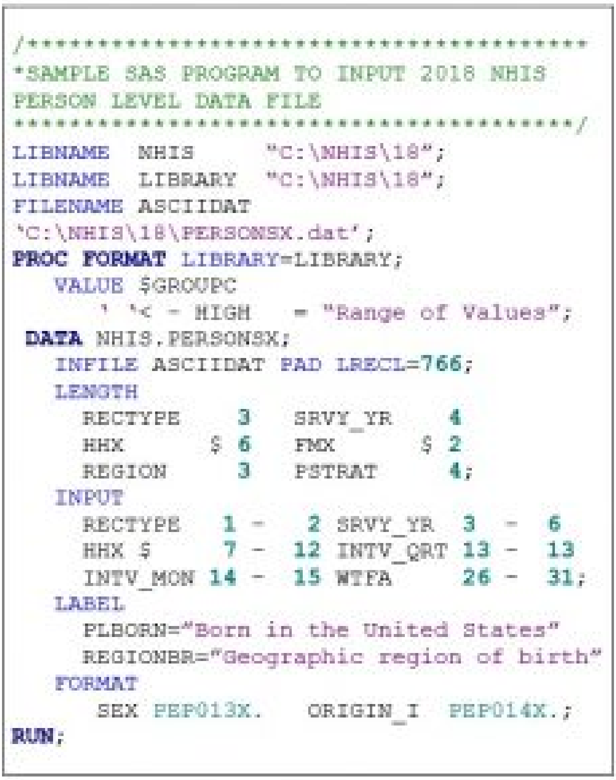
- Open the SAS Person Level Programming File
- Create a LIBNAME statement which houses the data and files associated with the analysis. It is recommended that you create the LIBNAME

statement as the survey name (e.g. “NHIS”) and use the same location that the data files are saved in on the C drive (e.g. “C:\NHIS\18”)

- Create a FILENAME statement which lets SAS know where the data file is stored (e.g. ‘C:\NHIS\18\PERSONSX.dat’)
- Modify or remove any instructions (/*green text*/) that you do not need in the programming file.
- Highlight all programming statements and click RUN.

The full SAS program for the person file is available for download in the Chapter 6 folder, in the Open ICPSR data repository.

Box 6.2. SAS Program to input 2018 NHIS person file



```
/*****************************************
*SAMPLE SAS PROGRAM TO INPUT 2018 NHIS
PERSON LEVEL DATA FILE
*****************************************/
LIBNAME  NHIS     "C:\NHIS\18";
LIBNAME  LIBRARY  "C:\NHIS\18";
FILENAME ASCIIDAT
'C:\NHIS\18\PERSONSX.dat';
PROC FORMAT LIBRARY=LIBRARY;
   VALUE $GROUPC
      ' '< - HIGH   = "Range of Values";
 DATA NHIS.PERSONSX;
   INFILE ASCIIDAT PAD LRECL=766;
   LENGTH
     RECTYPE     3   SRVY_YR    4
     HHX       $ 6   FMX      $ 2
     REGION      3   PSTRAT     4;
   INPUT
     RECTYPE   1 -   2 SRVY_YR  3  -  6
     HHX $     7 -  12 INTV_QRT 13 -  13
     INTV_MON 14 -  15 WTFA     26 -  31;
   LABEL
     PLBORN="Born in the United States"
     REGIONBR="Geographic region of birth"
   FORMAT
      SEX PEP013X.   ORIGIN_I  PEP014X.;
RUN;
```

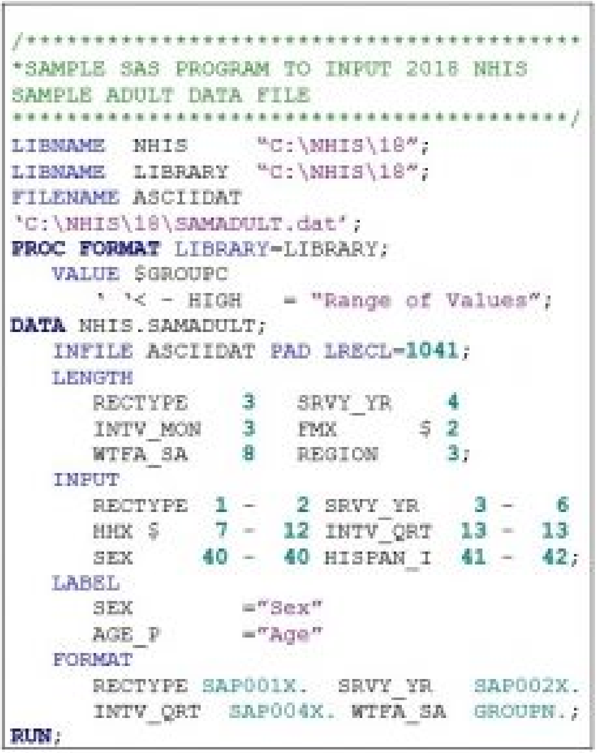
Sample Adult SAS Programming File

- Open the SAS Sample Adult Level Programming File.
- Create a LIBNAME statement which houses the data and files associated with the analysis. It is recommended that you create the LIBNAME statement as the survey name (e.g. "NHIS") and use the same location that the data files are saved in on the C drive (e.g. "C:\NHIS\18").

- Create a FILENAME statement which lets SAS know where the data file is stored (e.g. ‘C:\NHIS\18\SAMADULT.dat’).
- Modify or remove any instructions (/*green text*/) that you do not need in the programming file.
- Highlight all programming statements and click RUN.

The full SAS program for the sample adult file is available for download in the Chapter 6 folder, in the Open ICPSR data repository.

Box 6.3. SAS Program to input 2018 NHIS Sample Adult file



```
/**************************************
*SAMPLE SAS PROGRAM TO INPUT 2018 NHIS
SAMPLE ADULT DATA FILE
**************************************/
LIBNAME  NHIS     "C:\NHIS\18";
LIBNAME  LIBRARY  "C:\NHIS\18";
FILENAME ASCIIDAT
'C:\NHIS\18\SAMADULT.dat';
PROC FORMAT LIBRARY=LIBRARY;
   VALUE $GROUPC
      ' '< - HIGH   = "Range of Values";
DATA NHIS.SAMADULT;
   INFILE ASCIIDAT PAD LRECL=1041;
   LENGTH
      RECTYPE    3   SRVY_YR    4
      INTV_MON   3   FMX      $ 2
      WTFA_SA    8   REGION     3;
   INPUT
      RECTYPE  1 -   2 SRVY_YR    3 -   6
      HHX $    7 -  12 INTV_QRT  13 -  13
      SEX     40 -  40 HISPAN_I  41 -  42;
   LABEL
      SEX         ="Sex"
      AGE_P       ="Age"
   FORMAT
      RECTYPE SAP001X.  SRVY_YR   SAP002X.
      INTV_QRT  SAP004X. WTFA_SA  GROUPN.;
RUN;
```

Step 3: Combine Person Level and Sample Adult Data using MERGE statement

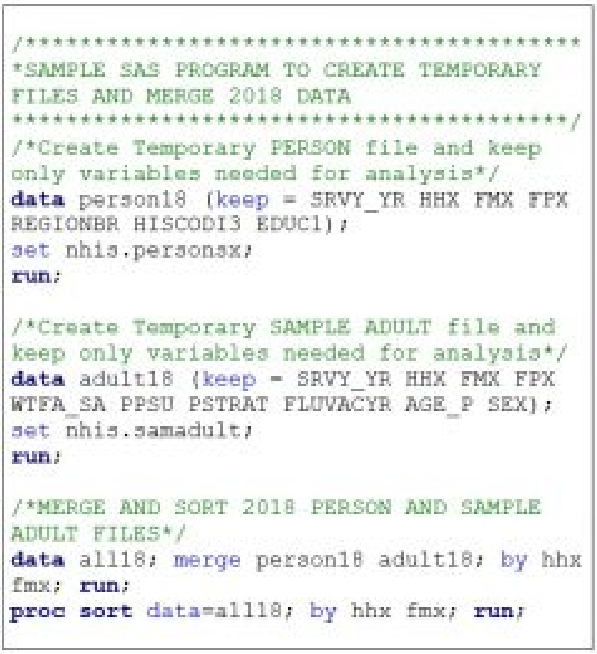
Once the data files have been input into SAS, the person and sample adult files must be combined. First, create a temporary file name for each file to indicate the type of data file and the year (e.g. person18 and samadult18). Second, sort each file by the household number (variable: HHX) and family number (variable: FMX) prior to merging. I also recommend using a KEEP statement to keep only the variables that you need for the analysis in your analytical dataset. Removing additional variables will allow the SAS program to run and present results faster. In this case study, I have kept the following variables (Table 6.3) to denote the survey design features and creation of the independent variable, dependent variable and selected covariates.

Table 6.3. Overview of variables used for NHIS case study

File	Variable Name	Variable Description
Design Variables		
Person and Sample Adult	SRVY_YR	Survey year
Person and Sample Adult	HHX	Household number
Person and Sample Adult	FMX	Family number
Person and Sample Adult	FPX	Person number
Sample Adult	WTFA_SA	Final annual weight
Sample Adult	PPSU	Primary sampling unit
Sample Adult	PSTRAT	Stratum
Independent Variable		
Person	REGIONBR	Region of birth
Person	HISCODI3	Race and ethnicity combined
Dependent Variable		
Sample Adult	FLUVACYR	Flu vaccine in past 12 months
Covariates		
Sample Adult	AGE_P	Age
Sample Adult	SEX	Sex
Person	EDUC1	Highest level of education

Third, merge and sort the combined dataset with only the variables from the person and sample adult files that are needed for analysis designed to meet your research aims. A sample SAS program for merging and sorting 2018 NHIS person and sample adult files is provided in Box 6.4.

Box 6.4. Sample SAS program to merge and sort NHIS person and sample adult data



```
/*****************************************
*SAMPLE SAS PROGRAM TO CREATE TEMPORARY
FILES AND MERGE 2018 DATA
*****************************************/
/*Create Temporary PERSON file and keep
only variables needed for analysis*/
data person18 (keep = SRVY_YR HHX FMX FPX
REGIONBR HISCODI3 EDUC1);
set nhis.personsx;
run;

/*Create Temporary SAMPLE ADULT file and
keep only variables needed for analysis*/
data adult18 (keep = SRVY_YR HHX FMX FPX
WTFA_SA PPSU PSTRAT FLUVACYR AGE_P SEX);
set nhis.samadult;
run;

/*MERGE AND SORT 2018 PERSON AND SAMPLE
ADULT FILES*/
data all18; merge person18 adult18; by hhx
fmx; run;
proc sort data=all18; by hhx fmx; run;
```

Step 4: Recode and rename variables

Questionnaire responses often need to be recoded or responses collapsed prior to conducting statistical analysis. For example, the NHIS has response options “7=Refused,” “8=Not ascertained,” and “9=Don’t know” for several questions. The responses are often removed and made “missing” prior to analysis. Furthermore, the numbers that represent certain values may need to be changed for easier interpretation of statistical analysis results. For example, NHIS has response options “1=Yes” and “2=No.” It is common practice to change “no” responses to 0, “0=No.” It is best practice to rename these recoded variables with a new variable name instead of replacing the original variable. Two or more variables may need to be combined in order to create the independent, dependent or other variables to answer study aims. In this case study, the variable created to make comparisons between US-born non-Hispanic Whites and foreign-born non-Hispanic Whites from Europe and the Middle East is created by combined two variables for 1) race and ethnicity and 2) region of birth. The variables recoded and renamed for analysis in this case study are provided in Table 6.4.

Table 6.4. Overview of NHIS variables recoded and renamed to meet research aims

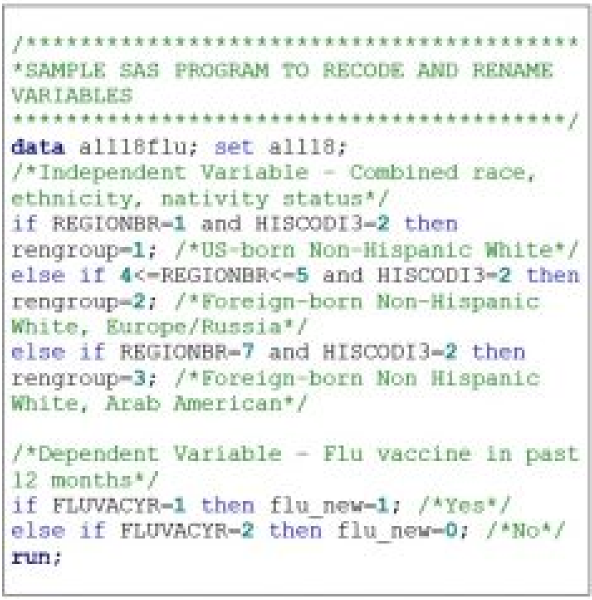
Question Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
Race and ethnicity	HISCODI3	1=Hispanic 2=Non-Hispanic White 3=Non-Hispanic Black 4=Non-Hispanic Asian 5-Non-Hispanic other races	RENGROUP (represents race, ethnicity and nativity group)	1=US-born non-Hispanic White 2=Foreign-born non-Hispanic White from Europe/Russia 3=Foreign-born non-Hispanic White from the Middle East (representing Arab Americans)
Region of birth	REGIONBR	01=United States 02=Mexico, Central America, Caribbean Islands 03=South America 04=Europe 05-Russia (and former USSR) 06=Africa 07=Middle East 08=Indian Subcontinent 09=Asia 10=SE Asia 11=Elsewhere 99=Unknown		
Flu vaccine, past 12 months	FLUVACYR	1=Yes 2=No 7=Refused 8=Not ascertained 9=Don’t know	FLU_NEW	0=No 1=Yes

Table 6.4 (continued) Overview of NHIS variables recoded and renamed to meet research aims

Question Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
Age	AGE_P	18-84=18-84 years 85=85+ years	AGE_NEW	1=18-34 years 2=35-54 years 3=55-64 years 4=65+ years
Highest level of education	EDUC1	00=Never attended/ kindergarten only 01-12=1st through 12th grade, no diploma 13=GED or equivalent 14=High school (HS) graduate 15=Some college, no degree 16-17=Associate's degree programs 18=Bachelor's degree 19=Master's degree 20-21=Doctorate 96=Child <5 97=Refused 98=Not ascertained 99=Don't know	EDUC_NEW	1=Less than HS graduate 2=HS graduate/ GED 3=Some college/ Associate's degree 4=Bachelor's degree or higher

A sample SAS program for recoding and renaming 2018 NHIS data for this case study is provided in Box 6.5.

Box 6.5. Sample SAS program to recode and rename NHIS variables

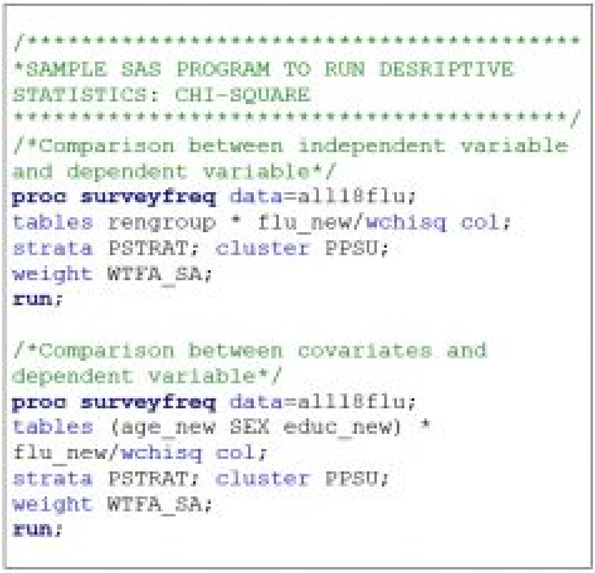


```
/*****************************************
*SAMPLE SAS PROGRAM TO RECODE AND RENAME
VARIABLES
*****************************************/
data all18flu; set all18;
/*Independent Variable - Combined race,
ethnicity, nativity status*/
if REGIONBR=1 and HISCODI3=2 then
rengroup=1; /*US-born Non-Hispanic White*/
else if 4<=REGIONBR<=5 and HISCODI3=2 then
rengroup=2; /*Foreign-born Non-Hispanic
White, Europe/Russia*/
else if REGIONBR=7 and HISCODI3=2 then
rengroup=3; /*Foreign-born Non Hispanic
White, Arab American*/

/*Dependent Variable - Flu vaccine in past
12 months*/
if FLUVACYR=1 then flu_new=1; /*Yes*/
else if FLUVACYR=2 then flu_new=0; /*No*/
run;
```

Step 5: Conduct Descriptive Statistical Analysis

Once all variables are recoded, collapsed, and renamed they can be used for statistical analysis. Statistical analysis should always start with descriptive analysis to describe the data source. Chi square analyses should be conducted to make comparisons between the independent variable, covariates, and dependent variables. It is important to remember that all analysis of NHIS data needs to be conducted with SAS survey procedures due to the complex sample design. Weighting (variable: WTFA_SA), clustering (variable: PPSU) and stratification (variable: PSTRAT) variables must be included in the programming statements. A sample SAS program for conducting chi-square tests using 2018 NHIS data for this case study is provided in Box 6.6.

Box 6.6. Sample SAS program for running descriptive statistics (chi-square)

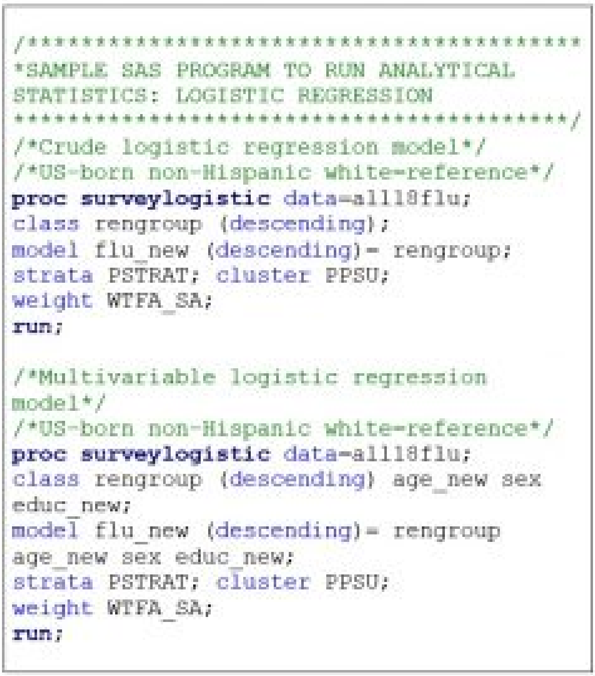


```
/*******************************************
*SAMPLE SAS PROGRAM TO RUN DESRIPTIVE
STATISTICS: CHI-SQUARE
*******************************************/
/*Comparison between independent variable
and dependent variable*/
proc surveyfreq data=all18flu;
tables rengroup * flu_new/wchisq col;
strata PSTRAT; cluster PPSU;
weight WTFA_SA;
run;

/*Comparison between covariates and
dependent variable*/
proc surveyfreq data=all18flu;
tables (age_new SEX educ_new) *
flu_new/wchisq col;
strata PSTRAT; cluster PPSU;
weight WTFA_SA;
run;
```

Step 6: Conduct Inferential Statistical Analysis

After calculating descriptive statistics, inferential statistical analysis can be conducted. Crude and multivariable logistic regression models can be calculated to determine associations between race, ethnicity, and nativity status and flu vaccine uptake among US- and foreign-born non-Hispanic Whites from Europe and the Middle East. Crude logistic regression models are used to determine the association between the independent and dependent variables without adjusting for other factors. Multivariable logistic regression models are used to determine associations between the independent and dependent variables after adjusting for potential covariates (e.g. age, sex, highest level of education). A reference category for the independent variable is needed. For this analysis, the reference group is US-born non-Hispanic Whites. Results from foreign-born non-Hispanic Whites from Europe and the Middle East are presented in comparison to US-born non-Hispanic Whites. A sample SAS program for conducting logistic regression analysis using 2018 NHIS data for this case study is provided in Box 6.7.

Box 6.7. Sample SAS program for running inferential statistics (logistic regression)



```
/*****************************************
*SAMPLE SAS PROGRAM TO RUN ANALYTICAL
STATISTICS: LOGISTIC REGRESSION
*****************************************/
/*Crude logistic regression model*/
/*US-born non-Hispanic white=reference*/
proc surveylogistic data=all18flu;
class rengroup (descending);
model flu_new (descending)= rengroup;
strata PSTRAT; cluster PPSU;
weight WTFA_SA;
run;

/*Multivariable logistic regression
model*/
/*US-born non-Hispanic white=reference*/
proc surveylogistic data=all18flu;
class rengroup (descending) age_new sex
educ_new;
model flu_new (descending)= rengroup
age_new sex educ_new;
strata PSTRAT; cluster PPSU;
weight WTFA_SA;
run;
```

6.7 SUMMARY

This chapter provided an overview of the NHIS and ways to conduct basic statistical analysis using 2018 public-use data files. The NHIS case study explored whether foreign-born Arab American adults were more or less likely to receive an annual flu vaccine in comparison to US-born non-Hispanic Whites. Sample SAS programming statements were provided for downloading and importing data files, merging data files, recoding and renaming variables and conduction categorical descriptive and inferential statistical analysis. The dataset and full SAS programming statements for the NHIS case study are available in the Chapter 6 folder, in the Open ICPSR data repository.

6.8 REFERENCES

1. National Center for Health Statistics. *Survey Description, National Health Interview Survey, 2018.*; 2019.
2. Blewett LA, Dahlen HM, Spencer D, Rivera Drew JA, Lukanen E. Changes to the Design of the National Health Interview Survey to Support Enhanced Monitoring of Health Reform Impacts at the State Level. *Am J Public Health.* 2016;106(11):1961-1966. doi:10.2105/AJPH.2016.303430
3. National Center for Health Statistics. NHIS – 2019 Questionnaire Redesign. Published November 27, 2019. Accessed September 2, 2020. https://www.cdc.gov/nchs/nhis/2019_quest_redesign.htm

4. National Center for Health Statistics. National Health Interview Survey, 2018. Public-use data file and documentation. Published 2019. Accessed August 10, 2020. https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm

5. Office of Management and Budget. Revisions to the Standards for the Classification of Federal Data on Race and Ethnicity. The White House. Published 1997. Accessed July 6, 2021. https://obamawhitehouse.archives.gov/omb/fedreg_1997standards

6. Dallo FJ, Kindratt TB. Disparities in vaccinations and cancer screening among U.S.- and foreign-born Arab and European American non-Hispanic White women. *Womens Health Issues*. 2015;25(1):56-62. doi:10.1016/j.whi.2014.10.002

7. Dallo FJ, Kindratt TB. Disparities in preventive health behaviors among non-Hispanic White men: heterogeneity among foreign-born Arab and European Americans. *Am J Mens Health*. 2015;9(2):124-131. doi:10.1177/1557988314532285

8. Dallo FJ, Kindratt TB, Snell T. Serious psychological distress among non-Hispanic whites in the United States: the importance of nativity status and region of birth. *Soc Psychiatry Psychiatr Epidemiol*. 2013;48(12):1923-1930. doi:10.1007/s00127-013-0703-1

9. Dallo FJ, Kindratt TB, Zahodne L. Prevalence of Self-Reported Cognitive Impairment among Arab American Immigrants in the United States.

Innovation in Aging. 2020;(igaa058). doi:10.1093/geroni/igaa058

10. Read JG, Amick B, Donato KM. Arab immigrants: a new case for ethnicity and health? *Soc Sci Med*. 2005;61(1):77-82. doi:10.1016/j.socscimed.2004.11.054
11. Read JG, Reynolds MM. Gender differences in immigrant health: the case of Mexican and Middle Eastern immigrants. *J Health Soc Behav*. 2012;53(1):99-123. doi:10.1177/0022146511431267
12. Reynolds MM, Chernenko A, Read JG. Region of origin diversity in immigrant health: Moving beyond the Mexican case. *Soc Sci Med*. 2016;166:102-109. doi:10.1016/j.socscimed.2016.07.018
13. Blumberg SJ, Parker JD, Moyer BC. National Health Interview Survey, COVID-19, and Online Data Collection Platforms: Adaptations, Tradeoffs, and New Directions. *Am J Public Health*. 2021;111(12):2167-2175. doi:10.2105/AJPH.2021.306516
14. Dallo FJ, Borrell LN. Self-reported diabetes and hypertension among Arab Americans in the United States. *Ethn Dis*. 2006;16(3):699-705.
15. National Center for Health Statistics. National Health Interview Survey, 2020. Public-use data file and documentation. Published 2021. Accessed August 10, 2020. https://www.cdc.gov/nchs/nhis/data-questionnaires-documentation.htm

16. NHIS – 2021 NHIS. Published February 8, 2022. Accessed June 28, 2022. https://www.cdc.gov/nchs/nhis/2021nhis.htm
17. National Center for Health Statistics. Overview of Weighting and Analytic Options for the 2020 National Health Interview Survey. Published 2021. https://ftp.cdc.gov/pub/Health_Statistics/NCHS/Dataset_Documentation/NHIS/2020/weighting-analysis-reference-508.pdf

6.9 COVID-19 PANDEMIC CHANGES

Sections 6.1 to 6.8 were written during the initial waves of the COVID-19 pandemic. Since the NHIS is an in-person survey conducted at participants' households, there were significant disruptions to the regular methodology due to stay-at-home orders and safety concerns for both participants and the NHIS field workers. With the declaration of a national emergency on March 13, 2020, changes were needed to adhere to essential work-based restrictions, collect data more effectively, and create survey content that could be used for evidence-based public health decision-making. This section will provide an overview of the data collection, survey content, and complex sample analytic changes due to disruptions by the COVID-19 pandemic.

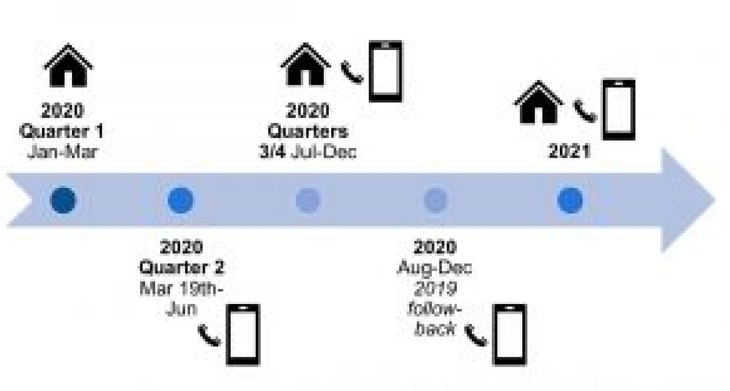
6.9.1 DATA COLLECTION CHANGES

On March 19, 2020, data collection methods temporality switched to telephone only.[14,15] Commercial address lists were used to locate telephone numbers for the addresses identified in the sampling frame. Telephone number from ~60% of households were obtained. Online searches were conducted to try to locate other numbers. With this change in methodology, the response rate lowered to 42% from April to June 2020 in comparison to 59% obtained from January to March 2020.[14] NHIS data users may be surprised that online methodologies were not adopted. Instead of adjusting the household-based sampling frame to an internet modality, the NCHS utilized its Research and Development Survey (RANDS) and the US Census Bureau modified its Household Pulse Survey (HPS) to provide real-time data on COVID-19 for public health and policy decision-making.[14]

In 2020, there were four data collection methods used by the NHIS. During quarter 1 (January through March 15th), data were collected using the CAPI system at participants' households. During quarter 2 (March 19th through June), surveys were collected by telephone. During quarters 3 and 4 (July through December), surveys were collected by telephone and in-person. To protect the safety of NHIS field workers and participants, an attempt was made to collect data from each household by telephone prior to conducting in-person interviews. If participants were not able to be reached by telephone, or if any recruitment or survey materials needed to be distributed, data were collected using in-person interviews. Safety precautions such as social distancing, masking, and collecting data outside were used to protect participants and NHIS field workers.[14] From August through December 2020, efforts were made to improve response rates. A "follow-back" survey was conducted by telephone with the previous years' adult participants. This method allowed for increasing the sample size but also allowed for comparisons to be made pre and post the initial waves of the pandemic. From January through

April 2021, participants were surveyed first by telephone and only in-person as needed only if they were not available by phone. Beginning May 2021, field workers were allowed to collect data in-person after assessing their own personal health risks and community transmission levels. An overview of the 2020-2021 data collection timeline is provided in Figure 6.1.

Figure 6.1. NHIS data collection changes during COVID-19 pandemic



6.9.2 SURVEY CONTENT CHANGES

The NHIS also made changes to the survey content to provide morbidity and health care services use data during the pandemic. In 2020, adults and children answered questions about COVID-19 diagnosis and testing. Participants were asked whether they delayed or skipped medical care or used telemedicine instead of in-person visits. Among adults only, questions were also added to determine changes and delays in cancer care and caregiving.[14] Questions also assessed social distancing measures in the workplace and changes in social and emotional support needs in the last 12 months.[14] In 2021, addition questions were added on COVID-19 vaccinations.[16]

6.9.3 COMPLEX SAMPLE ANALYTIC CHANGES

The changes in the data collection procedures led to new potential non-response biases. Once the data collection shifted to telephone, participants who were ages 65 and older, had a bachelor's degree level of education or higher, a family income greater than $75,000, owned a home, and resided in that home for 11 years or longer had higher response rates than other groups.[15] Changes were made to the NHIS weighting process from previous years. Weights were already calibrated to the US Census population estimates for age, sex, race, ethnicity, education, and geographic areas. From April 2020 onward, weights were adjusted to account for housing tenure.[14,16]

The 2020 NHIS public-use data files are separated to account for the changes in the complex sample. The four data files include the: 1) sample adult interview; 2) sample adult longitudinal; 3) sample adult partial; and 4) sample child interview. The sample adult interview file includes all adults who provided data during 2020. The sample adult longitudinal file online includes the 2019 follow-back sample. The sample adult partial file does not include the 2019 follow-back participants. The sample child interview file is the only file for children. An overview of the weighting variables used for each of these files is provided in Table 6.5.

Table 6.5. Overview of 2020 NHIS final weight variables for each file type

Sample Adult Interview	Sample Adult Longitudinal	Sample Adult Partial	Sample Child Interview
WTFA_A	WTFA_L	WTFA_P	WTFA_S

Details of the 2021 NHIS weighting procedures are not available as of this writing. More information on changes to the NHIS data collection procedures due to the COVID-19 pandemic are available on the website.[14,17]

CHAPTER 7.

