National Addiction & HIV Data Archive Program
http://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/36231

ICPSR | INTER-UNIVERSITY CONSORTIUM FOR POLITICAL AND SOCIAL RESEARCH

ICPSR 36231

Population Assessment of Tobacco and Health (PATH) Study [United States] Restricted-Use Files

United States Department of Health and Human Services. National Institutes of Health. National Institute on Drug Abuse

United States Department of Health and Human Services. Food and Drug Administration. Center for Tobacco Products

User Guide

Inter-university Consortium for Political and Social Research
P.O. Box 1248
Ann Arbor, Michigan 48106
www.icpsr.umich.edu

Terms of Use

The terms of use for this study can be found at:
http://www.icpsr.umich.edu/icpsrweb/ICPSR/studies/36231/terms

Information about Copyrighted Content

Some instruments administered as part of this study may contain in whole or substantially in part contents from copyrighted instruments. Reproductions of the instruments are provided as documentation for the analysis of the data associated with this collection. Restrictions on "fair use" apply to all copyrighted content. More information about the reproduction of copyrighted works by educators and librarians is available from the United States Copyright Office.

NOTICE
WARNING CONCERNING COPYRIGHT RESTRICTIONS

The copyright law of the United States (Title 17, United States Code) governs the making of photocopies or other reproductions of copyrighted material. Under certain conditions specified in the law, libraries and archives are authorized to furnish a photocopy or other reproduction. One of these specified conditions is that the photocopy or reproduction is not to be "used for any purpose other than private study, scholarship, or research." If a user makes a request for, or later uses, a photocopy or reproduction for purposes in excess of "fair use," that user may be liable for copyright infringement.

PATH
Population Assessment of Tobacco and Health
A collaboration between the NIH and FDA

Contract #: HHSN271201100027C

PATH Study Restricted Use Files
User Guide

Date: March 28, 2016

Prepared by:
Westat
An Employee-Owned Research Corporation
1600 Research Boulevard
Rockville, Maryland 20850-3129
(301) 251-1500

Table of Contents

Chapter **Page**

Appendix

Table

1 Introduction

This User Guide describes the data files included in the Restricted Use Files (RUFs) for the Population Assessment of Tobacco and Health (PATH) Study. This document includes information about the sample design, data collection methods, response rates, and weighting and imputation methods; it also provides guidance for data users on the variance estimation of complex survey data. The RUFs contain extensive data from the adult interview, youth and parent interviews, and biospecimen collection. Any information from the data files that might be used to directly identify respondents has been removed following a disclosure review.

1.1 Background of PATH Study

The PATH Study is a nationally representative, longitudinal cohort study of tobacco use and how it affects the health of people in the United States. The PATH Study is the first large joint research effort on this topic by the National Institutes of Health (NIH) and Food and Drug Administration (FDA) since the 2009 Family Smoking Prevention and Tobacco Control Act authorized the FDA to regulate tobacco products. The PATH Study will provide an empirical evidence base for developing, implementing, and evaluating regulations governing tobacco products by measuring the behavioral and health effects associated with changes in such regulations.

A four-stage stratified area probability sample design was used in Wave 1 of the PATH Study, with a two-phase design for sampling the adult cohort at the final stage. At the first stage, a stratified sample of geographical primary sampling units (PSUs) was selected, in which a PSU was a county or group of counties. For the second stage, within each selected PSU, smaller geographical segments were formed and then a sample of these segments was drawn. At the third stage, the sampling frame consisted of the residential addresses located in these segments. The fourth stage selected adults and youth from the sampled households identified at these addresses, with varying sampling rates for adults by age, race, and tobacco-use status. Adults were sampled in two phases: Phase 1 sampling used information provided by one adult household member in the household screener and Phase 2 sampling used information that the sampled adult provided in the Phase 2 screener at the beginning of the adult interview. Parents do not constitute a separate sample. Parents who provide permission

for their child to complete a youth interview are asked to complete a brief parent interview about their youth selected for the study.