MEDICAL EXPENDITURE PANEL SURVEY

7.1 INTRODUCTION

Chapter 7 covers the Medical Expenditure Panel Survey (MEPS). The MEPS comprises a series of surveys from families, individuals, employers and medical providers that have been collected by the Agency for Healthcare Research and Quality (AHRQ) since 1996. The purpose of the MEPS is to estimate and monitor national trends in health care use, medical costs and health insurance coverage.[1] This chapter includes details on: how data are collected; how data are made publicly available as machine-actionable data files; what variables must be included to address design features of the complex sample; the strengths and limitations of the survey; and practical tips for conducting statistical analysis; and how to answer research questions using a case study. The practical tips provided for analysis of MEPS data are based on the author's previous experiences analyzing MEPS data from 2001-2017 to answer questions related to associations between predisposing and enabling factors that contribute to morbidity, mortality and health services use. The MEPS case study will explore whether adults who perceive their physician provided quality patient-provider communication (PPC) are more or less likely to receive an annual flu vaccine in comparison those who did not receive quality PPC. The bulk of the chapter will comprise of section *7.6: MEPS Case Study* in order to give investigators hands-on practice downloading and cleaning large databases and conducting basic categorical data analysis using PROC SURVEYFREQ and PROC SURVEYLOGISTIC. The syntax provided was created for use with SAS 9.4.

7.2 DATA COLLECTION

The MEPS uses a longitudinal panel design collecting data from individuals and families from five rounds of interviews over a two-year period.[2] Data can be compiled for cross-sectional analyses or longitudinal, retrospective cohort designs. A subsample of household participants who filled out the previous year's National Health Interview Survey (NHIS) are selected for each panel. Oversampling of subgroups aligns with the NHIS nationally representative sample design.[2] For example, non-Hispanic Blacks and Hispanics were oversampled until 2016. Non-Hispanic Asians were oversampled from 2006-2015. In 2016, the new NHIS design was modified to increase precision of statewide estimates and these changes were reflected in the MEPS beginning in 2017.[3] Data for 2018 onward were not available as of this writing. Further details of the MEPS sampling design and data collection methods are reported elsewhere.[4]

7.3 DATA FILES

The MEPS is comprised of two main components: 1) household component and 2) medical provider component. Each component includes multiple subsections. The MEPS insurance component also collects health insurance plan offerings from employers in the US on an annual basis.[5] Each annual household consolidated and medical condition file represents data from two panels. Each longitudinal file represents data from one panel over a two-year period. An overview of file names for household consolidated, medical condition, and longitudinal files from 2001-2018 by panel and year are provided in Table 7.1.

TABLE 7.1. OVERVIEW OF 2001-2018 MEPS FILE NAMES BY YEAR AND PANEL

Year	Panels	Household	Medical Condition	Longitudinal	Years	Panel
2001	5/6	H60	H61	H65	2000-2001	5
2002	6/7	H70	H69	H71	2001-2002	6
2003	7/8	H79	H78	H80	2002-2003	7
2004	8/9	H89	H87	H86	2003-2004	8
2005	9/10	H97	H96	H98	2004-2005	9
2006	10/11	H105	H104	H106	2005-2006	10
2007	11/12	H113	H112	H114	2006-2007	11
2008	12/13	H121	H120	H122	2007-2008	12
2009	13/14	H129	H128	H130	2008-2009	13
2010	14/15	H138	H137	H139	2009-2010	14
2011	15/16	H147	H146	H148	2010-2011	15
2012	16/17	H155	H154	H156	2011-2012	16
2013	17/18	H163	H162	H164	2012-2013	17
2014	18/19	H171	H170	H172	2013-2014	18
2015	19/20	H181	H180	H183	2014-2015	19
2016	20/21	H192	H190	H193	2015-2016	20
2017	21/22	H201	H199	H202	2016-2017	21
2018	22/23	H209	H207	H210	2017-2018	22

7.3.1 HOUSEHOLD COMPONENT

An adult reference person from each household is identified to answer questions about adults and child members of the family.[2] The household component includes questions describing demographics (e.g. sex, age, race/ethnicity, education, marital status, nativity status, English language proficiency), income level, health insurance, access to health care, person-level health conditions (e.g. perceived health status and priority health conditions – such as diabetes, hypertension, cancer) preventive health services, limitations, and disabilities. MEPS also assesses personal characteristics of health care providers, including specialty type, race/ ethnicity, and sex, which can be used for studies evaluating patient-provider concordance. Child special health care needs, behavioral problems, and preventive care received at doctors' visits are also measured. Parents are also asked to answer questions on health care quality from the Consumer Assessment of Healthcare Providers and Systems (CAHPS©). Self-administered questionnaires (SAQ) are also collected on a periodic (2014 Preventive Care SAQ; 2016-2017 Cancer SAQ) and consistent annual (SAQ, Diabetes Care Survey) basis.[2]

7.3.1.a. Self-Administered Questionnaire (SAQ)

During rounds 2 and 4 of each panel, adult participants are asked to complete and mail back a paper-and-pencil SAQ which includes CAHPS© questions assessing health care quality received in the last 12 months. Sample questions include whether or not participants' received care right away, were offered help filling out forms, and how they rate their health care (scale 0-10) with 0 representing the worst health care possible and 10 representing the best health care received. Additional questions assess general health (e.g. smoking status), health status, non-specific psychological distress, and depression.[2]

7.3.1.b. Diabetes Care Survey (DCS)

During rounds 3 and 5, adults who reported that they were ever told by a doctor or other health professional that they ever had diabetes receive an additional paper-and-pencil questionnaire to complete and mail back to AHRQ.[2] The DCS asks questions on diabetes care treatment, monitoring, self-efficacy, and ways that individuals learn how to take care of their diabetes. For example, participants are asked whether their diabetes is being treated by diet modification, oral medication, and insulin injections. To evaluate diabetes care monitoring, participants are asked whether they have had their feet checked, eye exam, blood cholesterol checked, received a flu vaccine, and the number of times tested for hemoglobin A1c in the past year. To determine diabetes care self-efficacy, participants are asked to rate their confidence level in taking care of their diabetes from 1=not confident at all to 4=very confident.[2]

7.3.2 MEDICAL PROVIDER COMPONENT

After completion of household interviews, participants are asked to provide permission for their medical providers to be contacted to verify health information. Medical providers are contacted by telephone to confirm diagnostic and procedure codes, billing charges and payments, utilization (number of medical events) and dates of visits. From 2000-2016, International Classification of Diseases, Ninth Revision, Clinical Modification (ICD-9-CM) codes were assigned by professional coders based on household interview responses. From 2017 onward, coding transitioned to include International Classification of Diseases, Tenth

Revision, Clinical Modification (ICD-10-CM) codes.[6] Information on utilization (number of medical events), charges, and sources of payment are collected.[2] Additional details on prescriptions medicines are collected as part of the pharmacy subcomponent, including medicine names, drug details and national drug codes.[6]

7.3.3 LONGITUDINAL FILES

The MEPS is unique to other national health surveys because it collects data from its participants five times over a two-year period, instead of one point in time. This allow users to determine changes over the two-year time period and offers fewer temporal biases that may occur when comparing exposures and outcomes during the same year.[7]

7.4. STRENGTHS AND LIMITATIONS

A strength of the MEPS is that the survey collects data from multiple sources (individuals, medical providers, employers) on a large number of health services topics not covered by other surveys, including use of medical services, charges/source of payments, provider satisfaction, patient-provider communication, provider characteristics, and care coordination.[2] Other advantages include the ability to conduct longitudinal analysis with each panel and the use of cross-walk files to link to the previous year's NHIS data. A limitation of the MEPS is that the samples sizes overlap for two years due to the survey panel design. This means that there is a much smaller sample size in comparison to other surveys. For example, the NHIS had 78,132 total respondents in 2017 whereas the MEPS had only 30,716 respondents.[2,8] Although the MEPS collects information on a wide variety of physician characteristics, another limitation is the underrepresentation of physician assistants and nurse practitioners as advanced practice providers for workforce research. Limited information is collected on team-based practice and non-physician provider practices.[9]

7.5. DESIGN FEATURES

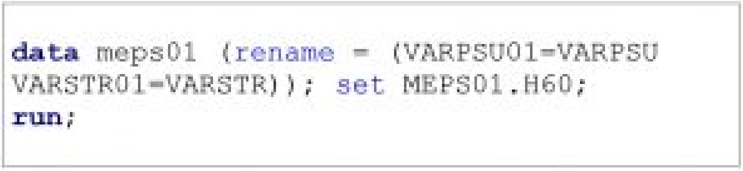
Data analysts must use special procedures to account for the complex sample design used by the MEPS. Researchers must include variables to adjust for the primary sampling units, clustering and weighting of each annual data files. MEPS provides a primary sampling unit and cluster variable for each year. Prior to 2002, MEPS included the data collection year in the variable name (VARPSU01, VARSTR01). Beginning in 2002, MEPS removed the data collection year from the variable name, which allows for a more streamlined approach when combining data across years. An overview of the primary sampling unit and cluster variables are provided in Table 7.2.

TABLE 7.2. OVERVIEW OF COMPLEX SAMPLE DESIGN VARIABLES ACROSS MEPS DATA COLLECTION YEARS

MEPS Survey Years	MEPS Primary Sampling Unit	MEPS Cluster
2001	VARPSU01	VARSTR01
2002-2018	VARPSU	VARSTR

For studies that combine data from 2001 with 2002 or later, researchers must change the variable name prior to combining the datasets. This can be done by using the RENAME statement. A sample SAS program for renaming variables is provided in Box 7.1.

Box 7.1. Rename variable prior to combining with additional years of MEPS data



```
data meps01 (rename = (VARPSU01=VARPSU
VARSTR01=VARSTR)); set MEPS01.H60;
run;
```

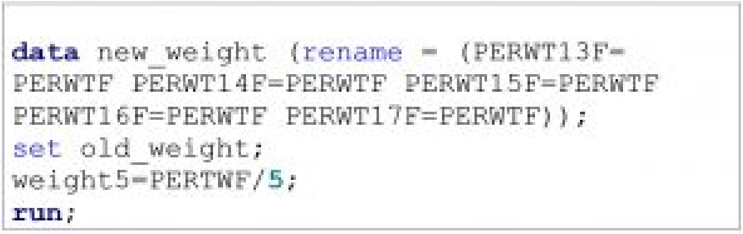
Several annual weights are provided for each file. The annual weights that include the year can be renamed prior to combining with additional years of data. The final person weight (e.g. PERWT17F) is used when analyzing general responses from the household consolidated and medical condition files. For research projects focused on questions gathering data directly from the SAQ (e.g. patient-provider communication quality), the self-administered questionnaire weight (e.g. SAQWT17F) should be used. For research projects focused on questions gathering data directly from individuals with diabetes (e.g. diabetes monitoring, self-efficacy), the final diabetes care supplement weight (e.g. DIABW17F) should be used. An overview of final annual weight variables for each file type for 2017 data are provided in Table 7.3.

Table 7.3. Overview of 2017 MEPS final annual weight variables for each file type

Weight Type	Person	Medical Condition	Longitudinal
Final Person Weight	PERWT17F	PERWT17F	LONGWT
Final Self-Administered Questionnaire Weight	SAQWT17F	–	LSAQWT
Final Diabetes Care Supplement Weight	DIABW17F	–	–

When combining multiple years of MEPS data, investigators must divide the total annual weight by the total number of years in the merge prior to conducting their statistical analysis. For example, if combining data from individuals from 2013-2017, all final person weight variables should be renamed and then divided by 5 in SAS. An example of this coding is provided in Box 7.2.

Box 7.2. SAS program to rename and create new weight variable for five years of MEPS data



```
data new_weight (rename = (PERWT13F=
PERWTF PERWT14F=PERWTF PERWT15F=PERWTF
PERWT16F=PERWTF PERWT17F=PERWTF));
set old_weight;
weight5=PERTWF/5;
run;
```

7.6. MEPS CASE STUDY

Previous studies have produced mixed results on associations between adults' perceptions of the qualities of patient-provider communication (PPC) received and their likelihood of receiving preventive services. MEPS measures several PPC qualities, including adults' perceptions of how often their health care providers "listened carefully to you," "spent enough time with you," and "explained things in a way that was easy to understand" in the last 12 months.[1] Starting in 2011, MEPS included additional PPC qualities which focus on measuring health literate practices of health care providers. These questions include how often in the last 12 months health care providers "give instructions on what to do about as specific illness or health condition," "how often these instructions were easy to understand" and whether the health care provider asked you "to describe how you were going to follow" their instructions.[2]

Kindratt and colleagues previous research has demonstrated that several qualities of PPC increased adults' likelihood of receiving breast, cervical and colorectal cancer screenings.[10–12] For example, using MEPS data, Kindratt and colleagues found that adults who reported their provider always gave specific instructions about what to do for their specific health conditions had higher odds of receiving breast (OR=1.19; 95% CI=1.06-1.33) and colorectal (OR=1.23; 95% CI=1.10-1.37) cancer screenings compared to those whose providers did not give specific instructions.[10] When disaggregated by race and ethnicity, non-Hispanic Black and Hispanic adults who reported their providers' exhibited all PPC qualities were more likely to receive colorectal cancer screenings compared to those who did not exhibit all qualities.[12] A study by Villani & Mortensen evaluated whether adults who reported their health care provider always showed respect, involved them in decision-making, and explained all options were more or less likely to receive an influenza vaccine compared to adults whose providers' did not exhibit these PPC qualities.[13] They found no association between each PPC quality and adults' likelihood of receiving a flu vaccine. More research evaluating other PPC qualities, such as health literate practices by providers (providing clear communication, practicing "teach-back" method), are needed to further explore this association. In this case study, we will determine the association between two PPC qualities and adults' likelihood of receiving an influenza vaccine using 2015 and 2016 MEPS household consolidated data.

7.6.1 SPECIFIC AIMS

- Aim 7.1: Compare sociodemographic and health-related characteristics of adults by influenza vaccine uptake
- Aim 7.2: Determine association between adults' perceptions of PPC qualities and their likelihood of receiving an influenza vaccine before and after controlling for covariates

7.6.2 METHODS

Complete the following steps to download, clean, recode and analyze MEPS data to answer the specific aims.

Step 1: Download Household Consolidated datasets and SAS programming files

The associations between PPC qualities and influenza vaccine uptake can be examined using data from the MEPS household consolidated data files.

- Go to the MEPS website for household full year consolidated data files.
- Click on "HC-181" for 2015 Full Year Consolidated Data File. Under "Data" and "File type***," click on "ZIP" next to "Data File, ASCII format" and save the file to your computer. A zip file will be downloaded which contains the household consolidated file. Open the zip file and save the data file to a permanent location on your computer. I recommend creating a folder on the 'C Drive' labeled MEPS and separated by each year (e.g. "15" for "2015') so that the location is consistent with the examples in this textbook.
- Under "Documentation" and "File Type," click on "TXT" next to the label "SAS Programming Statements." Open your SAS program then copy and paste the text into the SAS editor. Save the SAS programming statements in the same folder as the data file.
- Repeat the previous steps for obtaining, downloading and saving 2016 household consolidated data.

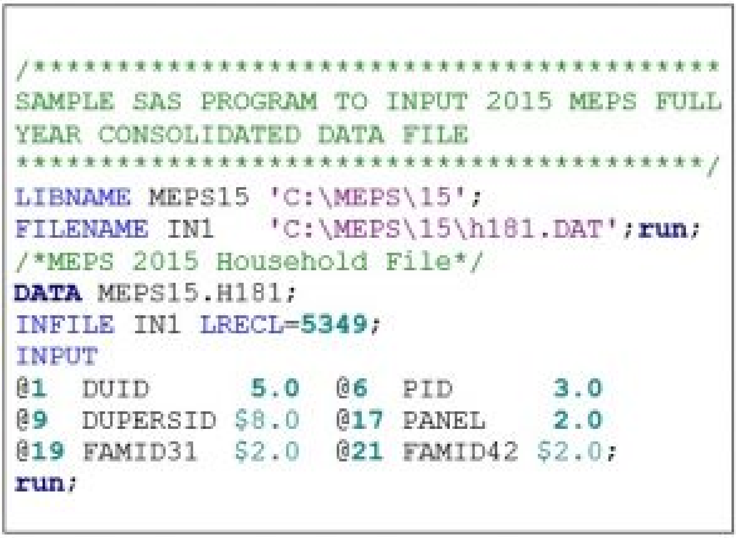
Step 2: Run SAS programming statements to create library and input 2015 and 2016 MEPS household consolidated data files

Sample SAS programs to create the libraries and input the 2015 and 2016 MEPS household consolidated data files are provided in Box 7.3 and Box 7.4, respectively. The full SAS programs are available for download in the chapter 7 folder of the Open ICPSR data repository.

2015 MEPS Full Year Consolidated SAS Programming File

- Open the SAS Full Year Consolidated Programming File
- Remove any unnecessary instructions within programming file. Labels and formats can be removed since all variables will be recoded when you complete your analysis.
- Create a LIBNAME statement which houses the data and files associated with the analysis. I recommend creating the LIBNAME statement as the survey name (e.g. "MEPS") and using the same location that the data files for saved in on the C drive (e.g. "C:\MEPS\15")
- Create a FILENAME statement which lets SAS know where the data file is stored (e.g. 'C:\MEPS\15\H181.dat')
- Highlight all programming statements and click RUN.

Box 7.3. SAS Program to input MEPS HC-181: 2015 Full Year Consolidated Data File



```
/*****************************************
SAMPLE SAS PROGRAM TO INPUT 2015 MEPS FULL
YEAR CONSOLIDATED DATA FILE
*****************************************/
LIBNAME MEPS15 'C:\MEPS\15';
FILENAME IN1   'C:\MEPS\15\h181.DAT';run;
/*MEPS 2015 Household File*/
DATA MEPS15.H181;
INFILE IN1 LRECL=5349;
INPUT
@1  DUID       5.0  @6  PID      3.0
@9  DUPERSID $8.0  @17 PANEL    2.0
@19 FAMID31  $2.0  @21 FAMID42 $2.0;
run;
```

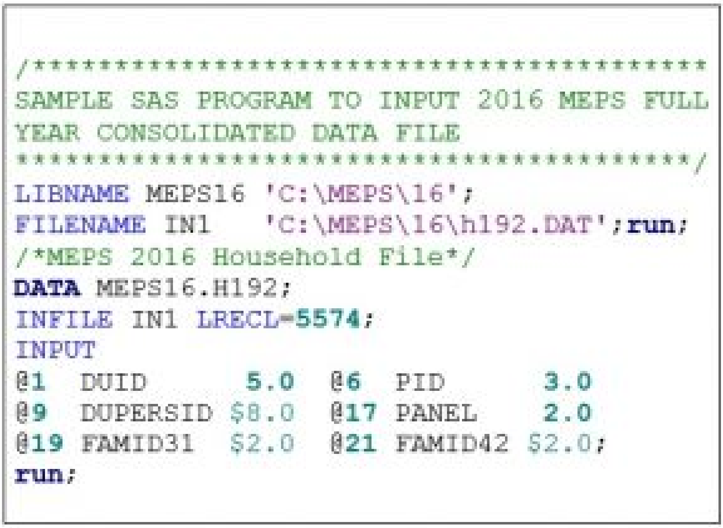
2016 MEPS Full Year Consolidated SAS Programming File

- Open the SAS Full Year Consolidated Programming File
- Remove any unnecessary instructions within programming file. Labels and formats can be removed since all variables will be recoded when you complete your analysis.
- Create a LIBNAME statement which houses the data and files associated with the analysis. I recommend creating the LIBNAME statement as the survey name (e.g. "MEPS") and using the same location that the data files are saved in on the C drive (e.g. "C:\MEPS\16")
- Create a FILENAME statement which lets SAS

know where the data file is stored (e.g. 'C:\MEPS\16\H192.dat')

- Highlight all programming statements and click RUN.

Box 7.4. SAS Program to input MEPS HC-192: 2016 Full Year Consolidated Data File



```
/*****************************************
SAMPLE SAS PROGRAM TO INPUT 2016 MEPS FULL
YEAR CONSOLIDATED DATA FILE
*****************************************/
LIBNAME MEPS16 'C:\MEPS\16';
FILENAME IN1   'C:\MEPS\16\h192.DAT';run;
/*MEPS 2016 Household File*/
DATA MEPS16.H192;
INFILE IN1 LRECL=5574;
INPUT
@1  DUID      5.0  @6  PID      3.0
@9  DUPERSID $8.0  @17 PANEL    2.0
@19 FAMID31  $2.0  @21 FAMID42 $2.0;
run;
```

Step 3: Combine 2015 and 2016 Full Year Consolidated Files using MERGE statement

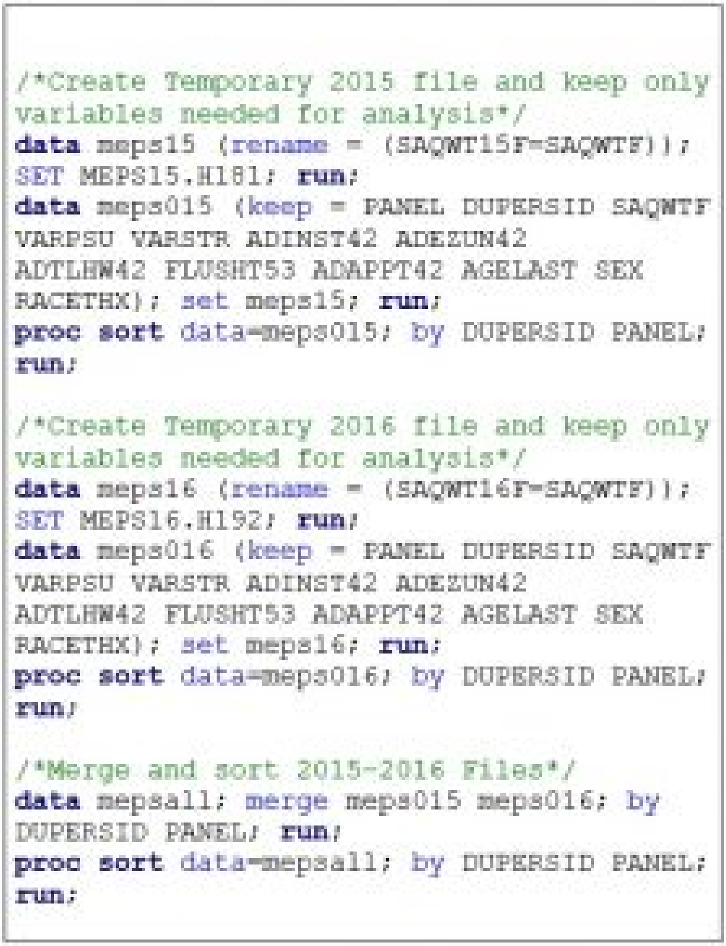
Once the data files have been input into SAS, the annual files must be combined. First, create a temporary file name for each file to indicate the type of data file and the year (e.g. meps15 and meps16). Second, sort each file by the participant id number (variable: DUPERSID) and panel (variable: PANEL) prior to merging. I also recommend using a KEEP statement to keep only the variables that you need for the analysis in your analytical dataset. Removing additional variables will allow the SAS program to run and present results faster. For MEPS, rename any variables that include the year in the name (e.g. PERWT15F= PERWTF, PERWT16F= PERWTF) prior to merging data across years. In this case study, I have kept the following variables (Table 7.4) to denote the survey design features and creation of the independent variable, dependent variable and selected covariates.

Table 7.4. Overview of variables used for MEPS case study

Variable Name	Variable Description
Design Variables	
PANEL	Panel number
DUPERSID	Person ID (Dwelling unit + person number)
SAQWT15F	Final SAQ person weight, 2015
SAQWT16F	Final SAQ person weight, 2016
VARPSU	Primary sampling unit
VARSTR	Stratum (clustering)
Independent Variables	
ADINST42	SAQ 12 Months: Dr. gave specific instructions (Rounds 4/2)
ADEZUN42	SAQ 12 Months: Dr. given instructions were easy to understand (Rounds 4/2)
ADTLHW42	SAQ 12 Months: Dr. asked to describe how you will follow instructions (Rounds 4/2)
Dependent Variable	
FLUSHT53	How long since last flu vaccination (Rounds 5/3)
Covariates/Inclusion Criteria	
ADAPPT42	SAQ 12 Months: # visits to medical office for care (Rounds 4/2)
AGELAST	Age
SEX	Sex
RACETHX	Race/Ethnicity

Third, merge and sort the combined dataset with only the variables from the 2015 and 2016 full consolidated files that are needed to meet your research aims. A sample SAS program for merging and sorting 2015-2016 MEPS full year consolidated files is provided in Box 7.5.

Box 7.5. Sample SAS program to merge and sort 2015-2016 MEPS full year consolidated data files



```
/*Create Temporary 2015 file and keep only
variables needed for analysis*/
data meps15 (rename = (SAQWT15F=SAQWTF));
SET MEPS15.H181; run;
data meps015 (keep = PANEL DUPERSID SAQWTF
VARPSU VARSTR ADINST42 ADEZUN42
ADTLHW42 FLUSHT53 ADAPPT42 AGELAST SEX
RACETHX); set meps15; run;
proc sort data=meps015; by DUPERSID PANEL;
run;

/*Create Temporary 2016 file and keep only
variables needed for analysis*/
data meps16 (rename = (SAQWT16F=SAQWTF));
SET MEPS16.H192; run;
data meps016 (keep = PANEL DUPERSID SAQWTF
VARPSU VARSTR ADINST42 ADEZUN42
ADTLHW42 FLUSHT53 ADAPPT42 AGELAST SEX
RACETHX); set meps16; run;
proc sort data=meps016; by DUPERSID PANEL;
run;

/*Merge and sort 2015-2016 Files*/
data mepsall; merge meps015 meps016; by
DUPERSID PANEL; run;
proc sort data=mepsall; by DUPERSID PANEL;
run;
```

Step 4: Recode and rename variables

Questionnaire responses often need to be recoded or responses collapsed prior to conducting statistical analysis. For example, the MEPS has response options "-9=Not ascertained," "-8=Don't know," "-7=Refused," and “-1=Inapplicable” for several questions. The responses are often removed and made “missing” prior to analysis. Furthermore, the numbers that represent certain values may need to be changed for easier interpretation of statistical analysis results. For example, MEPS has response options “1=Yes” and “2=No.” It is common practice to change “no” responses to 0, “0=No.” It is best practice to rename these recoded variables with a new variable name instead of replacing the original variable.

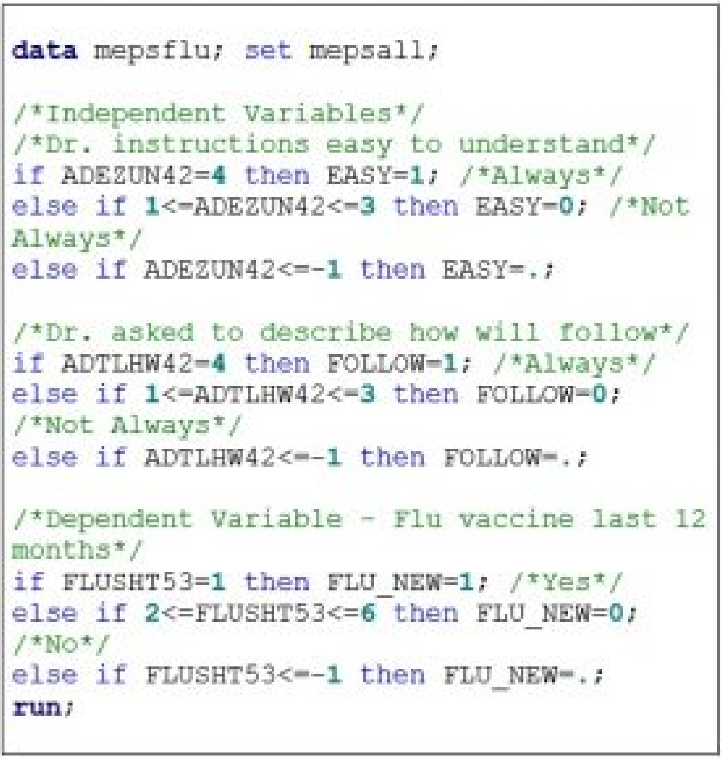
An overview of the variables recoded and renamed for analysis in this case study is provided in Table 7.5.

Table 7.5. Overview of MEPS variables recoded and renamed to meet research aims.

Question Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
Instructions easy to understand	ADEZUN42	-9=Not ascertained -1=Inapplicable 1=Never 2=Sometimes 3=Usually 4=Always	EASY	0=Not Always (combine never, sometimes, usually responses) 1=Always
Asked to describe how you will follow instructions	ADTLHW42	-9=Not ascertained -1=Inapplicable 1=Never 2=Sometimes 3=Usually 4=Always	FOLLOW	0=Not Always (combine never, sometimes, usually responses) 1=Always
How long since last flu vaccine	FLUSHT53	-9=Not ascertained -8=Don't know -7=Refused -1=Inapplicable 1=Within past year 2=Within 2 years 3=Within 3 years 4=Within 5 years 5=>5 years 6=Never	FLU_NEW	0=Not within past year (combine past 2 years, 3 years, 5 years, >5 years, never) 1=Within past year
Age	AGELAST	0-85 years	AGE_NEW	1=18-44 years 2=45-64 years 3=65 and older
Race/ Ethnicity	RACETHX	1=Hispanic 2=Non-Hispanic White 3=Non-Hispanic Black 4=Non-Hispanic Asian 5=Non-Hispanic other/multiple	RACEETH	1=Hispanic 2=Non-Hispanic White 3=Non-Hispanic Black 4=Non-Hispanic Other (including Asian/multiple)

A sample SAS program for recoding and renaming MEPS data for this case study is provided in Box 7.6.

Box 7.6. Sample SAS program to recode and rename MEPS variables



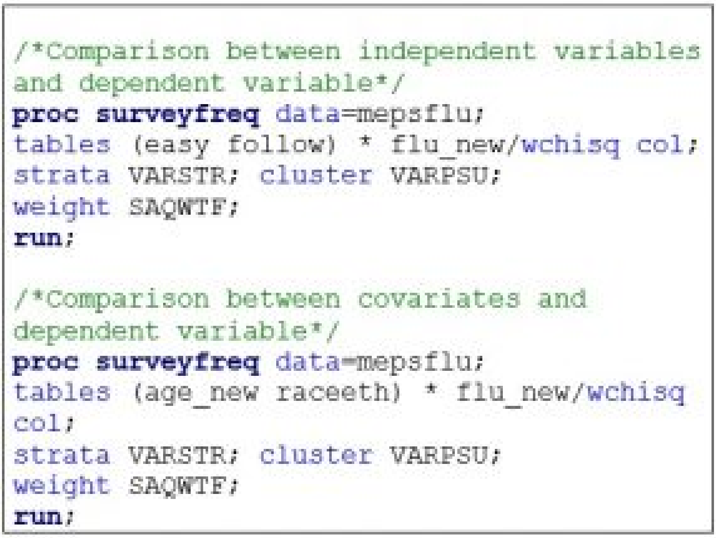
```
data mepsflu; set mepsall;

/*Independent Variables*/
/*Dr. instructions easy to understand*/
if ADEZUN42=4 then EASY=1; /*Always*/
else if 1<=ADEZUN42<=3 then EASY=0; /*Not
Always*/
else if ADEZUN42<=-1 then EASY=.;

/*Dr. asked to describe how will follow*/
if ADTLHW42=4 then FOLLOW=1; /*Always*/
else if 1<=ADTLHW42<=3 then FOLLOW=0;
/*Not Always*/
else if ADTLHW42<=-1 then FOLLOW=.;

/*Dependent Variable - Flu vaccine last 12
months*/
if FLUSHT53=1 then FLU_NEW=1; /*Yes*/
else if 2<=FLUSHT53<=6 then FLU_NEW=0;
/*No*/
else if FLUSHT53<=-1 then FLU_NEW=.;
run;
```

__Step 5: Conduct Descriptive Statistical Analysis__

Once all variables are recoded, collapsed, and renamed they can be used for statistical analysis. Statistical analysis should always start with descriptive analysis to describe the data source. Chi square analyses should be conducted to make categorical comparisons between the independent variable, covariates, and dependent variables. It is important to remember that all analysis of MEPS data needs to be conducted with SAS survey procedures due to the complex sample design. Weighting (variable: SAQWTF), primary sampling unit (variable: VARPSU) and cluster (variable: VARSTR) variables must be included in the programming statements.

A sample SAS program for conducting chi-square tests using MEPS data for this case study is provided in Box 7.7.

Box 7.7. Sample SAS program for running descriptive statistics (chi-square)

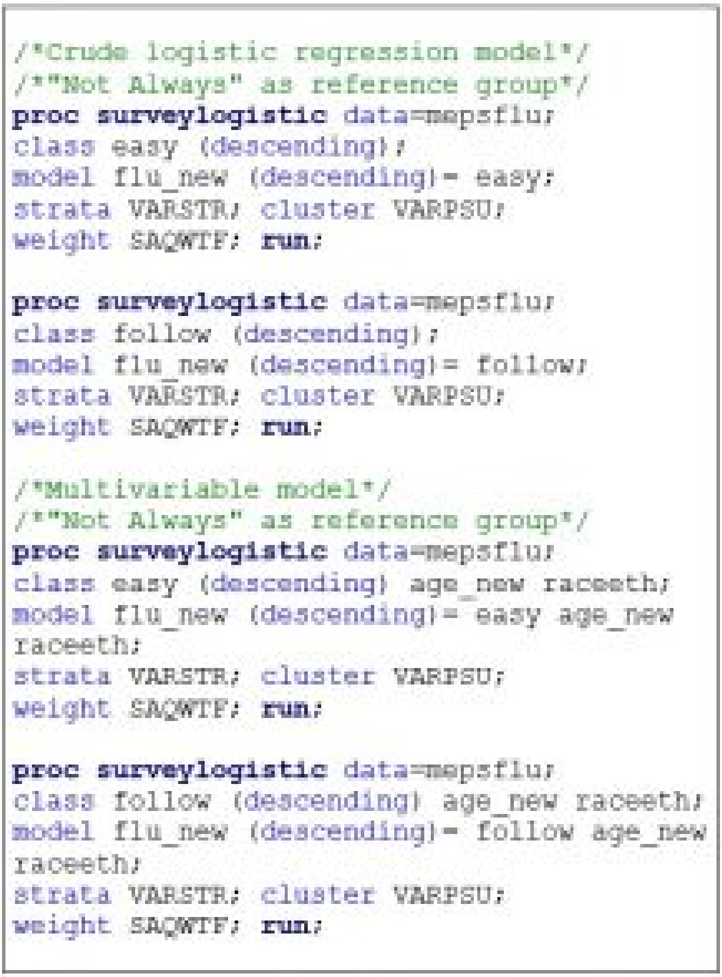


```
/*Comparison between independent variables
and dependent variable*/
proc surveyfreq data=mepsflu;
tables (easy follow) * flu_new/wchisq col;
strata VARSTR; cluster VARPSU;
weight SAQWTF;
run;

/*Comparison between covariates and
dependent variable*/
proc surveyfreq data=mepsflu;
tables (age_new raceeth) * flu_new/wchisq
col;
strata VARSTR; cluster VARPSU;
weight SAQWTF;
run;
```

Step 5: Conduct Inferential Statistical Analysis

After calculating descriptive statistics, inferential statistical analysis can be conducted. Crude and multivariable logistic regression models can be calculated to determine associations between each quality of PPC (gave instructions that were easy to understand, asked you to describe how you will follow instructions) and flu vaccine uptake in the last 12 months. Crude logistic regression models are used to determine the association between the independent and dependent variables without adjusting for other factors. Multivariable logistic regression models are used to determine associations between the independent and dependent variables after adjusting for potential covariates (e.g. age, race/ethnicity). Reference categories for the independent variables are needed. For this analysis, the reference group is "not always." Results compare adults who reported their health care provider "always" exhibited each PPC quality to those whose provider did "not always" exhibit each specific PPC quality. A sample SAS program for conducting logistic regression analysis using MEPS data for this case study is provided in Box 7.8.

Box 7.8. Sample SAS program for running MEPS inferential statistics (logistic regression)



```
/*Crude logistic regression model*/
/*"Not Always" as reference group*/
proc surveylogistic data=mepsflu;
class easy (descending);
model flu_new (descending)= easy;
strata VARSTR; cluster VARPSU;
weight SAQWTF; run;

proc surveylogistic data=mepsflu;
class follow (descending);
model flu_new (descending)= follow;
strata VARSTR; cluster VARPSU;
weight SAQWTF; run;

/*Multivariable model*/
/*"Not Always" as reference group*/
proc surveylogistic data=mepsflu;
class easy (descending) age_new raceeth;
model flu_new (descending)= easy age_new
raceeth;
strata VARSTR; cluster VARPSU;
weight SAQWTF; run;

proc surveylogistic data=mepsflu;
class follow (descending) age_new raceeth;
model flu_new (descending)= follow age_new
raceeth;
strata VARSTR; cluster VARPSU;
weight SAQWTF; run;
```

7.7 SUMMARY

This chapter provided an overview of the MEPS and ways to conduct basic statistical analysis using 2015-2016 public-use data files. The MEPS case study explored whether adults whose health care provider's always demonstrated clear communication (gave instructions that were easy to understand) and the teach-back method (asked them to describe how they will follow instructions) were more likely to receive a flu vaccine. Sample SAS programming statements were provided for downloading and inputting data files, merging data files, recoding and renaming variables, and conducting categorical descriptive and inferential statistical analysis. The dataset and full SAS programming statements for the MEPS case study are available n the Chapter 7 folder, in the Open ICPSR data repository.

7.8 REFERENCES

1. Cohen SB. Design strategies and innovations in the medical expenditure panel survey. Med Care. 2003;41(7 Suppl):III5-III12. doi:10.1097/01.MLR.0000076048.11549.71
2. Agency for Healthcare Research and Quality (AHRQ). MEPS HC-201: 2017 Full Year Consolidated Data File. Accessed September 2, 2020. https://meps.ahrq.gov/data_stats/download_data/pufs/h201/h201doc.pdf
3. National Center for Health Statistics. Survey Description, National Health Interview Survey, 2018.; 2019.
4. Agency for Healthcare Research and Quality (AHRQ). Medical Expenditure Panel Survey Home. Accessed September 2, 2020. https://meps.ahrq.gov/mepsweb/
5. Agency for Healthcare Research and Quality

(AHRQ). Medical Expenditure Panel Survey Insurance/Employer Component Overview. Accessed September 2, 2020. https://meps.ahrq.gov/mepsweb/survey_comp/Insurance.jsp

6. Agency for Healthcare Research and Quality (AHRQ). MEPS HC-199: 2017 Medical Conditions. Accessed September 2, 2020. https://meps.ahrq.gov/data_stats/download_data/pufs/h199/h199doc.pdf
7. MEPS HC-202: Panel 21 Longitudinal Data File. Accessed September 2, 2020. https://meps.ahrq.gov/data_stats/download_data/pufs/h202/h202doc.shtml
8. NHIS – 2017 Data Release. Published May 10, 2019. Accessed September 2, 2020. https://www.cdc.gov/nchs/nhis/nhis_2017_data_release.htm
9. Morgan PA, Strand J, Østbye T, Albanese MA. Missing in action: care by physician assistants and nurse practitioners in national health surveys. Health Serv Res. 2007;42(5):2022-2037. doi:10.1111/j.1475-6773.2007.00700.x
10. Kindratt TB. The Influence of Face-to-face and E-mail Patient-provider Communication on Breast, Cervical, and Colorectal Cancer Screenings. Texas Medical Center Dissertations (via ProQuest). Published online January 1, 2018:1-119.
11. Kindratt TB, Atem F, Dallo FJ, Allicock M, Balasubramanian BA. The Influence of Patient–Provider Communication on Cancer

Screening. Journal of Patient Experience. Published online May 11, 2020:2374373520924993. doi:10.1177/2374373520924993

12. Kindratt TB, Dallo FJ, Allicock M, Atem F, Balasubramanian BA. The influence of patient-provider communication on cancer screenings differs among racial and ethnic groups. Prev Med Rep. 2020;18:101086. doi:10.1016/j.pmedr.2020.101086
13. Villani J, Mortensen K. Patient-provider communication and timely receipt of preventive services. Prev Med. 2013;57(5):658-663. doi:10.1016/j.ypmed.2013.08.034
14. Zuvekas SH, Kashihara D. The Impacts of the COVID-19 Pandemic on the Medical Expenditure Panel Survey. Am J Public Health. 2021;111(12):2157-2166. doi:10.2105/AJPH.2021.306534

7.9 COVID-19 PANDEMIC CHANGES

Sections 7.1 to 7.8 were written during the initial waves of the COVID-19 pandemic. Since the MEPS conducts in-person surveys at participants' households, there were significant disruptions to the regular methodology due to stay-at-home orders and safety concerns for both participants and the data collectors.[14] With the declaration of a national emergency on March 13, 2020, changes were needed to adhere to essential work-based restrictions, collect data more effectively, and enhance survey content that could be used for evidence-based public health decision-making. All data collection procedures were halted on March 17, 2020, and in-person data collection was switched to telephone through Fall 2020.[14] New content was collected on telehealth visits, COVID-19 specific impacts on health care utilization, and COVID-19 vaccination uptake. Full details of the changes are reported elsewhere.[14]

CHAPTER 8.

HEALTH INFORMATION NATIONAL TRENDS SURVEY

8.1 INTRODUCTION

Chapter 8 covers the Health Information National Trends Survey (HINTS). The HINTS has been collected by the National Cancer Institute (NCI) since 2003 to monitor national trends in health communication, information technology use, and knowledge, attitudes, and practices towards cancer prevention and care.[1] The first iteration (HINTS 1) was collected in 2003. HINTS 2 and HINTS 3 were collected in 2005 and 2007-2008 respectively. Beginning with HINTS 4, data collection was split into cycles. HINTS 4 has four cycles collected in 2011 (Cycle 1), 2012 (Cycle 2), 2013 (Cycle 3) and 2014 (Cycle 4). In 2015, the United States Federal Drug Administration (FDA) partnered with the NCI to evaluate tobacco use and related communications, and public knowledge, beliefs, and behaviors regarding dietary supplements and medical products.[1] This chapter includes details on: how data are collected; how data are made publicly available as machine-actionable data files; what variables must be included to address design features of the complex sample; the strengths and limitations of the survey; and practical tips for conducting statistical analysis; and how to answer research questions using a case study. The practical tips provided for analysis of HINTS data are based on the author's previous experiences analyzing HINTS 4 data to answer questions related to associations between predisposing and enabling factors that contribute to morbidity, mortality and health services use. The HINTS case study will explore how e-mail communication between patients and health care providers between visits influences colon cancer screening uptake. The bulk of the chapter will comprise of section *8.6: Case Study* in order to give the reader hands-on practice downloading and cleaning large databases and conducting basic categorical data analysis using PROC SURVEYFREQ and PROC SURVEYLOGISTIC. The syntax provided was created for use with SAS 9.4.

8.2 DATA COLLECTION

Data collection methods for HINTS have evolved over the past 17 years to increase participation using information technology. Surveys can be completed in English or Spanish. The sampling frame includes two strata to ensure inclusion of minority and non-minority populations. The most recent iteration as of this writing (HINTS 5 Cycle 3) includes self-administered and web-based options. Prior to HINTS 5, self-administered questionnaires were collected by mail.[1] HINTS 5 Cycle 3 included as "web pilot" with two experimental conditions ("Web Option" or "Web Bonus"). Participants randomly selected for the Web Option group choose whether they wanted to complete the survey by paper or online. Other participants randomly selected for the Web Bonus group were given the same choices but offered a $10 incentive to complete the survey online. Participants recruited to complete the self-administered questionnaire by mail received the initial survey, a reminder postcard, and up to two follow-up mailings with additional copies of the survey. Participants recruited to complete the web-based options received instructions by mail with a website link and pin number/access code to complete the survey. It was requested that the adult with the "next-birthday" in the household complete the questionnaires. For example, if there were two adults in the household (one with a birthday in May, the other birthday was in October), the adult whose birthday was in May would be requested to fill out the survey if it was receiving in January. A $2 incentive was included with this questionnaire to promote participation. HINTS 5 Cycle 4 data were not available as of this writing. Further details of the HINTS sample design and data collection methods are reported on the HINTS website and published research.[1–4]

8.3 DATA FILE

The HINTS is comprised of one data file per cycle, which can be combined with other iterations and cycles to increase sample size. Each HINTS administration includes questions related to the measurement of the following core constructs: sociodemographics; technology use and access; health care use and access; health information-seeking; cancer prevention and screening knowledge and behavior; cancer risk perceptions; and cancer-related behavior.[2] The HINTS 3 Cycle 5 questionnaire includes questions related to the following 15 topics, represented by A through O (no Section I):[5]

- A) Looking for health information
- B) Using the internet to find information
- C) Your health care
- D) Medical records
- E) Caregiving
- F) Your overall health
- G) Health and nutrition
- H) Physical activity and exercise
- J) Sun & UV exposure
- K) Tobacco products
- L) Cancer screening and awareness
- M) Your cancer history
- N) Beliefs about cancer
- O) You and your household

8.4. STRENGTHS AND LIMITATIONS

A strength of the HINTS is its measurement of the rapidly changing health information and communication landscape using nationally representative samples. The HINTS is unique in its ability to provide data on cancer patients and survivors. The public-use data tools are easily and readily available on the website and several supporting documents are provided to support data users.[6] HINTS abides by the Findability, Accessibility, Interoperability, and Reusability (FAIR) principles for data management.[2] Several survey questions are collected across survey iterations and cycles allowing for the examination of trends over time or collapsing to increase the sample size. Despite its strengths as a cross-sectional survey, results from the HINTS cannot be used to determine causality. Questions measuring health behaviors during the same time period (e.g. last 12 months) may suffer from temporality bias due to the inability to determine whether an exposure (e.g. communication with health care provider by e-mail or patient portal) occurred before the outcome (e.g. cancer screenings). Due to the small sample size, statewide estimates cannot be determined. However, the HINTS allows for regional estimates by Census region and divisions. Similar to other national health surveys, response rates have decreased over time. HINTS results may be biased due to large numbers of missing data. However, the HINTS recommends jackknife weighting processes, instead of Taylor-Linearization methods, to address this bias.[7]

8.5. DESIGN FEATURES

Data analysts must use special procedures to account for the complex sample design used by the HINTS. Although analysts can use complex survey methods similar to other national health surveys (e.g. cluster, stratification variables), the recommended approach for analyzing HINTS data is to use jackknife replicate weights to ensure the computation of the correct variance estimates.[8] Each HINTS cycle includes a set of 50 replicate weights. A final sample weight is used to calculate population estimates and 50 replicate weights are used to calculate accurate standard errors of the estimates for the combined sample. A final sample weight is used to calculate population estimates and 50 replicate weights are used to calculate accurate standard errors of the estimates for the paper-only sample. Then, a final sample weight is used to calculate population estimates and 50 replicate weights are used to calculate accurate standard errors of the estimates for the web-option sample. Next, a final sample weight is used to calculate population estimates and 50 replicate weights are used to calculate accurate standard errors of the estimates for the web-bonus sample. Finally, a final sample weight is used to calculate population estimates and 150 replicate weights are used to calculate accurate standard errors of the estimates for the combined sample, controlling for group differences. Although not the recommended approach for population estimates, stratum and cluster variables are available for analysis using Taylor-series linearization methods.[8] An overview of the weights, cluster and stratum variables are provided in Table 8.1.

TABLE 8.1. OVERVIEW OF IMPORTANT ANALYTIC VARIABLES FOR WEIGHTING AND COMPLEX SAMPLE CHARACTERISTICS, HINTS 5 CYCLE 3

	Final Sample Weight	**Jackknife Replication Methods**	
		Replicate Weights	**Degrees of Freedom**
Combined Sample Taylor Linearization Methods *VAR_STRATUM and VAR_CLUSTER*	TG_all_FINWT0	TG_all_FINWT0 through TG_all_FINWT50	49
Paper-Only	TG1_all_FINWT0	TG1_all_FINWT0 through TG1_all_FINWT50	49
Web-Option	TG2_all_FINWT0	TG2_all_FINWT0 through TG2_all_FINWT50	49
Web-Bonus	TG3_all_FINWT0	TG3_all_FINWT0 through TG3_all_FINWT50	49
Combined Sample *(Control for Group Differences)*	NWGT0	NWGT0 through NWGT150	147

8.6. HINTS CASE STUDY: THE INFLUENCE OF E-MAIL PPC ON COLON CANCER SCREENING

Communicating with health care providers by e-mail, patient portals, mobile applications and text messaging has increased substantially over the past several years. E-mail patient-provider communication (PPC) describes the communication between health care providers and patients between visits using e-mail or direct communication through patient portals.[9] E-mail PPC use differs by age, sex, education, race/ethnicity, and history of chronic diseases such as diabetes and hypertension.[9] Few studies have evaluated how e-mail PPC may improve individual's likelihood of receiving preventive services. Kindratt and colleagues previous research found mixed results when evaluating the influence of e-mail PPC on preventive service utilization. Using National Health Interview Survey data, they found that adults who used e-mail to communicate with their provider had greater odds of reporting they received an influenza vaccine, mammogram, pap test, and colon cancer screening in the past 12 months.[10-11] However, using HINTS 4 Cycles 1 through 4 data, Kindratt and colleagues did not find any associations between e-mail PPC and cancer screening uptake.[12] In this case study, we will determine the association between e-mail PPC and adults' likelihood of receiving a colon cancer screening using HINTS 5 Cycle 3 data.

8.6.1 SPECIFIC AIMS

- Aim 8.1: Compare sociodemographic and health-related characteristics of adults who use e-mail to communicate with their health care provider

- Aim 8.2: Determine associations between e-mail PPC and adults' likelihood of receiving a colon cancer screening before and after controlling for covariates.

8.6.2 METHODS

Complete the following steps to download, clean, recode and analyze HINT 5 Cycle 3 data to answer the specific aims.

Step 1: Download HINTS 5 Cycle 3 Public Use Data Files

- Go to the HINTS website to access public use data for download
- Review "HINTS Data Terms of Use"
- Check the box at the bottom to indicate you will comply with terms of use
- Enter your e-mail address and click "Accept"
- You will be taken to an updated page with "Public Use Datasets"
- Under the heading "HINTS 5, Cycle 3 (2019) dataset, updated March 2020," click on the "SAS data and supporting documents" link to download a zip file
- Unzip the file and save the documents to your computer. I recommend creating a folder on the "C Drive" labeled HINTS and separated by each iteration (if using more than 1 iteration or cycle). This will be consistent with the location statements used in the textbook examples (HINTS 5.3 representing HINTS 5 Cycle 3).

The downloaded zip file should contain the following files:

1. HINTS 5 Cycle 3 Public Codebook
2. HINTS 5 Cycle 3 Public Format Assignments
3. HINTS 5 Cycle 3 Public Formats
4. HINTS 5 Cycle 3 Public History Document
5. HINTS 5 Cycle 3 Survey Overview & Data Analysis Recommendations
6. HINTS 5 Cycle 3 Methodology Report
7. HINTS 5 Cycle 3 SAS Public-Use Dataset
8. HINTS 5 Cycle3 Annotated Instruments English and Spanish
9. HINTS 5 Cycle3 Web Pilot Results Report

The most useful files for conducting the statistical analysis in SAS are the public codebook, public format assignments, public formats, and public use dataset files. The public codebook contains an overview of all variable names, labels, formats, response options, weighted and unweighted sample sizes and proportions. SAS programming statements (.sas files) are provided in the public format assignments and public formats files. The statements can be used to apply labels to your data file. For example, if you do not use the formatting files, your outputs will read "1" as responses instead of "yes" to indicate the actual responses from the survey. The public use dataset files include the responses to each variable in numerical format.

Step 2: Run SAS programming statements to create library and labels

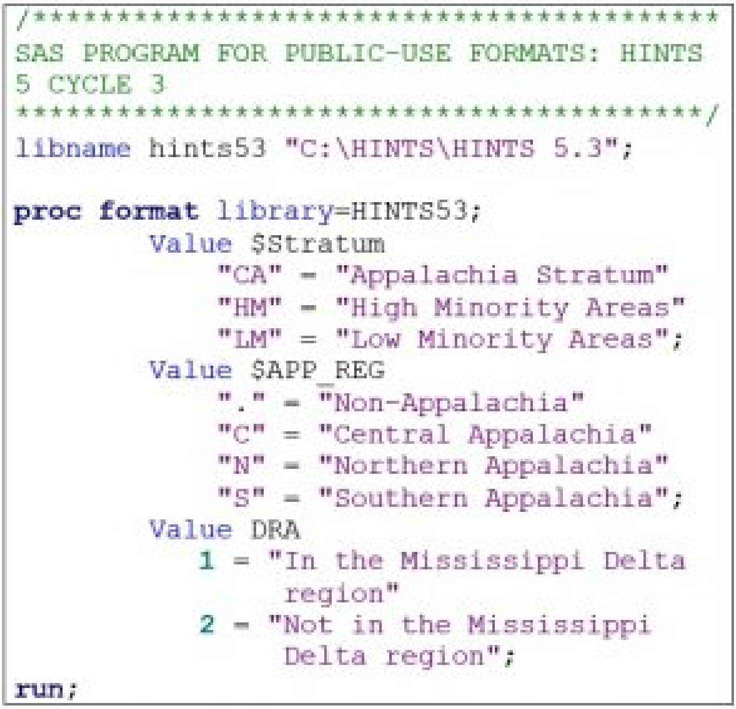
for HINTS 5 Cycle 3 Data

Sample SAS programs to create the libraries and format the HINTS 5 Cycle 3 data with labels are provided in Box 8.1 and Box 8.2, respectively. The full SAS programs are available for download in the chapter 8 folder in the Open ICPSR data repository.

HINTS 5 Cycle 3 SAS Public Formats File

- Open the HINTS 5 Cycle 3 Public Formats File
- Create a LIBNAME statement that houses the data and files associated with the analysis. I recommend creating the LIBNAME statement as the survey name (e.g. "HINTS") and using the same location that the data files for saved in on the C drive (e.g. "C:\HINTS\HINTS 5.3")
- Highlight all programming statements and click RUN.

Box 8.1. SAS program for public formats, HINTS 5 Cycle 3

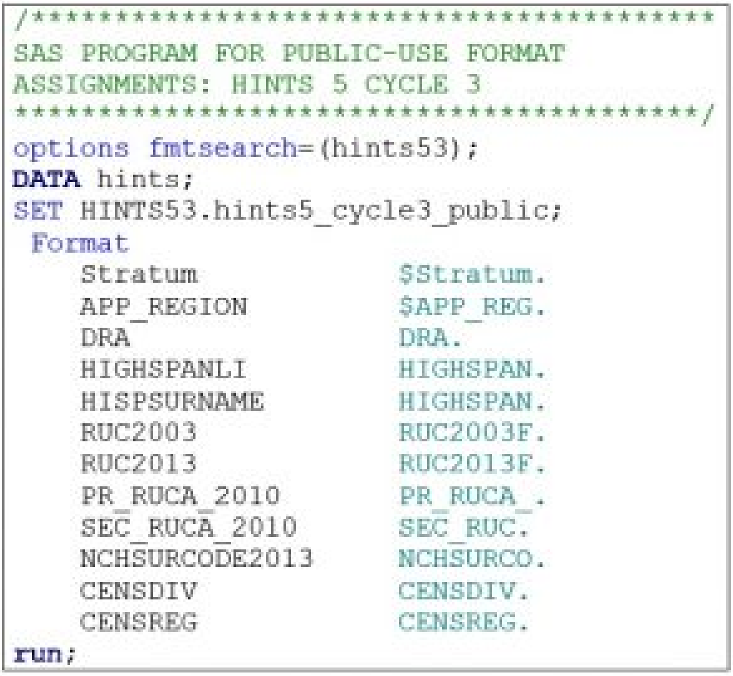


```
/****************************************
SAS PROGRAM FOR PUBLIC-USE FORMATS: HINTS
5 CYCLE 3
****************************************/
libname hints53 "C:\HINTS\HINTS 5.3";

proc format library=HINTS53;
        Value $Stratum
            "CA" = "Appalachia Stratum"
            "HM" = "High Minority Areas"
            "LM" = "Low Minority Areas";
        Value $APP_REG
            "." = "Non-Appalachia"
            "C" = "Central Appalachia"
            "N" = "Northern Appalachia"
            "S" = "Southern Appalachia";
        Value DRA
           1 = "In the Mississippi Delta
                region"
           2 = "Not in the Mississippi
                Delta region";
run;
```

HINTS 5 Cycle 3 SAS Public Format Assignments File

- Open the HINTS 5 Cycle 3 Public Format Assignments file
- Enter the LIBNAME in the “options fmtsearch” statement and update file name
- Highlight all programming statements and click RUN

Box 8.2. SAS program for public format assignments, HINTS 5 Cycle 3



```
/*****************************************
SAS PROGRAM FOR PUBLIC-USE FORMAT
ASSIGNMENTS: HINTS 5 CYCLE 3
*****************************************/
options fmtsearch=(hints53);
DATA hints;
SET HINTS53.hints5_cycle3_public;
 Format
    Stratum                $Stratum.
    APP_REGION             $APP_REG.
    DRA                    DRA.
    HIGHSPANLI             HIGHSPAN.
    HISPSURNAME            HIGHSPAN.
    RUC2003                RUC2003F.
    RUC2013                RUC2013F.
    PR_RUCA_2010           PR_RUCA_.
    SEC_RUCA_2010          SEC_RUC.
    NCHSURCODE2013         NCHSURCO.
    CENSDIV                CENSDIV.
    CENSREG                CENSREG.
run;
```

Step 3: Select Variables for HINTS 5 Cycle 3 Analysis

Once formats and labels have been assigned to the dataset, you can remove any variables that are not needed for your analysis. This will reduce the size of the dataset and make processing time quicker when running SAS programming statements. In this case study, I have kept the following variables (Table 8.2) to denote the survey design features and creation of the independent variable, dependent variable and selected covariates.

Table 8.2. Overview of variables used for HINTS case study

Variable Name	Variable Description
Design Variables	
TG_all_FINWT0	Final person-level sample weight – all modalities combined
TG_all_FINWT1 – TG_all_FINWT50	Final person-level replicate weights 1-50 – all modalities combined
Independent Variables	
Electronic_TalkDoctor	In the past 12 months have you used a computer, smart phone, or other electronic means to use e-mail or the internet to communicate with a doctor?
EverTestedColonCa	Have you ever had one of these tests to check for colon cancer?
Covariates/Inclusion Criteria	
FreqGoProvider	In the past 12 months, not counting times you went to an emergency room, how many times did you go to a doctor, nurse, or other health professional to get care for yourself?
Age	What is your age?
SelfGender	Self-reported gender
Race_Cat2	Derived variable to categorize responses given in O6 (Race)

Step 4: Recode and rename variables

Questionnaire responses often need to be recoded or responses collapsed prior to conducting statistical analysis. For example, the HINTS has response options "-9=Missing Data (Not ascertained)" and "-7=Missing data (Web-partial, Question Never Seen)" for several questions. The responses are often removed and made "missing" prior to analysis. Furthermore, the numbers that represent certain values may need to be changed for easier interpretation of statistical analysis results. For example, HINTS has response options “1=Yes” and “2=No.” It is common practice to change “no” responses to 0, “0=No.” It is best practice to rename these recoded variables with a new variable name instead of replacing the original variable.

An overview of the variables recoded and renamed for analysis in this case study is provided in Table 8.3.

Table 8.3. Overview of HINTS variables recoded and renamed to meet research aims

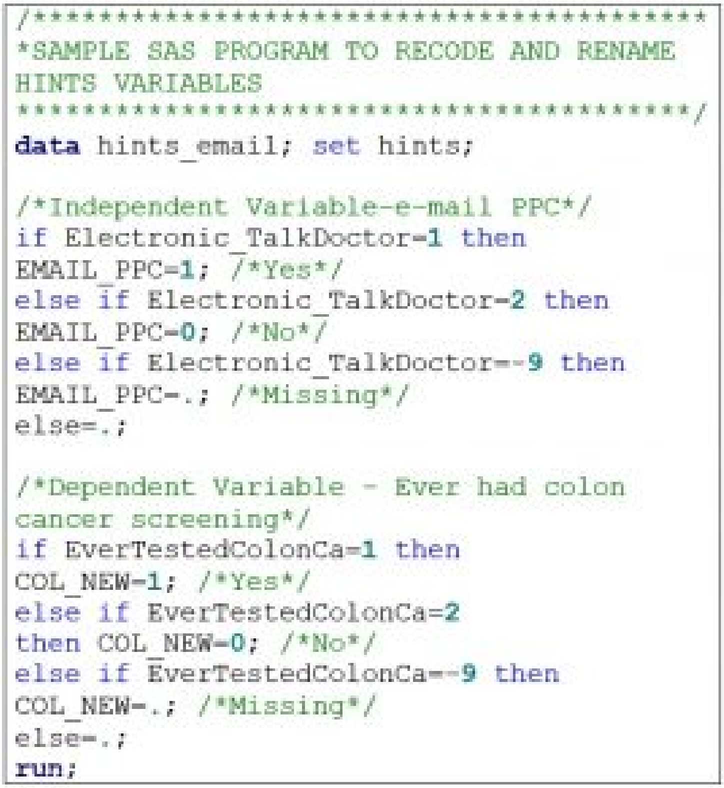
Question Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
In the past 12 months have you used a computer, smart phone, or other electronic means to use e-mail or the internet to communicate with a doctor or a doctor’s office?	Electronic_TalkDoctor	-9=Missing data (Not Ascertained) 1=Yes 2=No	EMAIL_PPC	0=No 1=Yes
Have you ever had one of these tests to check for colon cancer?	EverTestedColonCa	-9=Missing data -7=Missing data (Web partial) 1=Yes 2=No	COL_NEW	0=No 1=Yes

Table 8.3 (continued). Overview of HINTS variables recoded and renamed to meet research aims

Question Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
What is your age?	Age	-9=Missing data (Not Ascertained) -4=Unreadable non-conforming numeric response 18-98 years	AGE_NEW	1=50-59 years 2=60-69 years 3=70 and older
Self: Gender	SelfGender	-9=Missing data (Not Ascertained) -7=Missing data (Web partial – Question Never Seen) 1=Male 2=Female	GENDER	1=Male 2=Female

A sample SAS program for recoding and renaming HINTS 5 Cycle 3 data for this case study is provided in Box 8.3.

Box 8.3. Sample SAS program to recode and rename HINTS variables



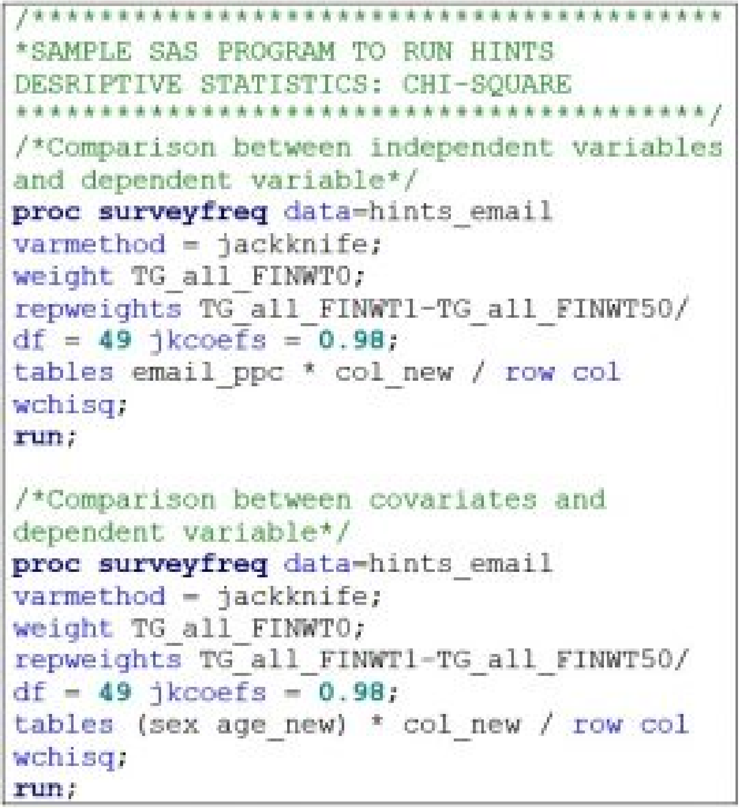
```
/****************************************
*SAMPLE SAS PROGRAM TO RECODE AND RENAME
HINTS VARIABLES
****************************************/
data hints_email; set hints;

/*Independent Variable-e-mail PPC*/
if Electronic_TalkDoctor=1 then
EMAIL_PPC=1; /*Yes*/
else if Electronic_TalkDoctor=2 then
EMAIL_PPC=0; /*No*/
else if Electronic_TalkDoctor=-9 then
EMAIL_PPC=.; /*Missing*/
else=.;

/*Dependent Variable - Ever had colon
cancer screening*/
if EverTestedColonCa=1 then
COL_NEW=1; /*Yes*/
else if EverTestedColonCa=2
then COL_NEW=0; /*No*/
else if EverTestedColonCa=-9 then
COL_NEW=.; /*Missing*/
else=.;
run;
```

Step 5: Conduct Descriptive Statistical Analysis

Once all variables are recoded, collapsed, and renamed they can be used for statistical analysis. Statistical analysis should always start with descriptive analysis to describe the data source. Chi square analyses should be conducted to make categorical comparisons between the independent variable, covariates, and dependent variables. It is important to remember that all analysis of HINTS data needs to be conducted with SAS Survey procedures due to the complex sample design. It is recommended to use the final sample weights (variable: TG_all_FINWT0), replicate weights (variables TG_all_FINWT1 to TG_all_FINWT50) and degrees of freedom (50-1=49) for the "delete one" jackknife replication method. More details on the replicate weighting are available in the HINTS 5 Cycle 3 Methodology Report.[8]

A sample SAS program for conducting chi-square tests using HINTS Cycle 3 data for this case study is provided in Box 8.4.

Box 8.4. Sample SAS program for running descriptive statistics (chi-square)

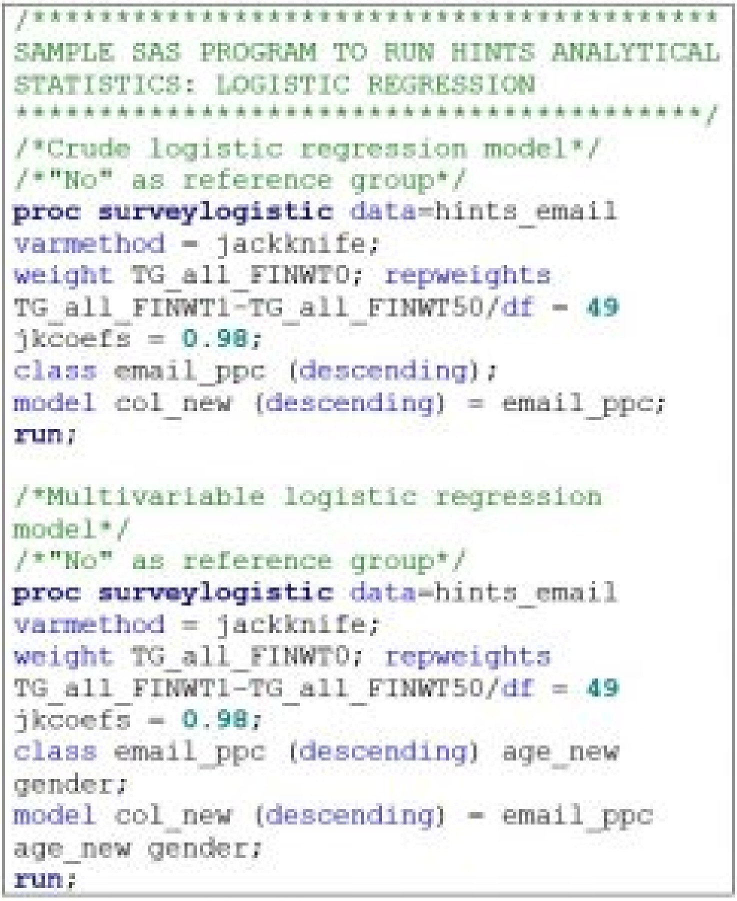


```
/****************************************
*SAMPLE SAS PROGRAM TO RUN HINTS
DESRIPTIVE STATISTICS: CHI-SQUARE
****************************************/
/*Comparison between independent variables
and dependent variable*/
proc surveyfreq data=hints_email
varmethod = jackknife;
weight TG_all_FINWT0;
repweights TG_all_FINWT1-TG_all_FINWT50/
df = 49 jkcoefs = 0.98;
tables email_ppc * col_new / row col
wchisq;
run;

/*Comparison between covariates and
dependent variable*/
proc surveyfreq data=hints_email
varmethod = jackknife;
weight TG_all_FINWT0;
repweights TG_all_FINWT1-TG_all_FINWT50/
df = 49 jkcoefs = 0.98;
tables (sex age_new) * col_new / row col
wchisq;
run;
```

Step 6: Conduct Inferential Statistical Analysis

After calculating descriptive statistics, inferential statistical analysis can be conducted. Crude and multivariable logistic regression models can be calculated to determine associations between e-mail PPC (using computer, smart phone, or other electronic means to use e-mail or the internet to communicate with a doctor or a doctor's office) and colon cancer screening. Crude logistic regression models are used to determine the association between the independent and dependent variables without adjusting for other factors. Multivariable logistic regression models are used to determine associations between the independent and dependent variables after adjusting for potential covariates (e.g. age, sex). A reference category for the independent variables is needed. For this analysis, the reference group is "No." Results compare adults did and did not use a computer, smart phone, or other electronic means to use e-mail or the internet to communicate with their doctor or doctor's office. A sample SAS program for conducting logistic regression analysis using HINTS 5 Cycle 3 data for this case study is provided in Box 8.5.

Box 8.5. Sample SAS program for running HINTS inferential statistics (logistic regression)



```
/*****************************************
SAMPLE SAS PROGRAM TO RUN HINTS ANALYTICAL
STATISTICS: LOGISTIC REGRESSION
*****************************************/
/*Crude logistic regression model*/
/*"No" as reference group*/
proc surveylogistic data=hints_email
varmethod = jackknife;
weight TG_all_FINWT0; repweights
TG_all_FINWT1-TG_all_FINWT50/df = 49
jkcoefs = 0.98;
class email_ppc (descending);
model col_new (descending) = email_ppc;
run;

/*Multivariable logistic regression
model*/
/*"No" as reference group*/
proc surveylogistic data=hints_email
varmethod = jackknife;
weight TG_all_FINWT0; repweights
TG_all_FINWT1-TG_all_FINWT50/df = 49
jkcoefs = 0.98;
class email_ppc (descending) age_new
gender;
model col_new (descending) = email_ppc
age_new gender;
run;
```

8.7 SUMMARY

This chapter provided an overview of the HINTS and ways to conduct basic statistical analysis using HINTS 5 Cycle 3 public-use data files. The HINTS case study explored whether adults who used e-mail to communicate with their health care provider were more or less likely to receive a colon cancer screening. Sample SAS programming statements were provided for downloading data files, labeling and formatting data files, recoding and renaming variables, and conducting categorical descriptive and inferential statistical analysis. The dataset and full SAS programming statements for the HINTS case study are available in the chapter 8 folder in the Open ICPSR data repository.

8.8 REFERENCES

1. National Cancer Institute. Health Information National Trends Survey (HINTS): Overview of the HINTS 5 Cycle 3 Survey and Data Analysis Recommendations, January 2020.
2. Finney Rutten LJ, Blake KD, Skolnick VG, Davis T, Moser RP, Hesse BW. Data Resource Profile: The National Cancer Institute's Health Information National Trends Survey (HINTS). *Int J Epidemiol*. 2020;49(1):17-17j. doi:10.1093/ije/dyz083
3. Finney Rutten LJ, Davis T, Beckjord EB, Blake K, Moser RP, Hesse BW. Picking up the pace: changes in method and frame for the health information national trends survey (2011-2014). *J Health Commun*. 2012;17(8):979-989. doi:10.1080/10810730.2012.700998
4. Nelson DE, Kreps GL, Hesse BW, et al. The Health Information National Trends Survey (HINTS): development, design, and dissemination. *J Health Commun*. 2004;9(5):443-460; discussion 81-84. doi:10.1080/10810730490504233

5. National Institutes of Health. Health Information National Trends Survey (HINTS): Annotated Form Cycle 3/Web Pilot, English Version.

6. National Cancer Institute. What is HINTS? https://hints.cancer.gov/

7. Maitland A, Lin A, Cantor D, et al. A Nonresponse Bias Analysis of the Health Information National Trends Survey (HINTS). *J Health Commun*. 2017;22(7):545-553. doi:10.1080/10810730.2017.1324539

8. Westat. *Health Information National Trends Survey 5 (HINTS 5) Cycle 3 Methodology Report, August 2019*.

9. Ye J, Rust G, Fry-Johnson Y, Strothers H. E-mail in patient-provider communication: a systematic review. *Patient Educ Couns*. 2010;80(2):266-273. doi:10.1016/j.pec.2009.09.038

10. Kindratt T, Callender L, Cobbaert M, Wondrack J, Bandiera F, Salvo D. Health information technology use and influenza vaccine uptake among US adults. *Int J Med Inform*. 2019;129:37-42. doi:10.1016/j.ijmedinf.2019.05.025

11. Kindratt TB, Allicock M, Atem F, Dallo FJ, Balasubramanian BA. Email Patient-Provider Communication and Cancer Screenings Among US Adults: Cross-sectional Study. *JMIR Cancer*. 2021;7(3):e23790. doi:10.2196/23790

12. Kindratt TB, Atem F, Dallo FJ, Allicock M, Balasubramanian BA. The Influence of Patient–Provider Communication on Cancer

Screening. *Journal of Patient Experience*. Published online May 11, 2020:2374373520924993. doi:10.1177/2374373520924993

CHAPTER 9.

BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM

9.1 INTRODUCTION

Chapter 9 covers the Behavioral Risk Factor Surveillance System (BRFSS). The BRFSS has been conducted by the Centers for Disease Control and Prevention (CDC) since 1984 to collect data on health behaviors, physical activity, diet, hypertension and preventive safety measures (e.g. seat-belt use) among US adults.[1]This chapter includes details on: how data are collected; how data are made publicly available as machine-actionable data files; what variables must be included to address design features of the complex sample; the strengths and limitations of the survey; and practical tips for conducting statistical analysis; and how to answer research questions using a case study. The practical tips provided for analysis of BRFSS data are based on the primary author's previous experiences analyzing 2014 and 2019 BRFSS data to answer questions related to associations between predisposing and enabling factors that contribute to morbidity, mortality and health services use. The BRFSS case study will explore how differences in caregiving experiences among urban and rural adults in Texas are moderated by race and ethnicity. The bulk of this chapter will comprise of section 9*.6: Case Study* in order to give investigators hands-on practice downloading and cleaning large databases and conducting basic categorical data analysis using PROC SURVEYFREQ and PROC SURVEYLOGISTIC. The syntax provided was created for use with SAS 9.4.

9.2 DATA COLLECTION

Since beginning in 1984, the BRFSS was expanded in 1988 to include optional modules, including chronic disease, health care access, and preventive services uptake. Some optional modules include data collection among children.[1] In 1993, the BRFSS was further expanded to become an annual national surveillance system. The BRFSS uses a random-digit-dialing cross-sectional study design to collect data using a computer-assisted telephone interview (CATI) system. The BRFSS is one of the largest health surveys collected worldwide with over 400,000 responses collected each year.[2] Prior to 2008, data were only collected from landline telephones. In 2008, the methodology was revised to conduct interviews using cell phones. Response rates are calculated for landline, cell phone, and combined responses for each state. National response rates are calculated as a median of response rates from each state. In 2019, the overall national combined response rate was 49.4% (range from (37.3%-73.1%).[3] Response rates were lowest in New York (39.5% landline, 33.5% cell phone, 37.3% overall) and highest in South Dakota (78.0%

landline, 33.5% cell phone, 73.1% overall). For Texas, the landline response rate was 50.6%, the cell phone response rate was 37.6%, and the combined response rate was 46.3%.[3] Further details of the BRFSS sampling design and data collection methods are reported on the BRFSS website.[4]

9.3 DATA FILES

The CDC provides a complete BRFSS public-use data file available annually, which can be combined with other years to increase sample sizes for the analysis of specific subgroups. The annual questionnaire includes three parts: 1) core component; 2) optional BRFSS modules; and 3) state-added questions.[5] In 2019, BRFSS grouped data into four datasets that combine landline and cellular responses. The main 2019 BRFSS Questionnaire data file can be used for research questions in the core section or common modules asked by the states. Separate BRFSS versions may be needed if states collected multiple versions of the questionnaires.[6]

9.3.1 CORE COMPONENT

The core component includes a standard set of questions used by all states. There is an annual core set of questions are asked every year and rotating core set of questions administered in odd- and even-numbered years. The core questions are selected from other established national surveys, including the National Health Interview Survey (NHIS) and National Health and Nutrition Examination Survey (NHANES) to allow for comparisons across survey methods.[5] Question topics include demographics, health-related perceptions, health conditions and health behaviors. Examples include questions examining health care access, health status, fruit and vegetable consumption, alcohol, and tobacco use.[5] Based on an agreement between state representatives and the CDC, states are required to administer the core component each year without modification.

9.3.2 OPTIONAL MODULES

Additional modules on specific topics are created each year. States vote to approve and adopt questions proposed by the CDC for inclusion in the optional modules. Then, states have the option to elect whether or not to use these questions with the core component on their annual survey.[5] Examples include questions on diabetes, skin cancer risk, shingles, cancer survivorship and the caregiving module used for the case study outlined in this chapter. Additional details on the optional modules are available on the BRFSS website.

9.3.3. STATE-ADDED QUESTIONS

States can also develop or obtain additional questions to be added to their BRFSS questionnaires for state programming purposes. In 2020, the Texas BRFSS added nine sections with state-added questions focused on 1) health access; 2) e-cigarette use; 3) reasons for not getting a shingles vaccine; 4) cancer survivorship; 5) food security; 6) oral health; 7) tobacco; 8) marijuana vaping; and 9) suicide attempts.[6] State-added questions are not available in the public use data files and must be obtained directly from the state BRFSS coordinators. More details on state-added questions for each state are available on each state-specific BRFSS website.

9.4. STRENGTHS AND LIMITATIONS

A strength of the BRFSS is a large sample size in comparison to other national datasets (e.g. NHIS or NHANES) due to the telephone data collection design. Another strengths is the ability for prevalence estimates for cancer screening and other health behaviors to be directly linked Healthy People 2020 and Healthy People 2030 objectives.[5,7] A limitation of the BRFSS is that all data are self-reported. The interviews are not conducted in person. There is an underrepresentation of rural counties in the sample design. The optional core modules are not administered by all US states and territories, which make it difficult to make comparisons between geographic contexts for those specific topics.[5]

9.5 DESIGN FEATURES

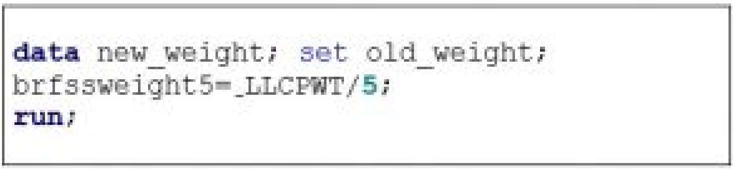
Data analysts must use special procedures to account for the complex sample design used by the BRFSS. Analytic procedures must include variables to adjust for the clustering, stratification, and weighting of each data file. The clustering (_PSU) and stratification (_STSTR) are the same regardless of the data file used. An overview of the BRFSS clustering, stratification, and weighting variables for 2019 are provided in Table 9.1.

TABLE 9.1. OVERVIEW OF BRFSS COMPLEX SAMPLE DESIGN VARIABLES

Data Description (Dataset)	Clustering	Stratification	Weight
2019 BRFSS Questionnaire Data (LLCP2019)	_PSU	_STSTR	_LLCPWT
2019 Combined Landline and Cell Phone Version 1 (LLCP18V1)	_PSU	_STSTR	LCPWTV1
2019 Combined Landline and Cell Phone Version 2 (LLCP18V2)	_PSU	_STSTR	LCPWTV2
2019 Combined Landline and Cell Phone Version 3 (LLCP18V3)	_PSU	_STSTR	LCPWTV3

When combining multiple years of NHIS data, investigators must divide the total annual weight by the total number of years (or multiply by 1/total years) in the merge prior to conducting their statistical analysis. For example, if combining the full questionnaire data from 2015-2019, _LLCPWT should be divided by 5 or multiplied by 1/5. An example of how to do this in SAS 9.4 is provided in Box 9.1.

Box 9.1. SAS program to create new weight variable for five BRFSS data collection periods



```
data new_weight; set old_weight;
brfssweight5=_LLCPWT/5;
run;
```

9.6 BRFSS CASE STUDY: CAREGIVING EXPERIENCES BY METRO AND NON-METRO GEOGRAPHIC CONTEXT AND RACE/ETHNICITY

In 2020, the National Alliance for Caregiving (NAC) and American Association of Retired Persons (AARP) estimated that 21% of adults in the US are informal caregivers, which has increased by 9.5 million since 2015.[9] Over 11 million unpaid individuals, family or friends, are caregivers for persons living with Alzheimer's disease and related dementias (ADRD). The prevalence of ADRD is highest among non-Hispanic Whites; however, the prevalence is increasing among racial and ethnically diverse older adults. Few studies have evaluated differences in ADRD caregiving by geographic context. Studies which have compared ADRD caregiving in metro and non-metro areas have highlighted unmet resource needs and support the lack of dementia-specific[10] and respite services.[11] In a recent study using data from National Study of Caregiving (NSOC), Kindratt and colleagues found that non-metro ADRD caregivers were less racially/ethnically diverse (82.7% White), and more were spouses/partners (20.2%).[12] Among racial/ethnic minority ADRD caregivers, non-metro context was associated with having more chronic conditions (p<.01), providing less care (p<.01), and not co-residing with care recipients (p<.001). Among White ADRD caregivers, non-metro context was associated with not reporting caregiving was more than they could handle (p<.05) and finding financial assistance for caregiving (p<.05). Non-metro minority ADRD caregivers had 3.09 times higher odds (95% CI=1.02-9.36) of reporting anxiety in comparison to metro minority ADRD caregivers. The BRFSS case study will extend this previous research by using BRFSS data from Texas to examine differences in ADRD caregiving experiences. Caregiving experiences that will be evaluated include managing personal care (e.g. medications, feeding, dressing, bathing) and the household (e.g. cleaning, managing money, preparing meals).

9.6.1 SPECIFIC AIMS

- Aim 9.1. Determine whether ADRD caregiving experiences differ across metro and non-metro geographic contexts among adults in Texas
- Aim 9.2. Determine whether the relationship between geographic context and ADRD caregiving experiences is moderated by the caregiver's race/ethnicity among metro and non-metro adults in Texas

9.6.2 METHODS

Complete the following steps to download, clean, recode and analyze 2019 BRFSS data to determine how associations between metro and non-metro geographic context and ADRD caregiver experiences are moderated by race and ethnicity. The full SAS programs are available for download in the chapter 9 folder in the Open ICPSR data repository.

Step 1: Download BRFSS 2019 data and SAS programming files

- Go to the 2019 BRFSS data website
- Click on “2019 BRFSS Data (ASCII)” under data files. This file contains the combined landline and cell phone data
- A zip file with a “.dat” data file will be downloaded to the “Downloads” folder on your computer
- Unzip the file and save the data file to your computer. I recommend creating a folder on the “C Drive” labeled BRFSS and separated by each year. This will be consistent with the location statements used in the textbook examples
- Under SAS resources, there are three syntax files that are useful for creating the dataset, formatting and labeling the variables in the data file
    - Click on “SASOUT19_LLCP.SAS” for the programming statements used to convert the “.dat” data file into a SAS data file and save to your computer. SAS statements from this file will be run first during Step 2.
    - Click on and save “Formats19 [SAS7BCAT Direct Download – 493 KB]” for programming statements used to generate the 2019 format library. Statements are available for 32-bit and 64-bit SAS. SAS statements from this file should be run second during Step 2.

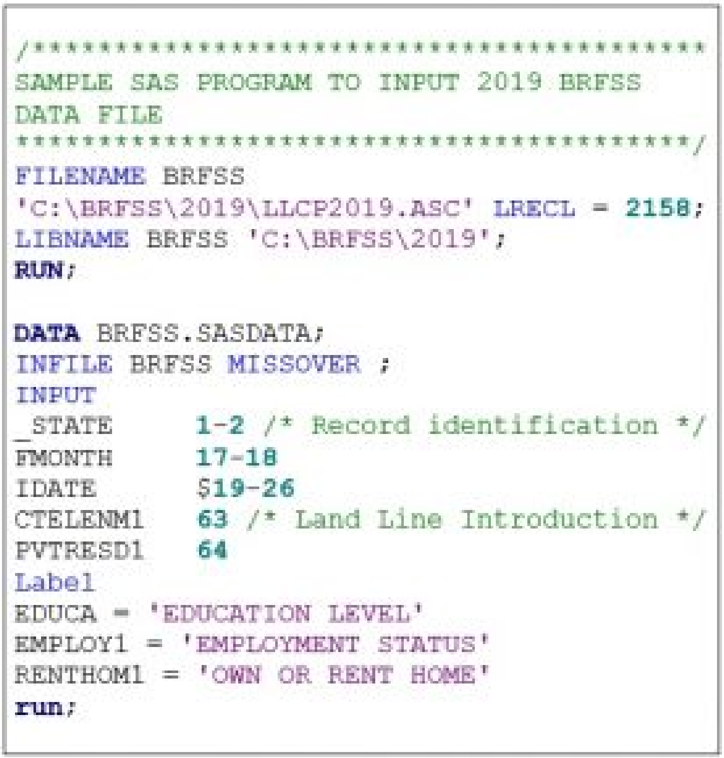
- ◦ Click on and save “Formas19.sas CDC” for format assignment statements. SAS statements from this file will be run third during Step 2.

Step 2: Run SAS programming statements to input data, create library, formats and labels for 2019 BRFSS data

Sample SAS programming statements to create the libraries and input the 2019 BRFSS data files are provided in Box 9.2. To create these programming statements, complete the following steps:

- Open “SASOUT19_LLCP.SAS” Programming File
- Create a LIBNAME statement which houses the data and files associated with the analysis. I recommend creating the LIBNAME statement as the survey name (e.g. “BRFSS”) and using the same location that the data files for saved in on the C drive (e.g. “C:\BRFSS\2019”)
- Create a FILENAME statement which lets SAS know where the data file is stored (e.g. ‘C:\BRFSS\2019\LLCP2019.ASC’)
- Modify or remove any instructions (/*green text*/) that you do not need in programming file
- Add filename to first DATA procedure (e.g. ‘data brfss.sasdata’)
- Add libname to the INFILE procedure (e.g. ‘brfss’)
- Highlight all programming statements and click RUN

Box 9.2. SAS Program to input 2019 BRFSS data file

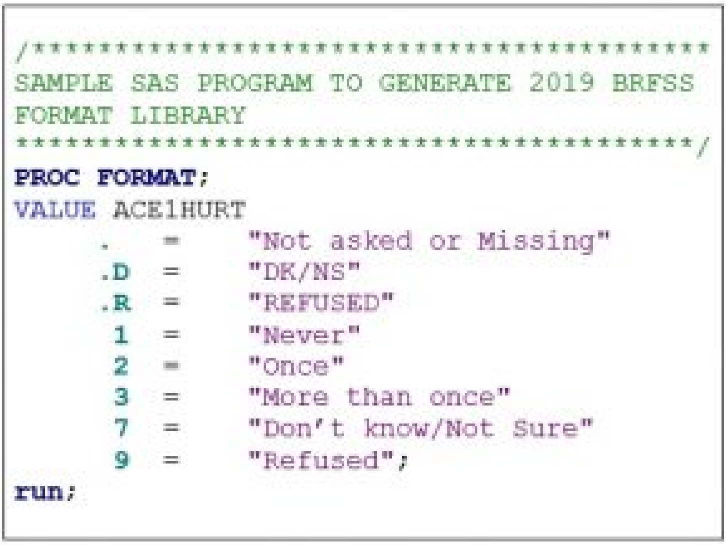


```
/****************************************
SAMPLE SAS PROGRAM TO INPUT 2019 BRFSS
DATA FILE
****************************************/
FILENAME BRFSS
'C:\BRFSS\2019\LLCP2019.ASC' LRECL = 2158;
LIBNAME BRFSS 'C:\BRFSS\2019';
RUN;

DATA BRFSS.SASDATA;
INFILE BRFSS MISSOVER ;
INPUT
_STATE     1-2 /* Record identification */
FMONTH     17-18
IDATE      $19-26
CTELENM1   63 /* Land Line Introduction */
PVTRESD1   64
Label
EDUCA = 'EDUCATION LEVEL'
EMPLOY1 = 'EMPLOYMENT STATUS'
RENTHOM1 = 'OWN OR RENT HOME'
run;
```

Sample SAS programming statements to generate the 2019 format library are provided in Box 9.3. To create these programming statements, complete the following steps:

- Open the “Formats19 [SAS7BCAT Direct Download – 493 KB]” Programming File
- Highlight all programming statements and click RUN

Box 6.3. SAS Program to generate BRFSS 2019 format library

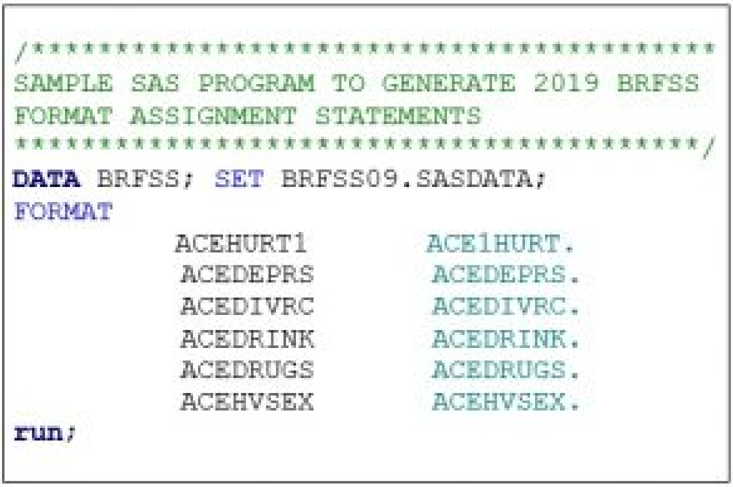


```
/****************************************
SAMPLE SAS PROGRAM TO GENERATE 2019 BRFSS
FORMAT LIBRARY
****************************************/
PROC FORMAT;
VALUE ACE1HURT
      .   =    "Not asked or Missing"
      .D  =    "DK/NS"
      .R  =    "REFUSED"
       1  =    "Never"
       2  =    "Once"
       3  =    "More than once"
       7  =    "Don't know/Not Sure"
       9  =    "Refused";
run;
```

Sample SAS programming statements to generate the 2019 BRFSS format assignment statements are provided in Box 9.4. To create these programming statements, complete the following steps:

- Open the “Formas19.sas CDC” Programming File
- Add “data” procedure statement at the top of the file
- Add “run” statement at the bottom of the file
- Highlight all programming statements and click RUN

Box 9.4. SAS Program to generate 2019 BRFSS format assignment statements

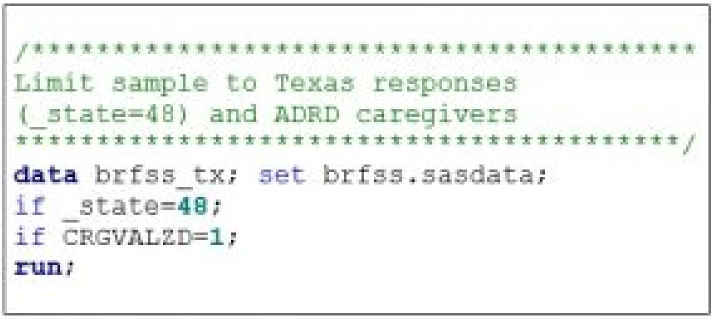


```
/*****************************************
SAMPLE SAS PROGRAM TO GENERATE 2019 BRFSS
FORMAT ASSIGNMENT STATEMENTS
*****************************************/
DATA BRFSS; SET BRFSS09.SASDATA;
FORMAT
          ACEHURT1       ACE1HURT.
          ACEDEPRS       ACEDEPRS.
          ACEDIVRC       ACEDIVRC.
          ACEDRINK       ACEDRINK.
          ACEDRUGS       ACEDRUGS.
          ACEHVSEX       ACEHVSEX.
run;
```

Step 3: Limit the dataset to respondents from Texas

Due to the large sample size and state-based probability sampling frame, analysts can be confident that their power will be sufficient for producing statewide estimates. In this case study, our sample is limited to respondents from Texas. The sample is also limited to ADRD caregivers. To produce statistical estimates for the state of Texas only, our data must be limited to response option "48" for variable "_state." Sample SAS programming statements for limiting the dataset by state are provided in Box 9.5.

Box 9.5. SAS Program to limit 2019 BRFSS data to only Texas respondents and ADRD caregivers



```
/****************************************
Limit sample to Texas responses
(_state=48) and ADRD caregivers
****************************************/
data brfss_tx; set brfss.sasdata;
if _state=48;
if CRGVALZD=1;
run;
```

Step 4: Select variables for analysis

Once formats been assigned to the dataset, you can remove any variables that are not needed for your analysis. This will reduce the size of the dataset and make processing time quicker when running SAS programming statements. In this case study, I have kept the following variables (Table 9.2) to denote the survey design features and creation of the independent variable, dependent variable, moderator, and selected covariates.

Table 9.2 Overview of variables used for BRFSS case study

Variable Name	Variable Description
Design Variables	
_PSU	Primary Sampling Unit
_STSTR	Sample Design Stratification
_LLCPWT	Weight: Land-line and cell
Inclusion Criteria	
_STATE	State code
CAREGIV1	Provided regular care: family/friend
CRGVALZD	Care recipient has Alzheimer's disease, dementia, or other cognitive impairment
Independent Variables	
_METSTAT	Metropolitan status
Dependent Variables	
CRGVHRS1	Hours provide care
CRGVPER1	Managed personal care
CRGVHOU1	Managed household tasks
Moderator	
IMPRACE	Imputed race/ethnicity value
Covariates	
CRGVREL3	Relationship to care recipient
_AGE_G	Imputed age in six groups
SEXVAR	Sex of respondent
EMPLOY1	Employment status

Step 5: Recode and rename variables

Questionnaire responses often need to be recoded or responses collapsed prior to conducting statistical analysis. For example, the BRFSS has response options “7=Don’t Know” and “9=Refused” for several questions. The responses are often removed and made “missing” prior to analysis. Furthermore, the numbers that represent certain values may need to be changed for easier interpretation of statistical analysis results. For example, BRFSS has response options “1=Yes” and “2=No.” It is common practice to change “no” responses to 0, “0=No.” It is best practice to rename these recoded variables with a new variable name instead of replacing the original variable.

An overview of the variables recoded and renamed for analysis in this case study is provided in Table 9.3.

Table 9.3. Overview of BRFSS variables recoded and renamed to meet research aims

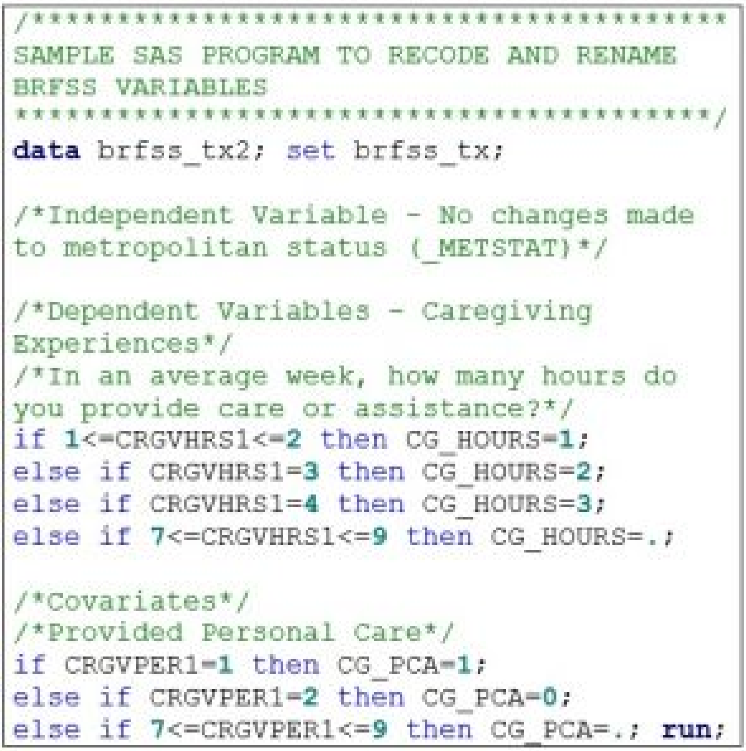
Question Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
Care recipient has ADRD	CRGVALZD	1=Yes 2=No 7=Don't know 9=Refused	CR_ADRD	0=No 1=Yes
Hours provided care in average week	CRGVHRS1	1=Up to 8 hours 2=9-19 hours 3=20-39 hours 4=40 hours or more 7=Don't know 9=Refused	CG_HOURS	1=0-19 hours 2=20-39 hours 3=40 hours or more
Managed personal care (medications, feeding, dressing, bathing)	CRGVPER1	1=Yes; 2=No 7=Don't know 9=Refused	CG_PCA	0=No 1=Yes
Managed household (cleaning, managing money, preparing meals)	CRGVHOU1	1=Yes; 2=No 7=Don't know 9=Refused	CG_HOUSE	0=No 1=Yes

Table 9.3 (continued). Overview of BRFSS variables recoded and renamed to meet research aims

Question Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
Relation of care recipient to caregiver	CRGVREL3	1=Mother 2=Father 3=Mother-in-law 4=Father-in-law 5=Child 6=Husband 7=Wife 8=Live in partner 9=Brother or brother-in-law 10=Sister or sister-in-law 11=Grandmother 12=Grandfather 13=Grandchild 14=Other relative 15=Non-relative	CG_RELATION	0=Other 1=Spouse or partner
Caregiver Sex	SEXVAR	1=Male 2=Female	CG_SEX (rename only)	1=Male 2=Female
Employment	EMPLOY1	1=Employed for wages 2=Self-employed 3=Out of work 1+ year 4=Out of work <1 year 5=Homemaker 6=Student 7=Retired 8=Unable to work 9=Refused	CG_WORK	1=Employed 2=Retired 3=Not working

A sample SAS program for recoding and renaming 2019 BRFSS data for this case study is provided in Box 9.6.

Box 9.6. Sample SAS program for recoding and renaming BRFSS variables



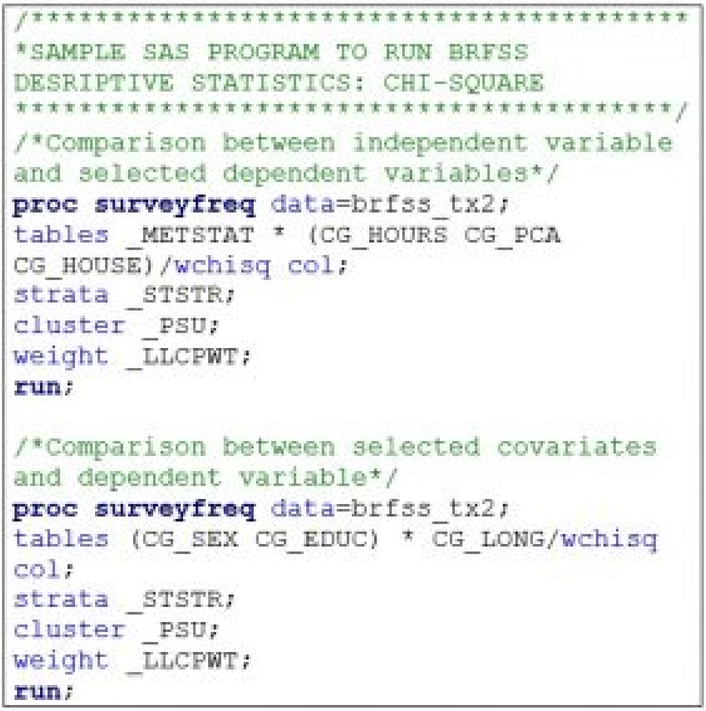
```
/******************************************
SAMPLE SAS PROGRAM TO RECODE AND RENAME
BRFSS VARIABLES
******************************************/
data brfss_tx2; set brfss_tx;

/*Independent Variable - No changes made
to metropolitan status (_METSTAT)*/

/*Dependent Variables - Caregiving
Experiences*/
/*In an average week, how many hours do
you provide care or assistance?*/
if 1<=CRGVHRS1<=2 then CG_HOURS=1;
else if CRGVHRS1=3 then CG_HOURS=2;
else if CRGVHRS1=4 then CG_HOURS=3;
else if 7<=CRGVHRS1<=9 then CG_HOURS=.;

/*Covariates*/
/*Provided Personal Care*/
if CRGVPER1=1 then CG_PCA=1;
else if CRGVPER1=2 then CG_PCA=0;
else if 7<=CRGVPER1<=9 then CG_PCA=.; run;
```

Step 6: Conduct descriptive statistical analysis

Once all variables are recoded, collapsed, and renamed they can be used for statistical analysis. Statistical analysis should always start with descriptive analysis to describe the data source. Chi square analyses should be conducted to make categorical comparisons between the independent variable, covariates, and dependent variables. It is important to remember that all analysis of BRFSS data needs to be conducted with SAS Survey procedures due to the complex sample design. Weighting (variable: _LLCPWT), primary sampling unit (variable: _PSU) and cluster (variable: _STSTR) variables must be included in the programming statements.

A sample SAS program for conducting chi-square tests using 2019 BRFSS data for this case study is provided in Box 9.7.

Box 9.7. Sample SAS program for running descriptive statistics (chi-square)



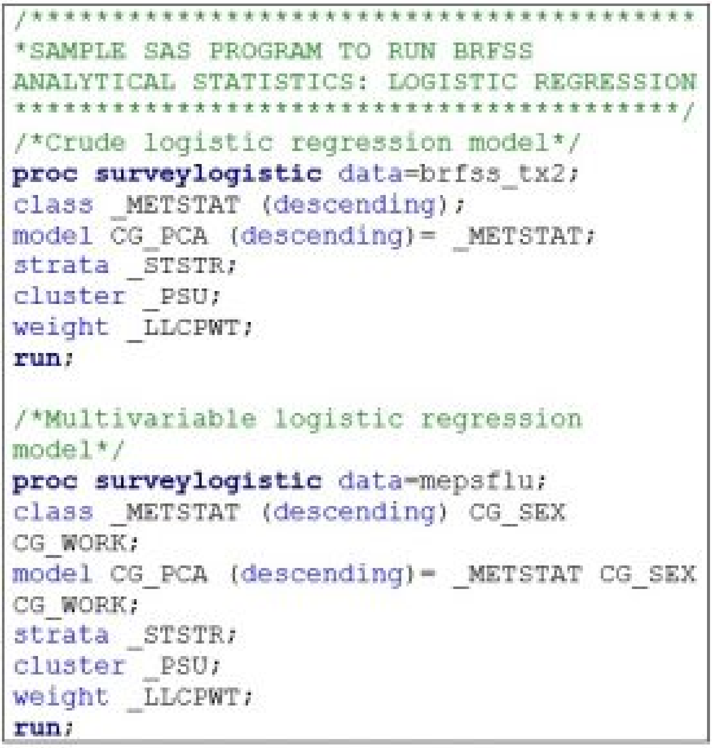
```
/****************************************
*SAMPLE SAS PROGRAM TO RUN BRFSS
DESRIPTIVE STATISTICS: CHI-SQUARE
****************************************/
/*Comparison between independent variable
and selected dependent variables*/
proc surveyfreq data=brfss_tx2;
tables _METSTAT * (CG_HOURS CG_PCA
CG_HOUSE)/wchisq col;
strata _STSTR;
cluster _PSU;
weight _LLCPWT;
run;

/*Comparison between selected covariates
and dependent variable*/
proc surveyfreq data=brfss_tx2;
tables (CG_SEX CG_EDUC) * CG_LONG/wchisq
col;
strata _STSTR;
cluster _PSU;
weight _LLCPWT;
run;
```

Step 7: Conduct inferential statistical analysis

After calculating descriptive statistics, inferential statistical analysis can be conducted. Crude and multivariable logistic regression models can be calculated to determine associations between metropolitan or non-metropolitan residential context and caregiver experiences. Crude logistic regression models are used to determine the association between the independent and dependent variables without adjusting for other factors. Multivariable logistic regression models are used to determine associations between the independent and dependent variables after adjusting for potential covariates (e.g. sex, employment status). A reference category for the independent variables are needed. For this analysis, the reference group is "Metropolitan" residence. Results compare caregiving experiences among caregivers who live in metropolitan (urban) and non-metropolitan (rural) geographic contexts.

A sample SAS program for conducting logistic regression analysis using 2019 BRFSS data for this case study is provided in Box 9.8.

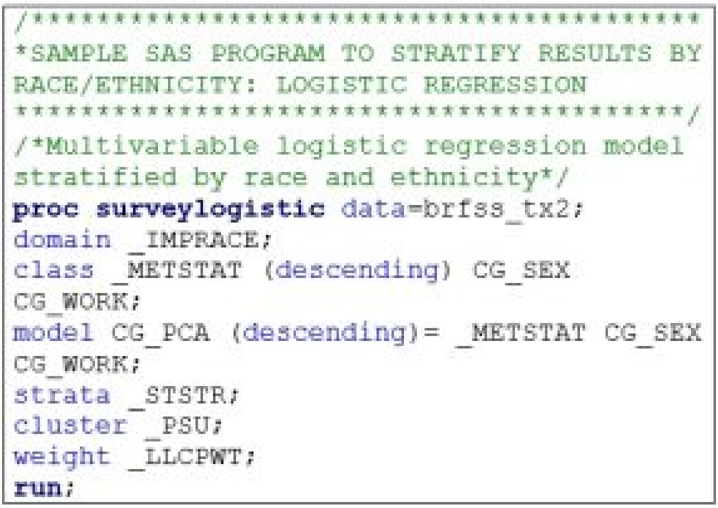
Box 9.8. Sample SAS program for running BRFSS inferential statistics (logistic regression)



```
/****************************************
*SAMPLE SAS PROGRAM TO RUN BRFSS
ANALYTICAL STATISTICS: LOGISTIC REGRESSION
****************************************/
/*Crude logistic regression model*/
proc surveylogistic data=brfss_tx2;
class _METSTAT (descending);
model CG_PCA (descending)= _METSTAT;
strata _STSTR;
cluster _PSU;
weight _LLCPWT;
run;

/*Multivariable logistic regression
model*/
proc surveylogistic data=mepsflu;
class _METSTAT (descending) CG_SEX
CG_WORK;
model CG_PCA (descending)= _METSTAT CG_SEX
CG_WORK;
strata _STSTR;
cluster _PSU;
weight _LLCPWT;
run;
```

In order to determine whether race/ethnicity is a moderator in the relation between geographic context and ADRD caregiver experiences, a "DOMAIN" statement in SAS must be used to present stratified results. Any example of where to include this statement is provided in Box 9.9. The _IMPRACE variable may need to be recoded to "White/Non-Hispanic" and "Other/Minority" due to small sample sizes of non-metro ADRD minority caregivers separated by specific races and ethnicities.

Box 9.9. Sample SAS program for running BRFSS inferential statistics (logistic regression) using DOMAIN statement to stratify results



```
/***************************************
*SAMPLE SAS PROGRAM TO STRATIFY RESULTS BY
RACE/ETHNICITY: LOGISTIC REGRESSION
***************************************/
/*Multivariable logistic regression model
stratified by race and ethnicity*/
proc surveylogistic data=brfss_tx2;
domain _IMPRACE;
class _METSTAT (descending) CG_SEX
CG_WORK;
model CG_PCA (descending)= _METSTAT CG_SEX
CG_WORK;
strata _STSTR;
cluster _PSU;
weight _LLCPWT;
run;
```

9.7 SUMMARY

This chapter provided an overview of the BRFSS and ways to conduct basic statistical analysis for one state using 2019 BRFSS public-use data files. The BRFSS case study explored whether geographic context was associated with caregiving experiences. Sample SAS programming statements were provided for downloading and inputting data files, merging data files, recoding and renaming variables, and conducting categorical descriptive and inferential statistical analysis. The dataset and full SAS programming statements for the BRFSS case study are available in the chapter 9 folder in the Open ICPSR data repository.

9.8 REFERENCES

1. Centers for Disease Control and Prevention. About the Behavioral Risk Factor Surveillance System (BRFSS). Published February 9, 2019. Accessed December 10, 2020. https://www.cdc.gov/brfss/about/about_brfss.htm
2. CDC – About BRFSS. Published February 9, 2019. Accessed December 10, 2020. https://www.cdc.gov/brfss/about/index.htm
3. Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System: 2019 Summary Data Quality Report.*; 2020. Accessed June 19, 2022. https://www.cdc.gov/brfss/annual_data/2019/pdf/2019-sdqr-508.pdf
4. CDC – BRFSS. Published August 31, 2020. Accessed June 19, 2021. https://www.cdc.gov/brfss/index.html
5. Centers for Disease Control and Prevention (CDC). *Behavioral Risk Factor Surveillance System Overview: BRFSS 2019.* Accessed June 17, 2021. https://www.cdc.gov/brfss/annual_data/2019/pdf/overview-2019-508.pdf
6. Centers for Disease Control and Prevention. *Behavioral Risk Factor Surveillance System (BRFSS) Complex Sample Weights and Preparing 2019 BRFSS Module Data for Analysis*. https://www.cdc.gov/brfss/annual_data/2019/pdf/Complex-Smple-Weights-Prep-Module-Data-Analysis-2019-508.pdf

7. Texas Department of State Health Services. *Texas Behavioral Risk Factor Surveillance System Questionnaire 2020.*; 2020. https://www.dshs.texas.gov/chs/brfss/attachments/2020_BRFSS_Survey.pdf

8. Data Sources – Healthy People 2030 | health.gov. Accessed June 21, 2021. https://health.gov/healthypeople/objectives-and-data/data-sources-and-methods/data-sources

9. Jr SM. Caregiving in the US 2020 | The National Alliance for Caregiving. Published May 11, 2020. Accessed June 25, 2021. https://www.caregiving.org/caregiving-in-the-us-2020/

10. Gibson A, Holmes SD, Fields NL, Richardson VE. Providing Care for Persons with Dementia in Rural Communities: Informal Caregivers' Perceptions of Supports and Services. *J Gerontol Soc Work*. 2019;62(6):630-648. doi:10.1080/01634372.2019.1636332

11. Kosloski K, Schaefer JP, Allwardt D, Montgomery RJV, Karner TX. The role of cultural factors on clients' attitudes toward caregiving, perceptions of service delivery, and service utilization. *Home Health Care Serv Q*. 2002;21(3-4):65-88. doi:10.1300/J027v21n03_04

12. Kindratt T, Sylvers D, Yoshikawa A, Anuarbe ML, Webster N, Bouldin E. ADRD Caregiving Experiences and Health by Race, Ethnicity and Care Recipient Geographic Context. *Innov Aging*. 2021;5(Suppl 1):990. doi:10.1093/geroni/

igab046.3557

CHAPTER 10.

NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY

10.1 INTRODUCTION

Chapter 10 covers the National Health and Nutrition Examination Survey (NHANES). The NHANES has been collected by the National Center for Health Statistics (NCHS) since 1960 to monitor and explore trends in the health status and nutritional status among all individuals in the United States (US).[1] It became formally known as the NHANES in 1999. A unique aspect of the NHANES in comparison to other surveys is that the NHANES collects data from both subjective interviews and objective physical examinations and laboratory tests. Objective measures are collected at mobile examination centers.[1] This chapter includes details on: how data are collected; how data are made publicly available as machine-actionable data files; what variables must be included to address design features of the complex sample; the strengths and limitations of the survey; and practical tips for conducting statistical analysis; and how to answer research questions using a case study. The practical tips provided for analysis of NHANES data are based on the author's previous experiences analyzing NHANES data to answer questions related to associations between predisposing and enabling factors that contribute to health behaviors, morbidity, mortality and health services use. The NHANES case study will explore racial and ethnic differences in 24-hour movement guideline adherence, specifically sedentary behavior guideline adherence. This objective is part of a series of research questions designed to evaluate how physical activity, sleep, and sedentary behavior are associated with cognitive health outcomes among adults in the US. The bulk of the chapter will comprise of section *10.6: NHANES Case Study* in order to give the readers hands-on practice downloading and cleaning large databases and conducting basic categorical data analysis using PROC SURVEYFREQ and PROC SURVEYLOGISTIC. The syntax provided was created for use with SAS 9.4 .

10.2 DATA COLLECTION

The NHANES uses a cross-sectional study design to collect data from personal interviews, physical examinations and laboratory data among noninstitutionalized adults and children. Data are compiled and released in 2-year cycles as public-use data files.[2] From 2011-2014, the complex sample design included 13 major strata, 4 minor strata and 8 primary sampling units (PSU).[2] The study design oversampled Hispanics, non-Hispanic Blacks and non-Hispanic Asians, persons below 130% of federal poverty level and persons ages 80 years and older. Data were collected from five US regions, with California separated as a distinct group. Starting in 2015, the design changed to 14 major strata, 4 minor strata and 4 PSUs.[3] Individuals were oversampled below 185% federal poverty level instead of 130% and California was not separated from all other states in the new design.

10.3 DATA FILES

The most recent NHANES iterations (2017-2020) included demographic, dietary, medical examination, laboratory, questionnaire, and limited access data.[4] Data were collected in participants' households as well as mobile examination centers.

10.3.1 Demographic Data

The demographic data file includes individual details on the participants' gender, age, marital status, language preference, race, and ethnicity.[4] Questions on place of birth are included and participants who report that they do not live in the US are asked about their citizenship status and how long they have lived in the US. While the NHANES collected data on countries of birth outside of the US, details on the countries of birth are not available to the public. The demographic data file also includes questions related to socioeconomic status, including the highest level of education, income, and questions about military service. Pregnancy status is reported among women ages 20 to 44 years. Primary sampling unit, cluster, and weighting variables are located with the demographic data for the 2017-2020 pre-pandemic data files.[4,5]

10.3.2 Dietary Data

The NHANES dietary interviews are conducted to obtain dietary data on food and beverage intake 24-hours prior to the first interview.[6] The dietary interviews are conducted in-person at mobile examination centers with follow-up phone calls. For participants less than six years old, dietary data are collected from a proxy adult. For children 6-8 and 9-11 years old, interviews are conducted with the child and a proxy adult. Child participants ages 12 and older completed interviews themselves. In-person interviews include several household items to be used for measuring food intake amounts. Participants were provided with these items to take home for the telephone interviews collected 3 to 10 days later.[6] There are several data files that include variables from the dietary interviews. NHANES provides data files for first- and second-day individual foods and nutrients. There are also data files that include information on 24-hour and 30-day dietary supplements.[6] The individual food data files include comprehensive responses on the types of food combinations eaten (e.g. cereal, soup, salad, tortilla products), its source (e.g. store – grocery/supermarket, restaurant with waiter/waitress, in K-12 school, or childcare center), and nutrients (e.g. total folic acid [mcg], potassium [mg], energy [kcal]).[6]

10.3.3. Examination Data

The NHANES examination files comprise of data from multiple procedures to measure the health of participants.[7] All examinations are conducted in the mobile examination centers. Examples of the examinations conducted include measurements of audiometry, anthropometry, body measures, balance, blood pressure, cardiovascular fitness, dermatology, muscle strength and oral health among others.[7]

10.3.4 Laboratory Data

The NHANES laboratory files comprise of multiple tests conducted on the biological specimens of participants.[8] All laboratory data are collected in the mobile examination centers and sent to a laboratory for testing. Laboratory tests are conducted using blood, urine and other biospecimens (e.g., hair, nasal swab, plasma).[8] Examples of laboratory tests conducted include cholesterol, folate, glycohemoglobin, insulin, plasma fasting glucose, and mercury, among others.[8]

10.3.5 Questionnaire Data

The NHANES laboratory files include self-reported data on questions regarding the health and wellness of participants.[9] Questionnaire data are collected in the household, mobile examination centers, and by telephone. Examples of questionnaire data collected range from socioeconomic status (occupation, income), acculturation, weight history, preventive health behaviors (e.g., immunizations, smoking, physical activity), health conditions (e.g., diabetes, kidney conditions, osteoporosis), food security and health care access and utilization.[9]

10.3.6 Limited Access Data

The NHANES limited access data files measure sensitive topics among youth and adults.[10] Data from questionnaires as well as biospecimens are collected with special precautions put in place so ensure confidentiality. Examples of limited access data include biospecimens collected to measure lead in blood and sexually transmitted diseases (e.g., chlamydia, herpes simplex viruses, HIV antibodies) and questionnaire data on drug use, alcohol use, and sexual behaviors among youth and adults.[10] Limited access data files are not available for public use. Researchers must apply for access through the National Center for Health Statistics' Research Data Centers.[10]

10.3.7 Linked Data

The NHANES can be linked to National Death Index (NDI) and Medicare data.[11] Efforts are underway to link NHANES data with Housing and Urban Development (HUD) and Medicaid data.[11] The purpose of linking NHANES and NDI data is to examine how multiple risk factors related to health and nutrition are associated with mortality. Linked NHANES/NDI data can be accessed as public-use data files or through a restricted data application. Public-use linked data files are only provided for adults and do not provide specific dates for birth, interviews and death, or specific causes of death beyond standard categories. Linkages with Medicare data allow for research focused on health status, health care costs, health care utilization, and prescription drug use among Medicare enrollees.[11] To access more specific details of the linked data, researchers must apply for access through the National Center for Health Statistics' Research Data Centers.[11]

10.3.8 Restricted Data

The NHANES restricts data on geography (Census 2010 Block ID), genetics (e.g. BRCA1 associated protein), and the exact dates of participants' interviews and examinations. To used these data, researchers must apply for access through the National Center for Health Statistics' Research Data Centers.[12]

10.4. STRENGTHS AND LIMITATIONS

There are several advantages to using the NHANES for research. First, the ability to validate self-reported and objective measurements through personal interviews, and physical examinations is a strength. A second strength is the ability to determine undiagnosed diseases with laboratory values, such as diabetes mellitus. A third strength is the ability to use acculturation variables collected among Hispanic and non-Hispanic Asian participants. However, there are some limitations of using NHANES data. First, there is a smaller annual sample size compared to other national surveys such as the National Health Interview Survey (NHIS). Second, the place of birth questions are limited to US- or foreign-born only and no data is collected on country of birth, which limits the ability for data disaggregation among foreign-born groups. Third, the large number of subsection files requires multiple merges of data files for each survey year.

10.5 DESIGN FEATURES

Data analysts must use special procedures to account for the complex sample design used by the NHANES. Survey procedures much include variables to adjust for the primary sampling units, stratification, and weighting of each cycle of continuous data files. Changes were made to the complex design variables for the 2017-2020 pre-pandemic data files because there were two cycles (2017-2018 and 2019-2020) and the 2019-2020 data file was incomplete.[5] For 2017-2020, the complex sample design variables are available in the "Demographic Variables and Sample Weight" file. Researchers must decide which weight to use based on the aims of their study using NHANES data. For research studies that only use data from the interviews, the interview weight is most appropriate. For research studies that use outcomes from the examination or laboratory data files, the mobile examination center weight is most appropriate. An overview of the primary sampling unit, stratum, and weighting variables for the 2017-2020 pre-pandemic data files are provided in Table 10.1.

Table 10.1. Overview of complex sample design variables for NHANES 2017-2020 pre-pandemic cycle

Primary Sampling Unit	Stratum	Weights	
		Interview	Mobile Examination Center
SDMVPSU	SDMVSTRA	WTINTPRP	WTMECPRP

10.6 NHANES CASE STUDY

In 2018, the Canadian Society for Exercise Physiology convened to develop the Canadian 24-Hour Movement Guidelines for adults ages 18-64 years and ages 65 and older. The guidelines integrate recommendations for sleep, physical activity, and sedentary behavior with the rationale that the combination of these behaviors throughout the day is associated with improved health outcomes.[13] Among adults ages 18-64 years old, it is recommended that individuals get 7 to 9 hours of good-quality sleep on a regular basis, with consistent bed and wake-up times. Adults are recommended to perform a variety of intensities and types of physical activity, including: 1) moderate to vigorous aerobic physical activities such that there is an accumulation of at least 150 minutes per week; 2) muscle strengthening activities using major muscle groups at least twice a week; and 3) several hours of light physical activities, including standing.[13] It is recommended that adults limit sedentary behavior to 8 hours or less (~480 minutes), including no more than 3 hours of recreational screen time and breaking up long periods of sitting as often as possible.[13] Recommendations differ slightly among older adults. Among adults ages 65 and older, it is recommended that individuals get 7 to 8 hours of good-quality sleep on a regular basis, with consistent bed and wake-up times. In addition to the physical activity recommendations for adults ages 18-65 years, older adults are recommended to engage in physical activities that challenge balance.[13] There are no differences in recommendations for sedentary behavior among adults ages 18-64 years or 65 years and older. While exercise physiology and health organizations in the US set goals and standards for each of these health behaviors, there is no effort for integration of these movements to examine health disparities. In this case study, we will determine racial and ethnic differences in sedentary behavior among US- and foreign-born Hispanics, non-Hispanic Whites, non-

Hispanic Blacks, and non-Hispanic Asians. Future research will incorporate physical activity and sleep behaviors.

10.6.1 SPECIFIC AIMS

- Aim 10.1: Compare the prevalence of adherence to 24-hour sedentary behavior guidelines in US adults by race, ethnicity, and nativity status
- Aim 10.2: Determine associations between race, ethnicity, and nativity and sedentary guideline adherence among racially and ethnically diverse foreign-born adults compared to their US-born counterparts

10.6.2 METHODS

Complete the following steps to download, clean, recode and analyze NHANES data to answer the specific aims.

Step 1: Download demographics, questionnaire, and examination datasets

Follow the steps below to download and store the necessary data files for the NHANES case study and import them into SAS 9.4.

- Create a folder in a permanent location to save your data files. I recommend creating a folder on the "C Drive" titled NHANES with a subfolder to identify the years (2017-2020). This will align with the examples in this textbook.
- Go to the NHANES 2017- March 2020 pre-pandemic data website

    - Under “Data, Documentation, Codebooks,” click “Demographics Data.”
    - Click on “P_DEMO Data [XPT – 3.4 MB]” under “Data File.” This file should automatically be downloaded and show up at the bottom of the browser or your “Downloads” folder. The file is a “SAS Xport Transport File” type. This data file can be opened in SAS and saved as a standard SAS file.
- Go back to the NHANES 2017- March 2020 pre-pandemic data website
    - Under “Data, Documentation, Codebooks,” click “Questionnaire Data.”
    - In the row for Physical Activity, click on “P_PAQ Data [XPT – 1.3 MB]” under “Data File.” This file should automatically be downloaded and show up at the bottom of the browser or your “Downloads” folder. The file is a “SAS Xport Transport File” type. This data file can be opened in SAS and saved as a standard SAS file
- Go back to the NHANES 2017- March 2020 pre-pandemic data website
    - Under “Data, Documentation, Codebooks,” click “Examination Data.”
    - In the row for Body Measures, click on “P_BMX Data [XPT – 2.4 MB]” under “Data File.” This file should automatically be downloaded and show up at the bottom

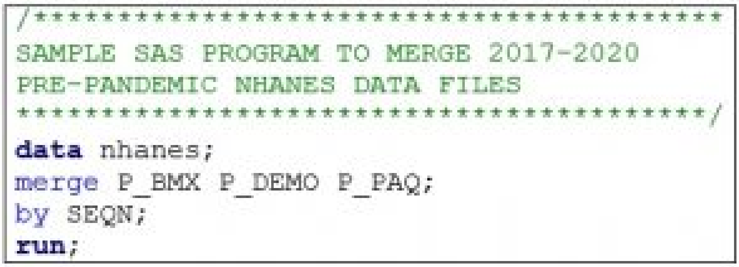
of the browser or your “Downloads” folder. The file is a “SAS Xport Transport File” type. This data file can be opened in SAS and saved as a standard SAS file.

Step 2: Open SAS transport files in SAS and save to permanent datasets for merge

Double click to open each of the SAS Xport transport files downloaded and saved on the C Drive in SAS.

- There should be three temporary work files created once these files are opened. The file names should be:
    - P_BMX
    - P_DEMO
    - P_PAQ
- Open a new SAS syntax editor file by clicking “New” where the blank white page is located.
- Enter the syntax to merge the data files provided in Box 10.1. The variable used to identify each participant is SEQN.

Box 10.1. Sample SAS program to merge 2017-2020 Pre-Pandemic NHANES data files



```
/***************************************
SAMPLE SAS PROGRAM TO MERGE 2017-2020
PRE-PANDEMIC NHANES DATA FILES
***************************************/
data nhanes;
merge P_BMX P_DEMO P_PAQ;
by SEQN;
run;
```

Step 3: Keep only the variables that you need

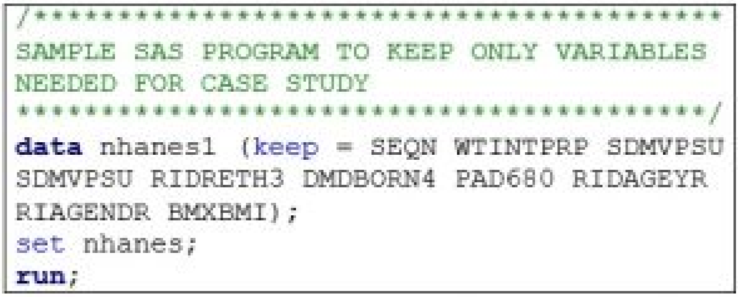
Once the data files are merged, it is recommended to keep only the variables needed for the analysis. Removing additional variables will allow the SAS program to run and present results faster. In this case study, I have kept the following variables (Table 10.2) to denote the survey design features and creation of the independent variables, dependent variables, and selected covariates.

Table 10.2. Overview of variables used for NHANES case study

Original File	Variable Name	Variable Description
	Design Variables	
All Files	SEQN	Respondent sequence number
P_DEMO	WTINTPRP	Full sample interview weight
P_DEMO	SDMVPSU	Masked variance pseudo-PSU
P_DEMO	SDMVSTRA	Masked variance pseudo-stratum
	Independent Variables	
P_DEMO	RIDRETH3	Race/Hispanic origin w/ NH Asian
P_DEMO	DMDBORN4	Country of birth
	Dependent Variable	
P_PAQ	PAD680	Minutes sedentary activity
	Covariates	
P_DEMO	RIDAGEYR	Age in years at screening
P_DEMO	RIAGENDR	Gender
P_BMX	BMXBMI	Body Mass Index (kg/ m^2)

A sample SAS program with a keep statement that includes only the variables needed is provided in Box 10.2.

Box 10.2. Sample SAS program to keep only variables needed for case study



```
/****************************************
SAMPLE SAS PROGRAM TO KEEP ONLY VARIABLES
NEEDED FOR CASE STUDY
****************************************/
data nhanes1 (keep = SEQN WTINTPRP SDMVPSU
SDMVPSU RIDRETH3 DMDBORN4 PAD680 RIDAGEYR
RIAGENDR BMXBMI);
set nhanes;
run;
```

Step 4: Recode and rename variables

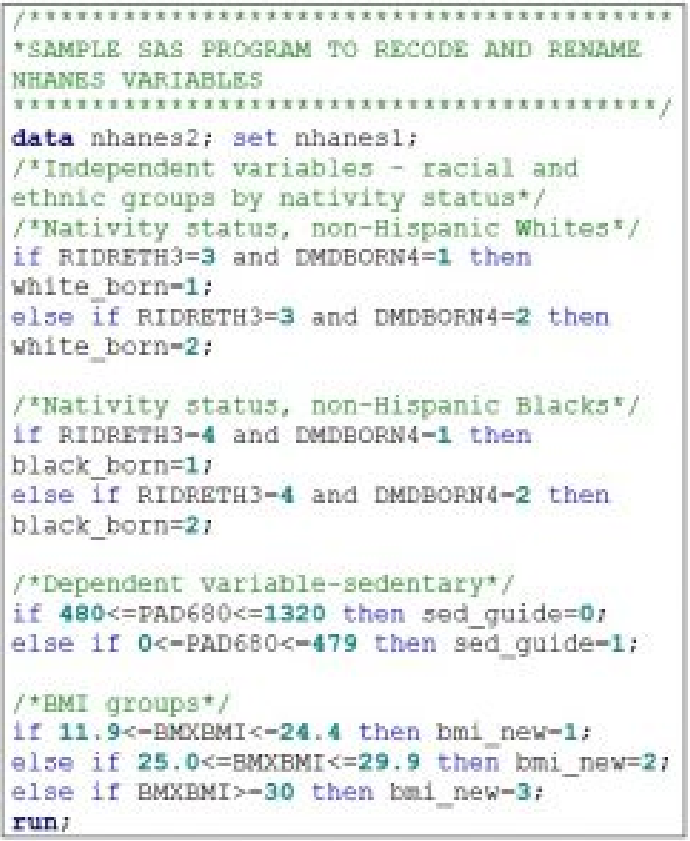
Questionnaire and examination responses often need to be recoded or responses collapsed prior to conducting statistical analysis. For example, the NHANES has response options "7777=Refused" and "9999=Don't know" for several questions. These responses are often removed and made "missing" prior to analysis. Furthermore, the numbers that represent certain values may need to be changed for easier interpretation of statistical analysis results. For example, NHANES has response options "1=Yes" and "2=No." It is common practice to change "no" responses to 0, "0=No." It is best practice to rename these recoded variables with a new variable name instead of replacing the original variable. Two or more variables may need to be combined in order to create the independent, dependent or other variables to answer study aims. In this case study, we will examine racial and ethnic differences in sedentary behavior by nativity status. Therefore, we will combine two variables for 1) race and ethnicity and 2) country of birth. An overview of the variables recoded and renamed for analysis in this case study is provided in Table 10.3.

Table 10.3 Overview of NHANES variables and recodes

Variable Description	Original Variable	Original Responses	Renamed Variable	Recoded Responses
Race/ Hispanic origin w/ non-Hispanic (NH) Asian	RIDRETH3	1=Mexican American 2=Other Hispanic 3=NH White 4=NH Black 5=None 6=NH Asian 7=Other, Multiple	WHITE_BORN	1=US-born White 2=Foreign-born White
Country of birth	DMDBORN4	1=Born in 50 US states/DC 2=Others 77=Refused 99=Don't know	BLACK_BORN	1=US-born Black 2=Foreign-born Black
			HISP_BORN	1=US-born Hispanic 2=Foreign-born Hispanic
			ASIAN_BORN	1=US-born Asian 2=Foreign-born Asian
Minutes sedentary behavior	PAD680	0-1320 Range of values 7777=Refused 9999-Don't know	SED_GUIDE	0=Not adherent (>=480 minutes) 1=Adherent (<480 minutes)
Age in years at screening	RIDAGEYR	0-79=0-79 years 80=80+ years	AGE_NEW	1=18-64 years 2=65+ years
Body Mass Index (kg/ m^2)	BMXBMI	11.9 to 92.3=Range of values	BMI_NEW	1=Healthy or underweight (BMI<=24.4) 2=Overweight (BMI>=25 and BMI<=29.9) 3=Obese (BMI>=30)

A sample SAS program for recoding and renaming NHANES data for this case study is provided in Box 10.3. All recodes are available in the full syntax file provided on the course website.

Box 10.3. Sample SAS program to recode and rename NHANES variables



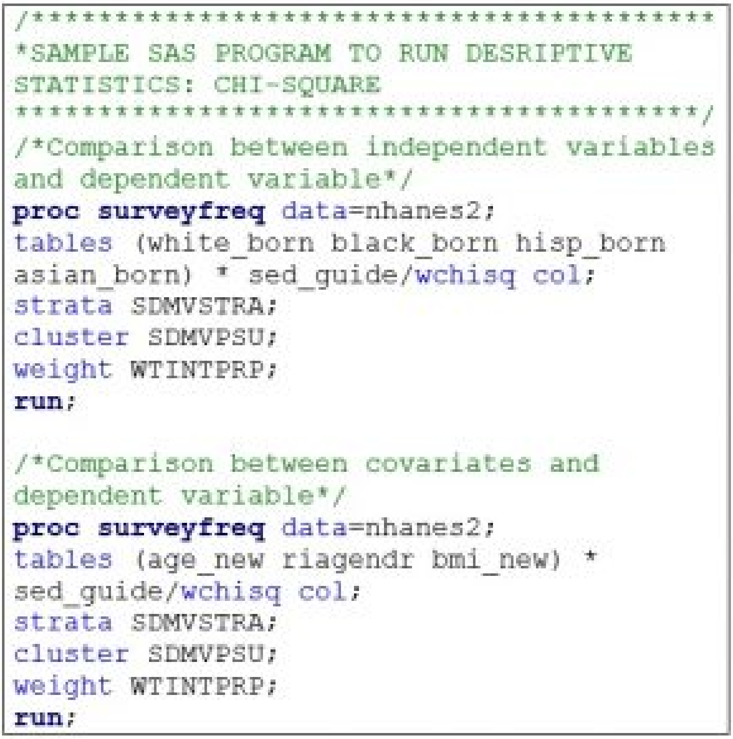
```
/*****************************************
*SAMPLE SAS PROGRAM TO RECODE AND RENAME
NHANES VARIABLES
*****************************************/
data nhanes2; set nhanes1;
/*Independent variables - racial and
ethnic groups by nativity status*/
/*Nativity status, non-Hispanic Whites*/
if RIDRETH3=3 and DMDBORN4=1 then
white_born=1;
else if RIDRETH3=3 and DMDBORN4=2 then
white_born=2;

/*Nativity status, non-Hispanic Blacks*/
if RIDRETH3=4 and DMDBORN4=1 then
black_born=1;
else if RIDRETH3=4 and DMDBORN4=2 then
black_born=2;

/*Dependent variable-sedentary*/
if 480<=PAD680<=1320 then sed_guide=0;
else if 0<=PAD680<=479 then sed_guide=1;

/*BMI groups*/
if 11.9<=BMXBMI<=24.4 then bmi_new=1;
else if 25.0<=BMXBMI<=29.9 then bmi_new=2;
else if BMXBMI>=30 then bmi_new=3;
run;
```

Step 5: Conduct Descriptive Statistical Analysis

Once all variables are recoded, collapsed, and renamed they can be used for statistical analysis. Statistical analysis should always start with descriptive analysis to describe the data source. Chi square analyses should be conducted to make comparisons between the independent variables, covariates, and dependent variables. It is important to remember that all analysis of NHANES data needs to be conducted with SAS survey procedures due to the complex sample design. Weighting (variable: WTINTPRP for questionnaire data), primary sampling unit (variable: SDMVPSU) and stratum (variable: SDMVSTRA) variables must be included in the programming statements.

A sample SAS program for conducting chi-square tests using 2017-2020 pre-pandemic NHANES data for this case study is provided in Box 10.4.

Box 10.4. Sample SAS program for running descriptive statistics (chi-square)

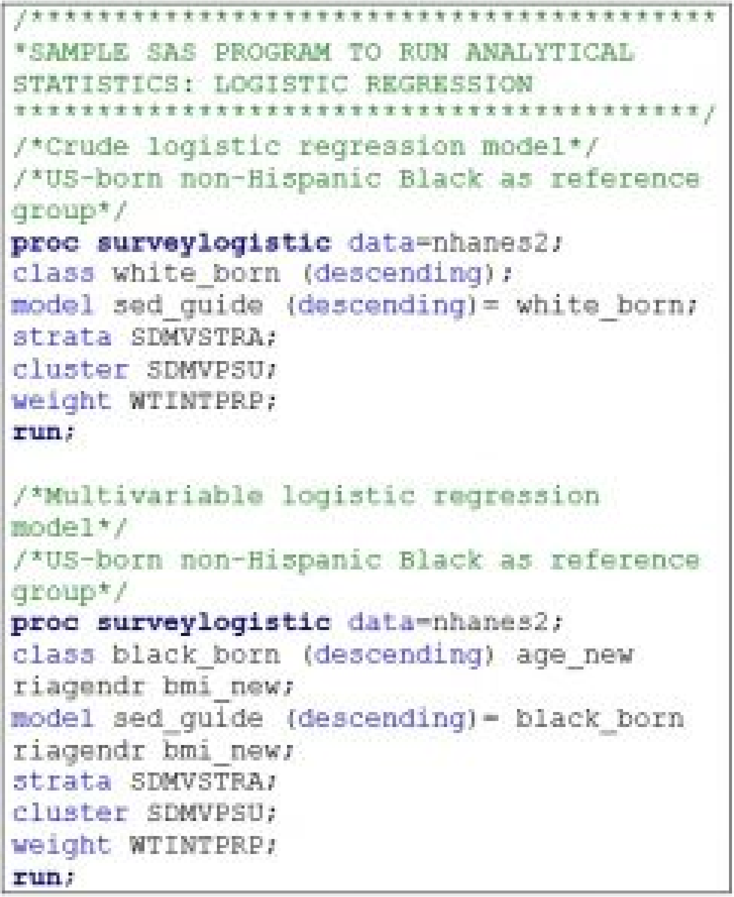


```
/*****************************************
*SAMPLE SAS PROGRAM TO RUN DESRIPTIVE
STATISTICS: CHI-SQUARE
*****************************************/
/*Comparison between independent variables
and dependent variable*/
proc surveyfreq data=nhanes2;
tables (white_born black_born hisp_born
asian_born) * sed_guide/wchisq col;
strata SDMVSTRA;
cluster SDMVPSU;
weight WTINTPRP;
run;

/*Comparison between covariates and
dependent variable*/
proc surveyfreq data=nhanes2;
tables (age_new riagendr bmi_new) *
sed_guide/wchisq col;
strata SDMVSTRA;
cluster SDMVPSU;
weight WTINTPRP;
run;
```

Step 6: Conduct Inferential Statistical Analysis

After calculating descriptive statistics, inferential statistical analysis can be conducted. Crude and multivariable logistic regression models can be calculated to determine associations between race, ethnicity, nativity status and sedentary guideline adherence. Crude or unadjusted logistic regression models are used to determine the association between the independent and dependent variables without adjusting for other factors. Multivariable or adjusted logistic regression models are used to determine associations between the independent and dependent variables after adjusting for potential covariates (e.g. age, gender, BMI). A reference category for the independent variable in needed. For this analysis, the reference group for each racial and ethnic group will be those born in the US (e.g. US-born non-Hispanic Whites, US-born non-Hispanic Blacks). A sample SAS program for conducting logistic regression analysis using 2017-2020 pre-pandemic NHANES data for this case study is provided in Box 10.5.

Box 10.5. Sample SAS program for running inferential statistics (logistic regression)



```
/****************************************
*SAMPLE SAS PROGRAM TO RUN ANALYTICAL
STATISTICS: LOGISTIC REGRESSION
****************************************/
/*Crude logistic regression model1*/
/*US-born non-Hispanic Black as reference
group*/
proc surveylogistic data=nhanes2;
class white_born (descending);
model sed_guide (descending)= white_born;
strata SDMVSTRA;
cluster SDMVPSU;
weight WTINTPRP;
run;

/*Multivariable logistic regression
model*/
/*US-born non-Hispanic Black as reference
group*/
proc surveylogistic data=nhanes2;
class black_born (descending) age_new
riagendr bmi_new;
model sed_guide (descending)= black_born
riagendr bmi_new;
strata SDMVSTRA;
cluster SDMVPSU;
weight WTINTPRP;
run;
```

10.7 SUMMARY

This chapter provided an overview of the NHANES and ways to conduct basic statistical analysis using 2017-2020 pre-pandemic public-use data files. The NHANES case study explored differences in sedentary guideline adherence among US- and foreign-born adults by race and ethnicity. Sample SAS programming statements were provided for downloading and inputting data files, merging data files, recoding and renaming variables and conducting categorical descriptive and inferential statistical analysis. The dataset and full SAS programming statements for the NHANES case study are available in the chapter 10 folder in the Open ICPSR data repository.

10.8 REFERENCES

1. NHANES – About the National Health and Nutrition Examination Survey. Published January 8, 2020. Accessed November 3, 2021. https://www.cdc.gov/nchs/nhanes/about_nhanes.htm
2. Chen TC, Parker JD, Clark J, Shin HC, Rammon JR, Burt VL. National Health and Nutrition Examination Survey: Estimation Procedures, 2011-2014. *Vital Health Stat 2.* 2018;(177):1-26.
3. Chen TC, Clark J, Riddles MK, Mohadjer LK, Fakhouri THI. National Health and Nutrition Examination Survey, 2015-2018: Sample Design and Estimation Procedures. *Vital Health Stat 2.* 2020;(184):1-35.
4. NHANES Questionnaires, Datasets, and Related Documentation. Accessed April 4, 2022. https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?Cycle=2017-2020
5. Centers for Disease Control and Prevention. NHANES Analytic Guidance and Brief Overview for the 2017-March 2020 Pre-pandemic Data Files. https://wwwn.cdc.gov/nchs/nhanes/

continuousnhanes/overviewbrief.aspx?Cycle=2017-2020

6. NHANES 2017-March 2020 Pre-Pandemic Dietary Data. Accessed April 4, 2022. https://wwwn.cdc.gov/nchs/nhanes/search/DataPage.aspx?Component=Dietary&Cycle=2017-2020
7. NHANES 2017-March 2020 Pre-Pandemic Examination Data. Accessed June 30, 2022. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Examination&Cycle=2017-2020
8. NHANES 2017-March 2020 Pre-Pandemic Laboratory Data. Accessed June 30, 2022. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Laboratory&Cycle=2017-2020
9. NHANES 2017-March 2020 Pre-Pandemic Questionnaire Data. Accessed June 30, 2022. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=Questionnaire&Cycle=2017-2020
10. NHANES 2017-March 2020 Pre-Pandemic Limited Access Data. Accessed June 30, 2022. https://wwwn.cdc.gov/nchs/nhanes/search/datapage.aspx?Component=LimitedAccess&Cycle=2017-2020
11. NCHS Data Linkage -Activities. Published February 14, 2022. Accessed June 30, 2022. https://www.cdc.gov/nchs/data-linkage/index.htm

12. RDC – Restricted Data – NHANES. Published August 27, 2021. Accessed June 23, 2022. https://www.cdc.gov/rdc/b1datatype/Dt1222.htm

13. Ross R, Chaput JP, Giangregorio LM, et al. Canadian 24-Hour Movement Guidelines for Adults aged 18-64 years and Adults aged 65 years or older: an integration of physical activity, sedentary behaviour, and sleep. *Appl Physiol Nutr Metab*. 2020;45(10 (Suppl. 2)):S57-S102. doi:10.1139/apnm-2020-0467

14. Paulose-Ram R, Graber JE, Woodwell D, Ahluwalia N. The National Health and Nutrition Examination Survey (NHANES), 2021-2022: Adapting Data Collection in a COVID-19 Environment. *Am J Public Health*. 2021;111(12):2149-2156. doi:10.2105/AJPH.2021.306517

10.9 COVID-19 PANDEMIC CHANGES

Sections 10.1 to 10.8 were written during the initial waves of the COVID-19 pandemic. Since the NHANES conducts in-person surveys at participants' households and mobile examination units, there were significant disruptions to the regular methodology due to stay-at-home orders and safety concerns for both participants and the data collectors. All data collection procedures were halted in March 2020. Several changes were made to the design to: ensure safety of the staff and participants; reduce response burden by only collecting essential data; and provide additional COVID-19 specific content.[14] Plans were made for data collection to begin in June 2021;

however, data collection procedures continue to remain halted as of this writing. Full details of the changes are reported elsewhere.[14]

CHAPTER 11.

DISSEMINATION

11.1 INTRODUCTION

The final step in the research process is to report your findings.[1] This chapter covers the dissemination of research studies using secondary data from national health surveys. It includes details on how to disseminate results by abstracts, presentations, and original research manuscripts. This chapter builds on previous curricula designed to train medical students[2] and physician assistant students[3] in research methods. The previous curricula materials have been modified and enhanced for researchers focused on disseminated results from secondary data analysis.

11.2 ABSTRACTS

Abstracts are brief summaries (typically 150-350 words) of preliminary findings or completed research projects. Abstracts are included at the beginning of most research manuscripts to inform the reader of the purpose, methods, most important findings, and implications of the research study. However, prior to being written to accompany manuscripts, abstracts can be written and submitted to professional organizations in calls for presentations during scientific sessions at professional meetings. Abstracts are usually structured like manuscripts using the following four sections: 1) introduction; 2) methods; 3) results; and 4) discussion (IMRAD).[4] The IMRAD structure aligns with the processes of scientific discovery and health research. The health research process includes identifying a study question (introduction), selecting the study approach, designing the study and collecting data (methods), analyzing data (results), and reporting findings (discussion).[1] However, some abstracts are unstructured without headings. Both formats include the same basic information about the research study. For abstracts using secondary data, it is important to include the source of the data as well as the years of data analyzed. Abstract guidelines for different professional meetings vary by organization. These guidelines should be provided with the call for abstracts shared on each professional meeting's website. An overview of some guidelines for public health research conferences that accept abstracts for studies using secondary data from national health surveys is provided in Table 11.1.

TABLE 11.1. SAMPLE ABSTRACT REQUIREMENTS FOR PROFESSIONAL MEETINGS ACCEPTING ABSTRACTS FOR RESEARCH USING SECONDARY DATA FROM NATIONAL HEALTH SURVEYS FOR PRESENTATIONS

	Academy Health	**American Public Health Association**	**Gerontological Society of America**
Word limit	**500**	**250**	**250**
Headings	Research Objective Study Design Population Studied Principal Findings Conclusions Implications for Policy or Practice	Background Methods Results Conclusions	None
Additional Requirements	None	At least 1 learning objective	At least 1 learning objective

**Note. This table has been adapted and updated from Table 11.1: Abstract requirements for common professional meetings in Kindratt & Kitzman-Ulrich's (2014) chapter on dissemination in Gimpel & Mokuria (eds.) Community Action Research in Family Medicine Residencies: A Community Medicine Handbook.*[5]

Abstracts accepted for conference presentation are usually published on the organizations' website as part of the program for the meeting. Some professional organizations partner with journals to publish abstracts is supplementary issues of their journals. For example, the Gerontological Society of America publishes abstracts from the previous year in a supplemental issue of Innovation in Aging. Examples of both primary and secondary research abstracts presented at professional meetings and published in journals by this textbook's primary author are provided in Table 11.2. Several primary studies were led and disseminated by students and residents.

Table 11.2. Examples of abstract publications[6-9] after presentation at professional meetings

Conference	Authors & Year	Title	Journal
2021 Gerontological Society of America Annual Scientific Meeting	Kindratt et al., 2021[6]	ADRD Caregiving Experiences and Health by Race, Ethnicity and Care Recipient Geographic Context	Innovation in Aging
2020 Academy Health Annual Research Meeting	Kindratt et al., 2020[7]	Cognitive disability among Arab Americans by nativity status: lack of evidence for the healthy migrant effect	Health Services Research
2019 Gerontological Society of America Annual Scientific Meeting	Dallo & Kindratt, 2019[8]	The epidemiology of Alzheimer's disease and related dementias among Arab Americans.	Innovation in Aging
2018 Food & Nutrition Conference & Expo	Xiao et al., 2018[9]	Teaching mobile health technology.	Journal of the Academy of Nutrition and Dietetics

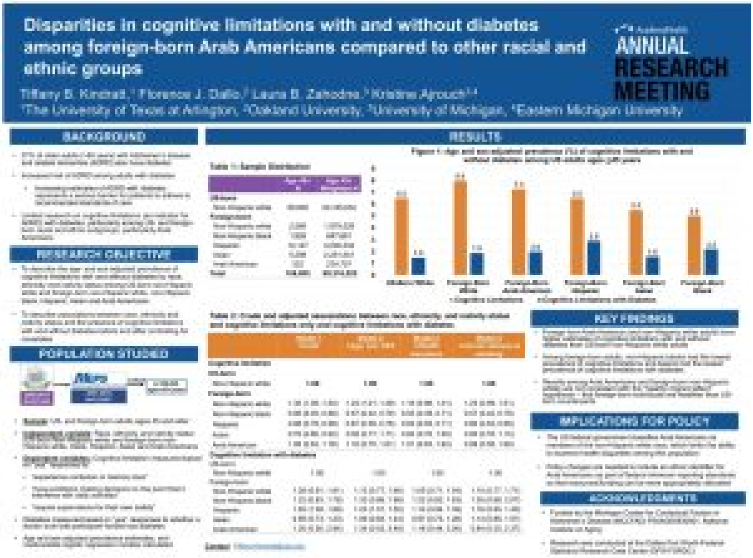
11.3 PRESENTATIONS

Part of the scientific process is to disseminate research by poster and oral presentation at professional scientific meetings prior to publication in peer reviewed journals. Poster presentations are often used for preliminary findings and oral presentations highlight the results from completed research studies. By presenting at professional scientific meetings, researchers are able to obtain feedback on their research methods and make changes as needed. Researchers may also network and learn about similar research studies being conducted that they may not know about because the results have not been published yet. A brief overview of poster and oral presentations with examples are provided in this section. Some details are also provided on the impact of the COVID-19 pandemic on presentations at professional scientific meetings.

11.3.1 POSTER PRESENTATIONS

Poster presentation sessions have been utilized at professional meetings in the US since the 1970s.[11] Poster presentations are a useful way for students to present their research findings in a way that is usually considered to be less stressful than a formal oral presentation in front of an audience. Poster presentations are often created as one slide using Microsoft PowerPoint. Most posters are large (3ft x 4ft) and landscape format. An example is provided in Figure 11.1.

Figure 11.1. Landscape poster example from 2021 Academy Health Annual Research Meeting



Disparities in cognitive limitations with and without diabetes among foreign-born Arab Americans compared to other racial and ethnic groups

Tiffany B. Kindratt,[1] Florence J. Dallo,[2] Laura B. Zahodne,[3] Kristine Ajrouch[3,4]

[1]The University of Texas at Arlington, [2]Oakland University, [3]University of Michigan, [4]Eastern Michigan University

BACKGROUND

RESULTS

RESEARCH OBJECTIVE

POPULATION STUDIED

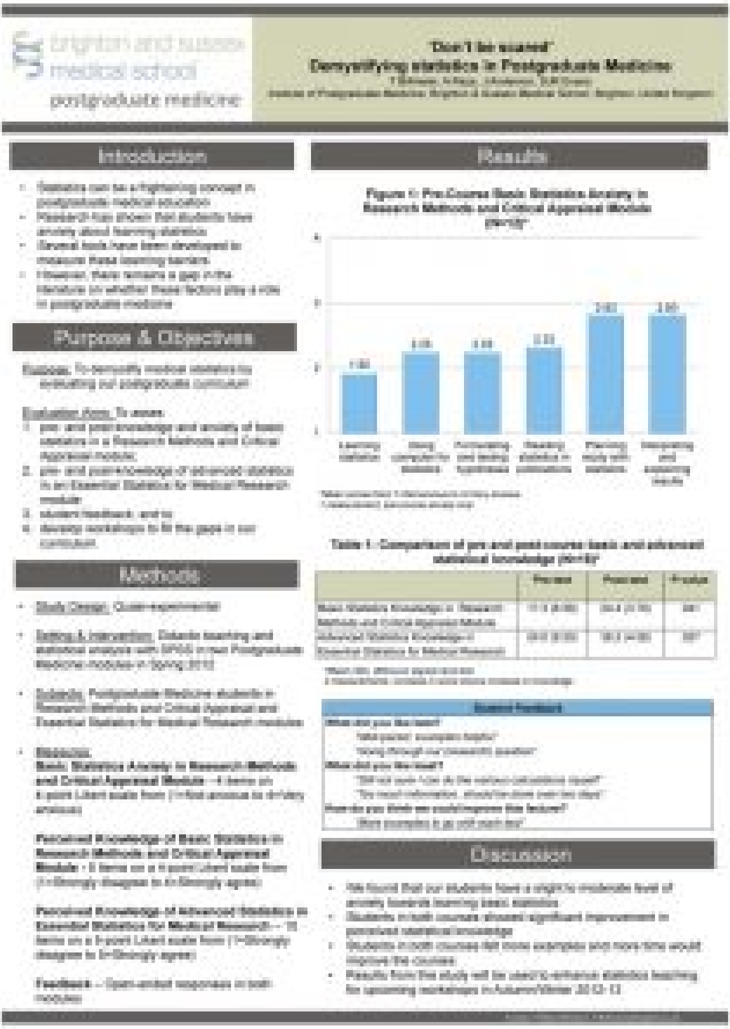
KEY FINDINGS

IMPLICATIONS FOR POLICY

ACKNOWLEDGEMENTS

It may be more common for conferences overseas to have posters presented in a vertical format. An example of a poster presentation from a conference in England is provided in Figure 11.2.

Figure 11.2. Vertical poster example from 2012 Association for the Study of Medical Education (ASME) Conference in Brighton, England



brighton and sussex medical school postgraduate medicine

"Don't be scared"
Demystifying statistics in Postgraduate Medicine

[illegible]

Introduction

- Statistics can be a frightening concept in postgraduate medical education
- Research has shown that students have anxiety about learning statistics
- Several tools have been developed to measure these learning barriers
- However, there remains a gap in the literature on whether these factors play a role in postgraduate medicine

Purpose & Objectives

Purpose: To demystify medical statistics by evaluating our postgraduate curriculum

Evaluation Aims: To assess:

1. pre- and post-knowledge and anxiety of basic statistics in a Research Methods and Critical Appraisal module
2. pre- and post-knowledge of advanced statistics in an Essential Statistics for Medical Research module
3. student feedback, and to
4. develop workshops to fill the gaps in our curriculum

Methods

- Study Design: Quasi-experimental
- Setting & Intervention: Didactic teaching and statistical analysis with SPSS in two Postgraduate Medicine modules in Spring 2012
- Subjects: Postgraduate Medicine students in Research Methods and Critical Appraisal and Essential Statistics for Medical Research modules
- Measures: Basic Statistics Anxiety in Research Methods and Critical Appraisal Module – [illegible]

Perceived Knowledge of Basic Statistics in Research Methods and Critical Appraisal Module – [illegible]

Perceived Knowledge of Advanced Statistics in Essential Statistics for Medical Research – [illegible]

Feedback – Open-ended responses in both modules

Results

Figure 1: Pre-Course Basic Statistics Anxiety in Research Methods and Critical Appraisal Module

Learning statistics
Using computers for statistics
Formulating and testing hypotheses
Reading statistics in publications
Planning study with statistics
Interpreting and explaining results

Table 1: Comparison of pre and post course basic and advanced statistical knowledge

	Pre-test	Post-test	P-value
Basic Statistics Knowledge in Research Methods and Critical Appraisal Module	[illegible]	[illegible]	[illegible]
Advanced Statistics Knowledge in Essential Statistics for Medical Research	[illegible]	[illegible]	[illegible]

Student Feedback

What did you like best?

What did you like least?

How do you think we could improve this lecture?

Discussion

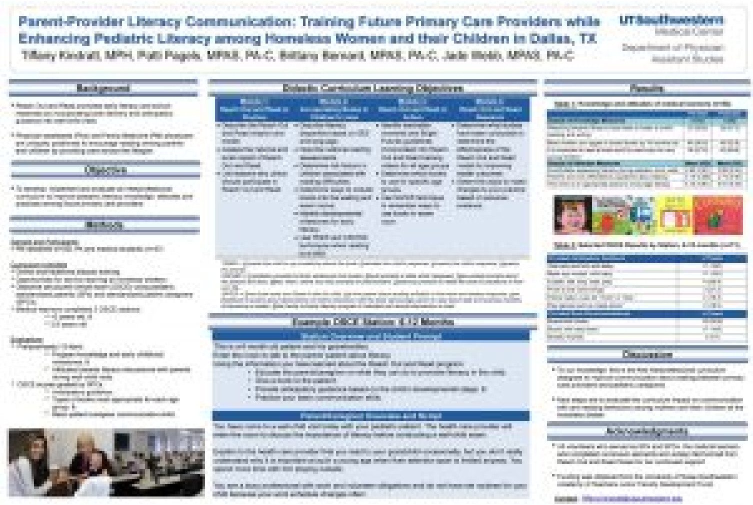
- We found that our students have a slight to moderate level of anxiety towards learning basic statistics
- Students in both courses showed significant improvement in perceived statistical knowledge
- Students in both courses felt more examples and more time would improve the courses
- Results from this study will be used to enhance statistics teaching for upcoming workshops in Autumn/Winter 2012-13

Some benefits of a poster presentation are that research findings can reach a larger audience and it may be easier to engage in a conversation and network with professional contacts. Poster presentations also allow the audience member to view multiple points of entry.[10] The audience member may view the title and go straight to the results section before having to view the introduction or methods of the presentation. An effective poster presentation is a condensed version of a full research abstract and may also be referred to as an illustrated abstract.[9] It should include all IMRAD sections as well an acknowledgment of the funding source (if applicable) and contact information for the principal investigator so that the viewer may contact them after the presentation. Additional tips on creating poster presentations include:

1. the less text the better
2. use a specific title related to the research aim (see Figure 11.1), or title that tries to catch the audience's attention (see Figure 11.2)
3. use bright colors
4. use pictures to portray a specific intervention or population of interest, particularly if conducting primary research with students or community members.

A photo release form is recommended if using pictures. An example of a poster presentation with pictures of physician assistant students and community member participants is provided in Figure 11.3.

Figure 11.3. Poster example using pictures of student and community member participants



Parent-Provider Literacy Communication: Training Future Primary Care Providers while Enhancing Pediatric Literacy among Homeless Women and their Children in Dallas, TX

Tiffany Kindratt, MPH, Patti Pagels, MPAS, PA-C, Brittany Bernard, MPAS, PA-C, Jade Webb, MPAS, PA-C

UTSouthwestern Medical Center

Department of Physician Assistant Studies

Background

[illegible]

Objectives

[illegible]

Methods

[illegible]

Didactic Curriculum Learning Objectives

[illegible]

Example OSCE Station: 4-12 Months

[illegible]

Results

[illegible]

Discussion

[illegible]

Acknowledgements

[illegible]

11.3.2 ORAL PRESENTATIONS

An oral or platform presentation allows the researcher to practice sharing their research findings in a formal presentation. Oral presentations at professional meetings typically range from 10-20 minutes with 5-10 minutes at the end designated for audience members to ask questions. Oral presentations are usually created using Microsoft PowerPoint slides or slides from a similar program. Slides should emphasize key points to engage with the audience. Similar to poster presentations, the oral presentation should include a combination of bulleted text, figures, tables and pictures. It is sometimes recommended that researchers create a poster presentation first and then transfer the information from each section to a new set of slides for the oral presentation. A good rule of thumb is to include 1 slide per minute of each presentation. Slides should use basic text and a large font of size 28 or larger for visual accessibility. The presenter should use the slides as a guide but not just read directly off each slide while presenting. Animations and transition slides can be used; however, they may be distracting and difficult for individuals with seeing disabilities. Since the short timeframe of oral presentations may not all allow the investigator to include all of the details of the research study, presenters may be interested in including some supplemental slides at the end in case the audience asks questions about any content that was not covered in the presentation.[10]

11.3.3 VIRTUAL AND ONLINE PRESENTATIONS

In March 2020, stay-at-home orders and safety precautions due to the COVID-19 pandemic forced most professional meetings to be virtual using online platforms, including social media (e.g. twitter presentations), video conferencing (e.g. Zoom), and other software systems.[11] Some presentations were required to be given live while others were pre-recorded and posted to YouTube or a meeting portal. There are several benefits and limitations to hosting virtual conferences.[12] Some of the benefits include the ability to reach a wider international audience and allow those who may not have the funding or other ability to attend in-person meetings due to obligations at home. Some of the limitations include technical issues while presenting, such as issues with lighting, webcams, and inconsistent internet access.[12] While presenting in virtual panel sessions, at least one or more of the presenters may have technical difficulties. Virtual conferences may also limit the ability of the presenter to fully engage in the presentation. For example, the presenter may have children at home due to school closures or have other work obligations that they may not have been required to attend if they were at an in-person meeting. Virtual conference attendance may allow some individuals greater comfort in networking activities while others may be more comfortable networking with their peers in-person. Regardless of a researcher's preference for in-person, virtual, or hybrid presentation formats, virtual poster and oral presentations allow researchers to share their work in additional accessible ways which will most likely continue to be offered as a mode of dissemination beyond the COVID-19 pandemic.[11,12]

11.4 MANUSCRIPTS

The final step in the research process is to write and submit a manuscript for peer-reviewed publication. The publication process is completed after conducting poster and oral presentations so that any feedback obtained can be incorporated into the final manuscript. Writing manuscripts using the scientific method is a teachable skill. Scientific writing is formulaic, comprising of short and concise simple sentences. Five principles of scientific writing include: 1) clarity; 2) simplicity; 3) conciseness; 4) exactness, and 5) authenticity.[2,3] Once the principal investigator decides on the journal to submit to, the manuscript should be tailored to that journal. Journals should provide instructions for authors on their specific formatting that should be followed before submission, including word count limits. Original research articles allow for word counts ranging from 2,000 to 5,000 words. Brief reports can limit word counts from 1,000 to 2,500 words. Letters to the editor are even shorter and may be limited to 500 words. Review articles are usually longer to allow for a comprehensive assessment of all research addressing the topic of interest in the study.

11.4.1 MANUSCRIPT SECTIONS

Manuscripts usually start with a brief abstract and then are usually structured using the following four sections: 1) introduction; 2) methods; 3) results; and 4) discussion (IMRAD).[4] As mentioned in the section on abstracts, this IMRAD structure aligns with the processes of scientific discovery and health research, which includes identifying a study question (introduction), selecting the study approach, designing the study and collecting data (methods), analyzing data (results), and reporting findings (discussion).[1]

Online platforms for journals allow for including supplemental material to complement these sections. However, not all journals follow this format. For example, in the journal Innovation in Aging, the methods section is replaced with "materials and methods." Each section usually requires subheadings to structure the material for the reader. These subheadings may be required by the journal but are usually designated by the authors. For example, the introduction may include a subheading for "literature review." The results may include subheadings for "selected characteristics" or "bivariate analysis." Finally, the discussion may include a subheading for "strengths and limitations" or "conclusions."

When writing research manuscripts, authors can benefit from the use of writing checklists. These checklists are sometimes required by journals. An example is the Strengthening the Reporting of Observational Studies in Epidemiology (STROBE) checklist which is used for writing observational studies.[13] The Consolidated Standards of Reporting Trial (CONSORT) statement provides guidance for reporting randomized clinical trials.[14] The Preferred Reporting Items for Systematic Reviews and Meta-Analyses (PRISMA) statement is used for systematically and accurately reporting reviews.[15]

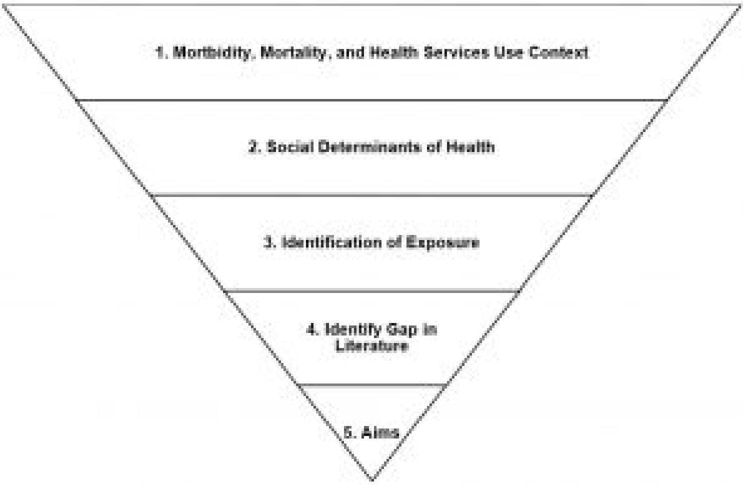
11.4.1.a Abstract

As mentioned earlier in this chapter, the abstract provides a brief summary of the article. There is a need to ensure consistency with the information in the abstract and the text of the manuscript.[4] The abstract can be the most important part of the research article because it is the first thing that readers will view prior to accessing and reading full articles. While some full-text articles require fees or library access to obtain, abstracts are available on the internet free to all readers. Abstracts may be written before the final manuscript as an outline for what to include in the formal paper. Another strategy is to write the abstract after completing the paper so that the writer can rephrase key points from the completed manuscript that may be the most impactful to potential readers.

11.4.1.b Introduction

The introduction starts the manuscript. A common framework used for developing an introduction section for research projects using data from national health surveys is an inverted pyramid or funnel approach (see Figure 11.4). The funnel approach allows the writer to describe the importance and rationale of the study from a broad perspective and then narrow it down to its specific aims.[16] With this approach, the introduction is separated into five sections that can be combined into 3 paragraphs or left separate depending on the article type (e.g., original research or brief report) and the variables of interest. The five sections include: 1) morbidity, mortality, and health services use context; 2) social determinants of health; 3) identification of exposure; 4) gap in the literature; and 5) specific aims of the study.

Figure 11.4. Overview of introduction section funnel outline for writing manuscripts using national health surveys


1. Mortbidity, Mortality, and Health Services Use Context
2. Social Determinants of Health
3. Identification of Exposure
4. Identify Gap in Literature
5. Aims

The first paragraph of the introduction should provide the broad context of the outcome. Its purpose is to contextualize the dependent variable or outcome. A question that the writer might address could be: "what is the morbidity and/or mortality of the disease and outcome under investigation?" Let's uses Alzheimer's disease and related dementias (ADRD) as the outcome. The writer may want to start the introduction by highlighting the burden of ADRD in the United States (over 6 million individuals), including the costs for treatment (over $350 million per year).[17] The second paragraph for many public health research studies focuses on social determinants of health. This paragraph starts to narrow down the context of the outcome by including details on the health disparities that exist as they relate to race, ethnicity, sex, gender, age or other social determinants of health. If ADRD is still our example, we may want to include studies that demonstrate differences in the prevalence of ADRD among non-Hispanic Black individuals compared to non-Hispanic White individuals. The third paragraph focuses on the exposure, or independent variable. It includes details on the exposure that is being investigated and what research has already been conducted on the relationship between the exposure and outcome. For studies looking at social determinants of health as an exposure, this may be combined with the previous paragraph. For example, if we look at racial and ethnic disparities in the prevalence of ADRD, we may want to go further and determine whether differences exist among foreign-born and US-born racial and ethnic minority groups since previous studies indicate that foreign-born individuals tend to have better health outcomes than their US-born counterparts. The fourth paragraph identifies the gap in the literature on the relation between the exposure and the outcome. The fifth paragraph outlines the specific aims or objectives of the study. This may also include the research questions or hypotheses depending on the project and journal requirements.

11.4.1.c Methods

The methods section should include the details of how the study was conducted and why the study was conducted in a particular fashion. It should be written in past tense since the methods have already been conducted.[18] The choice of active or passive voice when writing is dependent on where the paper is submitted. The methods section should provide enough detail that anyone who wanted to replicate the study could do so. This is particularly important when writing manuscripts using public-use data because any researcher could download the data and run the same analysis but yield different results. When writing articles using large data sources, it is recommended that there are specific subsections to describe the data source, participants, variables, statistical analysis, and institutional review board (IRB) approval process. It is recommended that the writer writes the answers to these questions delineated for each section in paragraph form.

Data Source

When reporting the results of a research study using national data, it is important to identify the national health survey that was used in the study. The years that were combined or specific data files (e.g., sample adult file for NHIS data or longitudinal file for MEPS data) should also be reported. Since the details of each survey have already been published on the survey website, this section should only include brief details on the history and purpose of the national health survey. The reader should also be referred to the website for more details and citations should be included for other studies that may have used the same methodology to create the sample with this data source. For example, Dallo and Kindratt used restricted NHIS data for several studies using the country of birth data from the NHIS. The original manuscript published included more details on the data sources and methods used[19] and the later studies[20,21] included a reference to the original study published. The level of details requested about the national health survey used is dependent on the peer reviewers and editor.

Participants

Based on the sampling frame, it is important to include the total number of participants who completed the national health survey used in the study. This section should include the details describing how the sample was limited based predetermined inclusion and exclusion criteria. The total number of unweighted participants that meet the inclusion and exclusion criteria should be included. It may also be beneficial to the reader if the total number of individuals that the unweighted sample represents when weighted is included.

Variables

The variables (or measures) section should include details on the questions used in your analysis to meet your study aims. This section may include subsections for independent variables (exposures or predictors), dependent variables (disease or outcomes), and covariates (including confounders, mediators, and other explanatory factors) selected for inclusion in the statistical analysis. Each section should include a description of the questions collected by the national health survey to gather data on each variable and ways that the investigator has changed the variable by combining it with other questions or removing missing responses. If applicable, these sections should include references to previous studies that have served as the basis for recoding variables or adjusting for specific covariates (e.g. confounders, other contributing factors) during the statistical analysis. As many details about the survey questions and response options should be included in this section in order to alleviate the readers concerns about potential information biases or concerns about internal validity.

Statistical Analysis

The statistical analysis section should include details on what analyses were conducted to meet the research aims. This section should include: 1) basic descriptive statistical procedures, such as frequencies, percentages, means and standard deviations; 2) comparative (or bivariate) statistical procedures, such as t-tests or chi square tests; and 3) inferential statistical procedures, such as regression analyses. Specific information about the weighting, stratification, and primary sampling units used for national health surveys should be included. It is common to cite the analytic guidelines for each specific national health survey in this section, specifically if changes have been made to these variables. The specific analytical software should be included to ensure that the proper procedures were used to account for the complex sample design.

Ethical Approval

Most scientific journals require a section or at least one sentence on ethical approval. The secondary analysis of public-use data is not considered human subjects research because the data extracted for analysis cannot be identified by the researcher. However, some universities require that research using public-use data from national health surveys go through a formal review process. These studies are often deemed as the "exempt" classification.

11.4.1.d Results

The results section is the "heart," or most important section of the paper.[22] It highlights the researchers' contribution to the scientific literature and overall general knowledge in the public health field. The results section should directly align with the objectives and hypotheses presented in the introduction, the methods outlined to meet those objectives in the methods section, and a comparison to other studies in the discussion section. Every research question in the introduction section should have a delineated method and related result. The results section should only present the main findings but not interpret them in the context of other literature.[23]

The results section includes a mixture of tables (figures, if applicable) and text. It should include all findings obtained while conducting the research study. At least one paragraph should be included for each table presented and tables should be referred to in the text. It may be useful to use headings that are similar to the table titles to direct the reader between the tables and text. Depending on the journal the article is being submitted to, authors may be requested to include a placeholder (e.g., insert table 1 here) in the document that is submitted for review. Results should be presented using at least two tables describing 1) the demographics of the sample and 2) the main analytic results. Tables should be presented by increasing complexity of the analysis (first demographics, second bivariate analysis, last multivariate analysis) and present large amounts of data in one place.[22,23] Each table should be able to stand alone with specific titles and footnotes to describe its contents.

Table 1

The first table usually provides demographic information about the sample. It may be presented with basic frequencies and percentages or including bivariate statistics that compare two groups. Key findings from table 1 should be included in written form in the text. Instead of including all of the results from the tables in the text, the author may want to only highlight statistically significant or clinically relevant results in the text. Examples of potential tables shells are provided below.

Table 11.3.a. Table shell representing how to present basic sample characteristics

	Unweighted N (%)	Weighted N (%)
Race, Ethnicity and Nativity		
US-Born Non-Hispanic Whites		
Foreign-born Non-Hispanic Whites		
Foreign-born Arab Americans		
Sex		
Male		
Female		

Table 11.3.b. Table shell representing ways to present bivariate analyses using sample characteristics

	Flu Vaccine Last 12 months		
	No **% (SE)**	**Yes** **% (SE)**	**p-value**
Race, Ethnicity and Nativity			
US-Born Non-Hispanic Whites			
Foreign-born Non-Hispanic Whites			
Foreign-born Arab Americans			
Sex			
Male			
Female			

Tables 2+

Table 2 and further tables include more complex statistical analysis. Since many national health surveys use cross-sectional designs, prevalence estimates may be calculated. Journal reviewers may be interested in seeing prevalence estimates reported in an adjusted analysis, such as the age- and sex- adjusted prevalence. These results are obtained by calculating predicted marginals using a LS MEANS statement with the regression analyses. Results may be presented in a table similar to Table 11.3.c, which demonstrates differences in age- and sex-adjusted prevalence estimates of cognitive health outcomes by race, ethnicity, and nativity status.

Table 11.3.c. Table shell for age- and sex-adjusted prevalence estimates

	US-Born	**Foreign-Born**	
Alzheimer's disease and related dementias (ADRD)			
ADRD with comorbid diabetes			
ADRD with comorbid hypertension			

Regression models may also be used for determining associations between the independent and dependent variables of interest before and after adjusting for covariates. Logistic, linear, multinomial, binomial, cox, Poisson, or other regression models may be presented. The explanations of each of these regressions models is beyond the scope of this textbook. An example table shell that may be used for logistic regression results is presented in Table 11.3.d.

Table 11.3.d. Table shell for logistic regression results

	Model 1 Crude OR (95% CI)	**Model 2 Adjusted for demographics OR (95% CI)**	**Model 3 Fully Adjusted OR (95% CI)**
Race, Ethnicity and Nativity			
US-Born Non-Hispanic Whites			
Foreign-born Non-Hispanic Whites			
Foreign-born Arab Americans			
**Include details of variables adjusted in footnotes underneath*			

Sensitivity Analysis

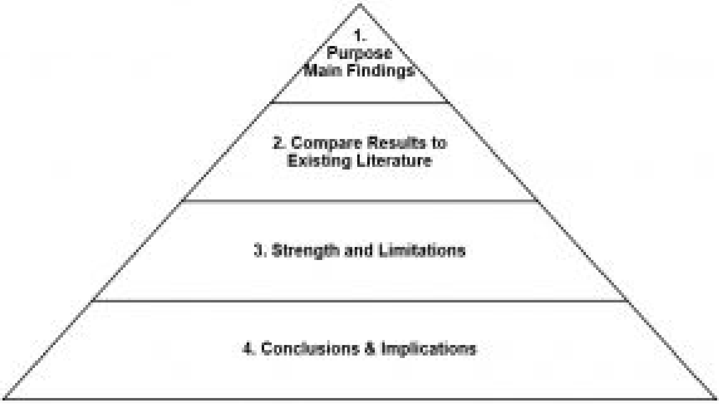
The methods and results sections may include details on a sensitivity analysis. Sensitivity analyses are used to determine whether a different model or set of assumptions will yield similar results.[24] For example, in research evaluating foreign-born Arab Americans using NHIS data, studies by Dallo and colleagues have used data from foreign-born individuals born in the Middle East to represent the Arab ethnicity.[25,26] Recent efforts to separate Arab Americans from non-Hispanic Whites in the US have recommended creating an inclusive racial/ethnic category for Middle Eastern and North African (MENA) populations.[27] Kindratt and colleagues expanded the formerly used Middle Eastern category to include non-Hispanic White Africans to represent

MENA individuals in a study evaluating cognitive limitations.[28] A sensitivity analysis was conducted to compare results from using the new MENA variable with results from using the Middle East only variable using logistic regression models. The odds ratios and 95% confidence intervals were overlapping, which indicated that the results were similar.[28] Results from a sensitivity analysis are sometimes included in the manuscript or included as supplementary material due to limitations on word count and the number of tables or figures allowed to be included with the text.

11.4.1.e Discussion

The purpose of the discussion section is to interpret the findings and provide meaning to the results in the context of the other medical literature. The discussion section should mirror the introduction section as a pyramid or reverse funnel (See Figure 11.5). Instead of starting off with the broad context and ending with a specific purpose and objectives, the discussion section starts off with specific results and expands to include the broader context throughout the section. With this approach, the discussion section is separated into four sections that can be expanded to five or more paragraphs depending on the article type (e.g., original research or brief report) and the exposures and outcomes of interest. The five sections include: 1) summary of the purpose and main findings; 2) comparison of results to existing literature; 3) strengths and limitations; and 4) conclusions and implications.

Figure 11.5. Overview of discussion section pyramid outline for writing manuscripts using national health surveys


1. Purpose Main Findings
2. Compare Results to Existing Literature
3. Strength and Limitations
4. Conclusions & Implications

The first paragraph of the discussion should restate the purpose of the study then briefly summarize the main findings. The implications of the main findings should be expanded on in the next section. The second section of the discussion should include a paragraph for each of the most relevant findings with references to compare and contrast the results with other studies. This section should include reasons why the results may be the same or different from other literature. The third section of the discussion should acknowledge the strengths and limitations of the study. For example, some strengths of studies using national health surveys may be that the sample was selected using a probability-based sample design versus a convenience sample. A strength may also be a large sample size. National health surveys also have many different content areas that allow for a broad assessment of other potential contributing factors related to the research question. Despite these strengths, there will also be limitations that need to be noted. For example, some limitations may include that the survey uses a cross-sectional design and causation cannot be established. Since the independent and dependent variables were measured at the same time, there may be no way of determining whether the independent variable (exposure) causes the dependent variable (disease/ outcome). There may also be information biases with self-reported data. For example, a limitation of self-reported data for cancer screenings is that the data may overrepresent or underrepresent screening estimates. Self-reported data may not be as accurate as other measures such as electronic medical records. The final paragraph of the discussion section is a conclusion. The conclusion should discuss the generalizability of the results and the impact the findings may have on potential interventions and policies. The manuscript should end with the "take home message" from the research and provide future directions and recommendations.

11.5 WRITING TIPS AND TRICKS

Here are some writing tips and tricks that may be useful.

- Start with the results section, then write the methods section. You may want to create your tables or figures first, then write 1-2 sentences about them.[23] By creating the tables and writing the results first, the introduction and discussion sections can be framed around the key findings.
- Select a similar article to use as a model for writing. If there is already an article published with the same dataset in the journal you plan to submit to, use it as a model for structuring your paper.

- If you wrote a grant proposal or IRB protocol, use the same information to start writing your paper. The proposals were probably written in future tense (e.g., “we will analyze” or “the data will be analyzed”) so change it to past tense (e.g., “we analyzed” or “the data were analyzed”).
- If you are unclear whether someone from the team qualifies as an author, check to see if they meet the ICJME guidelines on authorship.[4] If you have co-authors, make sure they are contributing to the manuscript and you are not doing everything yourself.
- Use a referencing software like Endnote or Zotero to manage your in-text citations and references at the end of the paper.
- When all else fails, hand write your manuscript on a piece of paper – not type it.

11.6 SUMMARY

In summary, this chapter provided an in-depth overview of disseminating research by presentation and publication. Dissemination is the final step in the research process and is essential when conducting research using national health surveys. Several examples have been provided to demonstrate different types of presentations created and sections of manuscripts written by the primary author of this textbook.

11.7 REFERENCES

1. Jacobsen KH. *Introduction to Health Research Methods: A Practical Guide*. 3rd ed. Jones & Bartlett

Learning; 2021.

2. Dehaven MJ, Gimpel NE, Dallo FJ, Billmeier TM. Reaching the underserved through community-based participatory research and service learning: description and evaluation of a unique medical student training program. *J Public Health Manag Pract*. 2011;17(4):363-368. doi:10.1097/PHH.0b013e3182214707
3. Kindratt TB. Research Extension Experience in Directed Studies: Solidifying Evidence-Based Medicine Competencies Through Research Participation. *J Physician Assist Educ*. 2020;31(1):36-41. doi:10.1097/JPA.0000000000000291
4. International Committee of Medical Journal Editors. Recommendations for the Conduct, Reporting, Editing, and Publication of Scholarly Work in Medical Journals. Accessed December 8, 2020. http://www.icmje.org/recommendations/
5. Kindratt T, Kitzman-Ulrich H. Dissemination: Reporting and writing. In: *Community Action Research in Family Medicine Residencies: A Community Medicine Handbook.* University of Texas Southwestern Medical Center Print Shop; 2014:133-150.
6. Kindratt T, Sylvers D, Yoshikawa A, Anuarbe ML, Webster N, Bouldin E. ADRD Caregiving Experiences and Health by Race, Ethnicity and Care Recipient Geographic Context. *Innov Aging*. 2021;5(Suppl 1):990. doi:10.1093/geroni/igab046.3557

7. Kindratt TB, Dallo FJ, Zahodne LB. Cognitive Disability Among Arab Americans By Nativity Status: Lack of Evidence for the Healthy Migrant Effect. *Health Services Research*. 2020;55(S1):21-21. doi:10.1111/1475-6773.13352

8. Dallo F, Kindratt T. The epidemiology of Alzheimer's disease and related dementias among Arab Americans. *Innov Aging*. 2019;3(Suppl 1):S463. doi:10.1093/geroni/igz038.1731

9. Xiao C, Kindratt T, Rodder S. Teaching Mobile Health Technology. *J Acad Nutr Diet*. 2018;118(9 Supplement):A34. doi:10.1016/j.jand.2018.06.138

10. Hess GR, Tosney KW, Liegel LH. Creating effective poster presentations: AMEE Guide no. 40. *Med Teach*. 2009;31(4):319-321. doi:10.1080/01421590902825131

11. Price M. Scientists discover upsides of virtual meetings. *Science*. 2020;368(6490):457-458. doi:10.1126/science.368.6490.457

12. Rubinger L, Gazendam A, Ekhtiari S, et al. Maximizing virtual meetings and conferences: a review of best practices. *Int Orthop*. 2020;44(8):1461-1466. doi:10.1007/s00264-020-04615-9

13. Vandenbroucke JP, Elm E von, Altman DG, et al. Strengthening the Reporting of Observational Studies in Epidemiology (STROBE): Explanation and Elaboration. *PLOS Medicine*. 2007;4(10):e297. doi:10.1371/journal.pmed.0040297

14. Schulz KF, Altman DG, Moher D, CONSORT

Group. CONSORT 2010 statement: updated guidelines for reporting parallel group randomised trials. *PLoS Med*. 2010;7(3):e1000251. doi:10.1371/journal.pmed.1000251

15. Page MJ, McKenzie JE, Bossuyt PM, et al. The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *BMJ*. 2021;372:n71. doi:10.1136/bmj.n71

16. Bahadoran Z, Jeddi S, Mirmiran P, Ghasemi A. The Principles of Biomedical Scientific Writing: Introduction. *Int J Endocrinol Metab*. 2018;16(4):e84795. doi:10.5812/ijem.84795

17. Alzheimer's Association. *2021 Alzheimer's Disease Facts and Figures*. Alzheimer's Association; 2021:108. Accessed April 15, 2021. https://www.alz.org/media/Documents/alzheimers-facts-and-figures.pdf

18. Ghasemi A, Bahadoran Z, Zadeh-Vakili A, Montazeri SA, Hosseinpanah F. The Principles of Biomedical Scientific Writing: Materials and Methods. *Int J Endocrinol Metab*. 2019;17(1):e88155. doi:10.5812/ijem.88155

19. Dallo FJ, Kindratt TB. Disparities in preventive health behaviors among non-Hispanic White men: heterogeneity among foreign-born Arab and European Americans. *Am J Mens Health*. 2015;9(2):124-131. doi:10.1177/1557988314532285

20. Dallo FJ, Kindratt TB. Disparities in Chronic Disease Prevalence Among Non-Hispanic Whites: Heterogeneity Among Foreign-Born Arab and

European Americans. *J Racial Ethn Health Disparities*. 2016;3(4):590-598. doi:10.1007/s40615-015-0178-8

21. Kindratt TB, Dallo FJ, Roddy J. Cigarette Smoking among US- and Foreign-Born European and Arab American Non-Hispanic White Men and Women. *J Racial Ethn Health Disparities*. 2018;5(6):1284-1292. doi:10.1007/s40615-018-0476-z
22. Bahadoran Z, Mirmiran P, Zadeh-Vakili A, Hosseinpanah F, Ghasemi A. The Principles of Biomedical Scientific Writing: Results. *Int J Endocrinol Metab*. 2019;17(2):e92113. doi:10.5812/ijem.92113
23. Iskander JK, Wolicki SB, Leeb RT, Siegel PZ. Successful Scientific Writing and Publishing: A Step-by-Step Approach. *Prev Chronic Dis*. 2018;15:E79. doi:10.5888/pcd15.180085
24. Szklo M, Nieto FJ. *Epidemiology: Beyond the Basics*. 3rd ed. Jones & Bartlett Learning; 2014.
25. Dallo FJ, Kindratt TB, Snell T. Serious psychological distress among non-Hispanic whites in the United States: the importance of nativity status and region of birth. *Soc Psychiatry Psychiatr Epidemiol*. 2013;48(12):1923-1930. doi:10.1007/s00127-013-0703-1
26. Dallo FJ, Kindratt TB, Zahodne L. Prevalence of Self-Reported Cognitive Impairment among Arab American Immigrants in the United States. *Innovation in Aging*. 2020;(igaa058). doi:10.1093/geroni/igaa058

27. Awad GH, Abuelezam NN, Ajrouch KJ, Stiffler MJ. Lack of Arab or Middle Eastern and North African Health Data Undermines Assessment of Health Disparities. *Am J Public Health*. 2022;112(2):209-212. doi:10.2105/AJPH.2021.306590

28. Kindratt TB, Dallo FJ, Zahodne LB, Ajrouch KJ. Cognitive Limitations Among Middle Eastern and North African Immigrants. *J Aging Health*. Published online May 23, 2022:8982643221103712. doi:10.1177/08982643221103712

CHAPTER 12.

CONCLUSIONS

12.1 OVERVIEW

This textbook sought to train future public health professionals, specifically Master of Public Health (MPH) students, how to conduct basic applied data analysis using secondary data collected from national health surveys. The goal was to eliminate gaps in knowledge, skills and analytical abilities that may prohibit MPH graduates from being successful in entry-level public health practice and research-focused positions. A brief recap of what was covered in each chapter of this textbook is provided in the following sections. Results from each of the case studies covered in Chapters 6-10 are also provided.

12.2 INTRODUCTION AND BASIC APPLIED DATA ANALYSIS RECAP

The first section included three chapters. Chapter 1 provided an overview of the textbook by outlining its purpose to train future public health professionals in the knowledge and skills to conduct applied secondary data analysis using national health surveys. Chapter 2 provided a general overview of the surveys used for the case studies presented in this textbook, including the National Health Interview Survey (NHIS) in Chapter 6, Medical Expenditure Panel Survey (MEPS) in Chapter 7, Health Information National Trends Survey (HINTS) in Chapter 8, Behavior Risk Factor Surveillance System (BRFSS) in Chapter 9, and the National Health and Nutrition Examination Survey (NHANES) in Chapter 10. Chapter 3 included a literature review of previous studies that have used national health surveys to answer public health and health services research-related research questions.

The second section included two chapters. Chapter 4 reviewed basic statistical functions commonly used for public health research questions. While this textbook was written for learners with some background knowledge of research methods and epidemiologic study designs, this chapter included basic terminology on types of data collected, descriptive (frequencies/percentages, means/ standard deviations) and analytical statistical procedures (chi square, logistic regression) used for analysis of national health surveys. Chapter 5 included details on additional survey design features needed to be considered when analyzing complex surveys, including weights, primary sampling units, and stratum variables. Data from the NHIS were used for SAS programming examples in these chapters.

12.3 NATIONAL HEALTH INTERVIEW SURVEY (NHIS) RECAP

Chapter 6 covered the background and details on how to obtain and analyze NHIS data. The objective of the NHIS case study was to explore whether Arab American adults were more or less likely to receive an annual flu vaccine in comparison to other racial/ethnic groups, such as other non-Hispanic Whites using 2018 person and sample adult files. The following specific aims were examined using chi square tests and logistic regression analyses.

- Aim 6.1: Compare socioeconomic and health-related characteristics of Arab Americans compared to US-born and foreign-born non-Hispanic Whites from Europe and Russia (including former USSR countries)
- Aim 6.2: Determine associations between region of birth and flu vaccine uptake among Arab Americans compared to US-born non-Hispanic Whites

Table 12.1 provides results from the chi square tests used to meet specific aim 6.1. Unweighted frequencies and weighted percentages are presented. There were statistically significant differences in flu vaccine uptake by region of birth, age, sex, and highest level of education among non-Hispanic Whites (all p's<.05). Adults who had a flu vaccine in the last 12 months were more likely to be US-born, ages 35-54 years, female, and have a bachelor's degree or higher level of education.

TABLE 12.1. SOCIODEMOGRAPHIC CHARACTERISTICS OF NON-HISPANIC WHITE ADULTS IN THE US BY FLU VACCINE UPTAKE IN THE LAST 12 MONTHS, NHIS 2018

	No N(%)	Yes N(%)	p-value
Region of birth among non-Hispanic Whites			.0476
United States	18,303 (95.9)	17,436 (96.9)	
Europe/Russia	513 (3.0)	406 (2.4)	
Arab/Middle East	182 (1.1)	90 (0.7)	
Age			<.0001
18-34 years	10,109 (39.4)	5,004 (25.1)	
35-54 years	13,270 (40.9)	8,508 (34.5)	
55-64 years	4,316 (11.7)	4,318 (16.7)	
65+ years	3,844 (8.0)	8,292 (23.7)	
Sex			<.0001
Male	15,127 (51.3)	10,774 (43.7)	
Female	16,412 (48.7)	15,348 (56.3)	
Highest level of education			<.0001
<High school	9,526 (32.9)	6,209 (26.5)	
High school diploma or GED	5,734 (20.0)	4,378 (17.8)	
Some college or Associate degree	7,387 (25.3)	6,011 (24.4)	
Bachelor’s degree or higher	6,829 (21.8)	8,020 (31.3)	

Table 12.2 provides results from logistic regression analyses used for meeting specific aim 6.2. Odds ratios (OR) and the corresponding 95% confidence intervals (CI) are presented. The reference group is US-born non-

Hispanic Whites. In the unadjusted model, foreign-born Arab Americans had 0.65 times lower odds of receiving a flu vaccine in the past 12 months compared to US-born Whites. However, because the confidence interval crosses the line of no effect at 1.00, the comparison is not statistically significant (95% CI=0.39, 1.10). This result differs from non-Hispanic Whites from Europe/Russia who had 0.78 times lower odds (95% CI=0.62, 0.97) of reporting a flu vaccine compared to US-born non-Hispanic Whites. For Arab Americans, results were statistically significantly different than US-born non-Hispanic Whites in the adjusted model (Model 2). After adjusting for age, sex, and education, foreign-born Arab Americans had 0.55 times lower odds (95% CI=0.32, 0.94) of reporting a flu vaccine in the past 12 months compared to US-born non-Hispanic Whites. The odds were lower than foreign-born non-Hispanic Whites from Europe/Russia (OR=0.55 Arab compared to OR=0.69 Europe/Russia) and results were statistically significantly lower than US-born non-Hispanic Whites. This result highlights the need to separate Arab American individuals from other non-Hispanic Whites so that their health outcomes are not masked under the White racial group.

TABLE 12.2. CRUDE AND MULTIVARIABLE LOGISTIC REGRESSION RESULTS, NHIS 2018

	Model 1 **Unadjusted** **OR (95% CI)**	**Model 2** **Adjusted for age, sex, and** **education** **OR (95% CI)**
Region of birth among non-Hispanic Whites		
United States	1.00	1.00
Europe/Russia	0.78 (0.62, 0.97)	0.69 (0.55, 0.86)
Arab/Middle East	0.65 (0.39, 1.10)	0.55 (0.32, 0.94)

12.4 MEDICAL EXPENDITURE PANEL SURVEY (MEPS) RECAP

Chapter 7 covered the background and details on how to obtain and analyze MEPS data. The objective of the MEPS case study was to explore whether adults who perceived their physician provided quality patient-provider communication (PPC) were more or less likely to receive an annual flu vaccine in comparison to those who did not receive quality PPC using household level in-person and self-administered questionnaire data. The following specific aims were examined using chi square tests and logistic regression analyses.

- Aim 7.1: Compare sociodemographic and health-related characteristics of adults by influenza vaccine uptake
- Aim 7.2: Determine association between adults' perceptions of PPC qualities and their likelihood of receiving an influenza vaccine before and after controlling for covariates

Two PPC qualities that were examined in this case study were whether instructions given to patients were easy for them to understand and whether the health care provider asked the patient to "teach-back," or describe how they will follow the instructions given to them. Table 12.3 provides results from the chi square tests used to meet specific aim 7.1. Unweighted frequencies and weighted percentages are presented. There were no statistically significant differences in flu vaccine uptake for either PPC quality evaluated. However, there were statistically significant differences in flu vaccine uptake by age and race/ethnicity (both p's<.0001). Adults who did not receive a flu vaccine in the last 12 months were more likely to be younger (ages 18-44 years). Non-Hispanic Black and Hispanic adults had higher estimates of not receiving a flu vaccine compared to non-Hispanic White adults and non-Hispanic adults of other or multiple races, inclusive of non-Hispanic Asians.

TABLE 12.3. PATIENT-PROVIDER COMMUNICATION QUALITIES AND SOCIODEMOGRAPHIC CHARACTERISTICS BY FLU VACCINE UPTAKE IN THE LAST 12 MONTHS IN THE US, MEPS 2015-2016

	Flu vaccine in last 12 months		p-value
	No N(%)	Yes N(%)	
Instructions provided easy to understand			.8679
Not Always	2,049 (31.1)	2,462 (30.9)	
Always	4,226 (68.9)	5,088 (69.1)	
Asked to describe how you will follow instructions			.4991
Not Always	4,105 (69.0)	5,003 (69.7)	
Always	2,156 (31.0)	2,529 (30.3)	
Age			<.0001
18-44 years	10,107 (55.6)	4,740 (33.2)	
45-64 years	5,937 (33.9)	4,765 (34.4)	
65+ years	1,731 (10.5)	4,231(32.4)	
Race/Ethnicity			<.0001
Hispanic	5,774 (18.6)	3,238 (12.5)	
Non-Hispanic White	6,738 (59.6)	6,611 (68.7)	
Non-Hispanic Black	3,569 (13.2)	2,338 (9.7)	
Non-Hispanic Other (including Asian/Multiple)	1,694 (8.6)	1,549 (9.1)	

Table 12.4 provides results from the logistic regression results used to meet specific aim 7.2. Odds ratios (OR) and the corresponding 95% confidence interval (CI) are presented. The reference group are those who did "not always" perceive their health care provider exhibited each

PPC quality. In the unadjusted models, there were no statistically significant differences between adults who perceived their health care provider always provided instructions that were easy to understand or asked them to describe how they will follow instructions with those who did not. Results remained statistically insignificant after adjusting for age and race/ethnicity for both PPC qualities. All logistic regression results were not statistically significant because the 95% confidence intervals cross the line of no effect at 1.00.

Table 12.4. Crude and multivariable logistic regression results, MEPS 2015-2016

	Model 1 **Unadjusted** OR (95% CI)	**Model 2** **Adjusted for age and race/ethnicity** OR (95% CI)
Instructions provided were easy to understand		
Not Always	1.00	1.00
Always	1.01 (0.92, 1.10)	1.06 (0.96, 1.16)
Asked to describe how you will follow instructions (teach-back)		
Not Always	1.00	1.00
Always	0.97 (0.88, 1.07)	1.03 (0.93, 1.14)

12.5 HEALTH INFORMATION NATIONAL TRENDS SURVEY (HINTS) RECAP

Chapter 8 covered the background and details on how to obtain and analyze HINTS data. The objective of the HINTS case study was to explore associations between e-mail PPC and colon cancer screening uptake using HINTS 5 Cycle 3 data. The following specific aims were examined using chi square tests and logistic regression analyses.

- Aim 8.1: Compare sociodemographic and health-related characteristics of adults who use e-mail to communicate with their health care provider
- Aim 8.2: Determine associations between e-mail PPC and adults' likelihood of receiving a colon cancer screening before and after controlling for covariates

Table 12.5 provides results from the chi square tests used to meet specific aim 8.1. Unweighted frequencies and weighted percentages are presented. There were no statistically significant differences in colon cancer screening uptake among adults who did and did not use e-mail to communicate with their health care provider. Furthermore, there were no statistically significant differences by gender. Older adults (ages 60-69 years and ages 70+ years) were more likely to receive a colon cancer screening than adults ages 50-59 years (p<.0001).

TABLE 12.5. E-MAIL PPC AND SOCIODEMOGRAPHIC CHARACTERISTICS BY COLON CANCER SCREENING UPTAKE IN THE LAST 12 MONTHS IN THE US, HINTS 5 CYCLE 3

	Colon cancer screening in last 12 months		p-value
	No N(%)	Yes N(%)	
Communicate with health care provider by e-mail			.2261
No	1,097 (55.2)	1,853 (57.8)	
Yes	886 (44.8)	1,342 (42.2)	
Gender			.3134
Male	767 (50.3)	1,316 (48.1)	
Female	1,097 (49.7)	1,676 (51.9)	
Age			<.0001
50-59 years	305 (69.1)	717 (38.0)	
60-69 years	163 (18.9)	1,077 (32.0)	
70+ years	131 (12.0)	1,166 (30.0)	

Table 12.6 provides results from the logistic regression analyses used to meet specific aim 8.2. Odds ratios (OR) and the corresponding 95% confidence intervals (CI) are presented. The reference group includes those who did not communicate with their health care provider by e-mail. In the unadjusted model, there were no statistically significant differences in colon cancer screening among adults who did and did not use e-mail to communicate with their health care provider. Results were not statistically significant because the 95% confidence intervals cross the line of no effect at 1.00. However, adults who used e-mail to communicate with their health

care provider had 2.15 times greater odds (95% CI=1.45, 3.19) of receiving a colon cancer screening after adjusting for gender and age.

Table 12.6. Crude and multivariable logistic regression results, HINTS 5 Cycle 3

	Model 1 **Unadjusted** OR (95% CI)	**Model 2** **Adjusted for gender and age** OR (95% CI)
Communicate with health care provider by e-mail		
No	1.00	1.00
Yes	0.90 (0.76, 1.07)	2.15 (1.45, 3.19)

12.6 BEHAVIORAL RISK FACTOR SURVEILLANCE SYSTEM (BRFSS) RECAP

Chapter 9 covered the background and details on how to obtain and analyze BRFSS data. The objective of the BRFSS case study was to explore whether differences in Alzheimer's disease and related dementia (ADRD) caregiving experiences among urban (metro) and rural (non-metro) adults in Texas were moderated by race and ethnicity. The differences obtained among metro and non-metro adults reported collectively were stratified by racial and ethnic groups. Data from the 2019 BRFSS were used to fulfil the aims. The following specific aims were examined using chi square tests and logistic regression analyses.

- Aim 9.1. Determine whether ADRD caregiving experiences differ across metro and non-metro geographic contexts among adults in Texas

- Aim 9.2. Determine whether the relationship between geographic context and ADRD caregiving experiences is moderated by the caregiver's race/ethnicity among metro and non-metro adults in Texas

Table 12.7 provides results from the chi square tests used to meet specific aim 9.1. Unweighted frequencies and weighted percentages are presented. There were no statistically significant differences in race/ethnicity, caregiver sex, caregiver relationship to care recipient or caregiver employment status by geographic context, among metro and non-metro ADRD caregivers (all p's>.05).

TABLE 12.7. SELECTED CHARACTERISTICS OF ADRD CAREGIVERS BY GEOGRAPHIC CONTEXT, BRFSS 2019

	Care Recipient Geographic Context		**p-value**
	Metro ***N (weighted %)***	**Non-Metro** ***N (weighted %)***	
Race/Ethnicity			0.2318
Non-Hispanic White (Majority group)	142 (49.2)	60 (69.42)	
Other Races (All minority groups)	79 (50.8)	14 (30.58)	
Caregiver Sex			0.2986
Male	65 (35.9)	22 (48.88)	
Female	156 (64.1)	52 (51.12)	
Caregiver Relationship to Care Recipient			0.1021
Mother, Father, In-laws	80 (39.66)	30 (50.84)	
Child	24 (11.37)	5 (1.11)	
Husband, wife, live in partner	33 (9.79)	16 (22.16)	
Other relative	40 (24.21)	11 (13.50)	
Non-relative/family friend	42 (14.96)	11 (12.39)	
Caregiver Employment Status			0.1039
Employed	98 (60.65)	27 (39.80)	
Retired	71 (18.84)	33 (41.22)	

Table 12.8 provides results from the logistic regression analyses used to meet specific aim 9.2 among non-Hispanic Whites (the majority group). Odds ratios (OR) and the corresponding 95% confidence intervals (CI) are presented. The reference group includes those whose

care recipient lives in a metro (urban) area. There were no statistically significant differences in household or personal caregiving experiences among non-Hispanic White ADRD caregivers from metro and non-metro geographic contexts. All logistic regression results were not statistically significant because the 95% confidence intervals cross the line of no effect at 1.00.

Table 12.8. Crude and adjusted logistic regression results for non-Hispanic White ADRD caregivers, BRFSS 2019 Texas

	Model 1 **Unadjusted** ***OR (95% CI)***	**Model 2** **Adjusted for sex, work, relationship** ***OR (95% CI)***
Household Caregiving Experiences		
Geographic Context		
Metro	1.00	1.00
Non-Metro	0.92 (0.26, 3.17)	0.72 (0.22, 2.34)
Personal Caregiving Experiences		
Geographic Context		
Metro	1.00	1.00
Non-Metro	2.26 (0.76, 6.71)	1.91 (0.59, 6.17)

Table 12.9 provides results from the logistic regression analyses used to meet specific aim 9.2 among caregivers from minority groups, including non-Hispanic Blacks, Hispanics, non-Hispanic Asians, and all others. Odds ratios (OR) and the corresponding 95% confidence intervals (CI) are presented. The reference group includes those whose care recipient lives in a metro (urban) area. There were no statistically significant differences in household or personal caregiving experiences among minority ADRD caregivers from metro and non-metro

geographic contexts. All logistic regression results were not statistically significant because the 95% confidence intervals cross the line of no effect at 1.00.

Table 12.9. Crude and adjusted logistic regression results for minority ADRD caregivers, BRFSS 2019 Texas

	Model 1 **Unadjusted** ***OR (95% CI)***	**Model 2** **Adjusted for sex, work, relationship** ***OR (95% CI)***
Household Caregiving Experiences		
Geographic Context		
Metro	1.00	1.00
Non-Metro	5.82 (0.73, 46.59)	3.05 (0.09, 103.33)
Personal Caregiving Experiences		
Geographic Context		
Metro	1.00	1.00
Non-Metro	9.80 (0.98, 97.73)	3.73 (0.32, 43.15)

12.7 NATIONAL HEALTH AND NUTRITION EXAMINATION SURVEY (NHANES) RECAP

Chapter 10 covered the background and details on how to obtain and analyze NHANES data. The objective of the NHANES case study was to determine racial and ethnic differences in sedentary behavior guideline adherence among US- and foreign-born Hispanics, non-Hispanic Whites, non-Hispanic Blacks, and non-Hispanic Asians. Data from the NHANES 2017-March 2020 pre-pandemic data files were used to fulfil the aims. The following specific aims were examined using chi square tests and logistic regression analyses.

- Aim 10.1: Compare the prevalence of adherence to 24-hour sedentary behavior guidelines in US adults by race, ethnicity, and nativity status
- Aim 10.2: Determine associations between race, ethnicity, and nativity and sedentary guideline adherence among racially and ethnically diverse foreign-born adults compared to their US-born counterparts

Table 12.10 provides results from the chi square tests used to meet specific aim 10.1. Unweighted frequencies and weighted percentages are presented. There were statistically significant differences in sedentary guideline adherence by nativity status among non-Hispanic Black and Hispanic adults. Fewer foreign-born non-Hispanic Black adults were adherent to sedentary behavior guidelines than US-born non-Hispanic Black adults (p=.0127). However, the pattern differed among Hispanic adults. More foreign-born Hispanic adults were adherent to sedentary behavior guidelines compared to US-born Hispanic adults (p=.0003). There were no statistically significant differences among non-Hispanic Whites or Asian adults by nativity status (both p’s>.05). There were no differences by age or gender. However, there was a statistically significant difference in sedentary guideline adherence by BMI (p<.0001).

TABLE 12.10. SELECTED CHARACTERISTICS OF ADULTS BY 24-HOUR MOVEMENT SEDENTARY GUIDELINE ADHERENCE, NHANES 2017-MARCH 2020 PRE-PANDEMIC DATA

	Sedentary Behavior Guideline		p-value
	Not adherent N (Weighted %)	Adherent N (Weighted %)	
Non-Hispanic White			.2880
US-born White	1,078 (95.7)	2,125 (94.4)	
Foreign-born White	44 (4.3)	96 (5.6)	
Non-Hispanic Black			.0127
US-born Black	665 (91.5)	1,639 (88.7)	
Foreign-born Black	56 (8.5)	169 (11.3)	
Hispanic			.0003
US-born Hispanic	205 (62.8)	601 (38.7)	
Foreign-born Hispanic	147 (37.2)	1149 (61.3)	
Non-Hispanic Asian			.7071
US-born Asian	56 (15.4)	94 (14.4)	
Foreign-born Asian	315 (84.6)	698 (85.6)	
Age			.0727
18-64 years	2,051 (81.3)	5,177 (78.9)	
65+ years	666 (18.7)	1,716 (21.1)	
Gender			.4735
Male	1,312 (47.6)	3,374 (48.7)	
Female	1,405 (52.4)	3,519 (51.3)	
BMI			<.0001
<25.00 BMI	524 (22.0)	1,558 (26.4)	
25.00-29.99 BMI	701 (28.2)	2,045 (34.0)	
30.00+ BMI	1,164 (49.8)	2,487 (39.6)	

Table 12.11 provides results from the logistic regression analyses used to meet specific aim 10.2. Odds ratios (OR) and the corresponding 95% confidence interval (CI) are

presented. There were no differences in sedentary guideline adherence by nativity status among non-Hispanic White or Asian adults. Foreign-born non-Hispanic Black had 1.37 times higher odds (95% CI=1.09, 1.71) of meeting sedentary behavior guidelines compared to US-born non-Hispanic Black adults in the unadjusted model. Results remained statistically significant after adjusting for age, gender, and BMI (OR=1.32; 95% CI=1.03, 1.69). Similar results were found for Hispanic adults. Hispanic adults had 2.67 times higher odds (95% CI=1.93, 3.70) of meeting sedentary behavior guidelines compared to US-born Hispanic adults in the unadjusted model. The odds increased to 2.85 (95% CI=2.02, 4.02) in the adjusted model.

Table 12.11. Unadjusted and adjusted logistic regression results, NHANES 2017-March 2020 pre-pandemic data

	Model 1 **Unadjusted** ***OR (95% CI)***	**Model 2** **Adjusted for age, gender, BMI** ***OR (95% CI)***
Race, Ethnicity, Nativity Status		
Non-Hispanic White		
US-born White	1.00	1.00
Foreign-born White	1.31 (0.78, 2.22)	1.35 (0.82, 2.21)
Non-Hispanic Black		
US-born Black	1.00	1.00
Foreign-born Black	1.37 (1.09, 1.71)	1.32 (1.03, 1.69)
Hispanic		
US-born Hispanic	1.00	1.00
Foreign-born Hispanic	2.67 (1.93, 3.70)	2.85 (2.02, 4.02)
Non-Hispanic Asian		
US-born Asian	1.00	1.00
Foreign-born Asian	1.08 (0.70, 1.67)	0.96 (0.59, 1.54)

12.8 DISSEMINATION RECAP

Chapter 11 covers the dissemination of research studies using secondary data from national health surveys. It includes details on how to disseminate results by abstracts, presentations, and original research manuscripts. Examples of poster presentations are provided as well as a thorough overview of writing each section of scientific manuscripts (Abstract, Introduction, Methods, Results, Discussion).

12.9 SUMMARY

The examples used in this textbook stem from previous studies and the current research laboratory focus of its primary author, Tiffany Kindratt, PhD, MPH. There is a wide range of research topics covered that may be of interest for undergraduate, graduate, and doctoral level students interested in national health surveys. Since all of the examples include SAS statistical software, future versions of this textbook and companion files will take into account other statistical software programs.

LINKS BY CHAPTER

Links by chapter are provided in the order that they appear in the chapter.

CHAPTER 1

- None

CHAPTER 2

- NHIS Website (https://www.cdc.gov/nchs/nhis/index.htm)
- MEPS Website (https://www.meps.ahrq.gov/mepsweb/index.jsp)
- HINTS Website (https://hints.cancer.gov/)
- BRFSS Website (https://www.cdc.gov/brfss/index.html)
- NHANES Website (https://www.cdc.gov/nchs/nhanes/index.htm)
- American Community Survey (ACS) (https://www.census.gov/programs-surveys/acs)
- Health and Retirement Study (HRS) (https://hrs.isr.umich.edu/welcome-health-and-

retirement-study)

- National Death Index (NDI) (https://www.cdc.gov/nchs/ndi/index.htm)
- National Health and Aging Trends Survey (NHATS) (https://www.cdc.gov/nchs/ndi/index.htm)
- National Longitudinal Study of Adolescent to Adult Health (Add Health) (https://addhealth.cpc.unc.edu/)
- National Study of Caregiving (NSOC) (https://www.nhats.org/researcher/nsoc)
- National Survey of Family Growth (NSFG) (https://www.cdc.gov/nchs/nsfg/index.htm)
- National Vital Statistics System (NVSS) (https://www.cdc.gov/nchs/nvss/index.htm)
- Youth Risk Factor Surveillance System (YRFSS) (https://www.cdc.gov/healthyyouth/data/yrbs/index.htm)

CHAPTER 3

- None

CHAPTER 4

- 2018 NHIS data release website (https://www.cdc.gov/nchs/nhis/nhis_2018_data_release.htm)
- Open ICPSR data repository (https://doi.org/10.3886/E172301V1)

CHAPTER 5

- 2018 NHIS data release website (https://www.cdc.gov/nchs/nhis/nhis_2018_data_release.htm)
- Open ICPSR data repository (https://doi.org/10.3886/E172301V1)

CHAPTER 6

- 2018 NHIS data release website (https://www.cdc.gov/nchs/nhis/nhis_2018_data_release.htm)
- Open ICPSR data repository (https://doi.org/10.3886/E172301V1)

CHAPTER 7

- MEPS website (https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_results.jsp?cboDataYear=All&cboDataTypeY=1%2CHousehold+Full+Year+File&buttonYearandDataType=Search&cboPufNumber=All&SearchTitle=Consolidated+Data)
- 2016 household consolidated data (https://meps.ahrq.gov/mepsweb/data_stats/download_data_files_detail.jsp?cboPufNumber=HC-192)
- Open ICPSR data repository (https://doi.org/10.3886/E172301V1)

CHAPTER 8

- HINTS website
- Open ICPSR data repository (https://doi.org/10.3886/E172301V1)

CHAPTER 9

- BRFSS website (https://www.cdc.gov/brfss/questionnaires/modules/category2019.htm)
- Open ICPSR data repository (https://doi.org/10.3886/E172301V1)
- 2019 BRFSS data website (https://www.cdc.gov/brfss/annual_data/annual_2019.html)

CHAPTER 10

- NHANES 2017- March 2020 pre-pandemic data website (https://wwwn.cdc.gov/nchs/nhanes/continuousnhanes/default.aspx?cycle=2017-2020)
- Open ICPSR data repository (https://doi.org/10.3886/E172301V1)

CHAPTER 11

- None

CHAPTER 12

- None

IMAGE CREDITS

All images are provided by the author and have the same license as the book.