Wave 1 data collection took place from September 12, 2013, to December 14, 2014. Interviews were conducted with 32,320 adults (ages 18 and older), 13,651 youth (ages 12 to 17) and parents of 13,588 of those youth. Computer-assisted personal interviewing (CAPI) was used for the household screener and the parent interview and audio computer-assisted self-interviewing (ACASI) was used for the adult and youth interviews. Each of the 32,320 adults who completed the adult interview was asked to provide biospecimens. The first annual follow-up wave (referred to as Wave 2) began approximately seven weeks before the end of Wave 1; the second annual follow-up wave (Wave 3) began approximately two weeks before the end of Wave 2. Each of these subsequent waves will be conducted using data collection methods similar to those used in Wave 1.

1.2 Restricted Use Data Contents

The restricted use data consist of the following files that will be included for each wave:

- A data file containing adult questionnaire responses, derived variables, corresponding weight variables, and related external data, referred to as the "adult data file" in the sections below.
- A data file containing youth questionnaire responses, parent questionnaire responses about the youth, derived variables, corresponding weight variables, and related external data, referred to as the "youth/parent data file" in the sections below.
- A set of files containing data from biospecimen collections, including various laboratory result data, corresponding weight variables, and metadata.

Additionally, a master linkage file is included with each respondent's unique identifier (PERSONID) and indicators as to which files in the list above contain data pertaining to that respondent within a wave and across waves as subsequent waves of data become available.

The adult data file and the youth/parent data file have similar content. They both include the questionnaire items from the interview, some derived variables such as tobacco-use definitions, and other variables such as census geographic information, an indicator as to whether a sampling area is urban, and some census block-level demographic characteristics. The data files are accompanied by annotated instruments for each interview. The annotated instruments provide a crosswalk between

the variables and the corresponding questions asked in each interview. These documents can serve as references for data users to interpret variables for analyses. (See Chapter 5 for details about the adult data file and youth/parent data file.) The biospecimen files are described in separate documents and will be included with the other restricted use files once they become available.

Chapter 2 of this User Guide contains information about the multi-stage sample design. Chapter 3 describes the data collection methods and response rates. Chapter 4 discusses weighting procedures and imputation. Chapter 5 describes the structure of the adult data file and youth/parent data file, including the record identifier, variable naming convention, and variable values and labels.

2 Overview of Wave 1 Sample Design

The PATH Study is a nationally representative, longitudinal cohort study of 45,971 adults and youth ages 12 years and older in the United States. Wave 1 data collection was conducted from September 12, 2013 to December 14, 2014, using ACASI to collect information on tobacco-use patterns, risk perceptions, and attitudes towards current and newly emerging tobacco products. The PATH Study's design allows for the longitudinal assessment of patterns of use of a spectrum of tobacco products, including initiation, cessation, relapse, and transitions between products, as well as factors associated with use patterns.

2.1 Target Population

The PATH Study's Wave 1 target population is the civilian household population 12 years of age or older in the United States (all 50 States and the District of Columbia). College students were sampled through their permanent residence rather than at their dormitory. Active-duty members of the military (Army, Navy, Marines, Air Force, and Coast Guard) were excluded, as were all persons living in institutional and non-institutional group quarters other than college dormitories. Spouses and children of active-duty military living off post in the 50 States and D.C. were included.

2.2 Multi-Stage Sampling

The Wave 1 sample was selected using a four-stage, stratified probability sample design involving the selection of (1) 156 primary sampling units (PSUs) consisting of counties or groups of contiguous counties, (2) 6,049 second-stage sampling units (referred to as segments), (3) 166,088 mailing addresses, and (4) 76,539 sampled persons within households occupying dwelling units at sampled addresses. The sampling rates for adults varied by age, race, and tobacco-use status. Two-phase sampling was used for adult selection within households to correct for potential misreporting of the tobacco-use status of other adult members of the household by the household screener respondent. The sampling rates for the two phases were designed to achieve sufficiently large sample sizes for young adults (ages 18 to 24) and adult tobacco users of all ages. Up to two youths ages 12 to 17 were

sampled within each household. In addition, a "shadow sample" of youth ages 9 to 11 was selected for use as a refresher sample for the youth cohort in later waves.

2.2.1 PSU Sampling

At the first stage, a stratified sample of 156 PSUs was selected using probability proportional to size (PPS) sampling. The measure of size (MOS) was defined as a weighted sum of estimated PSU population counts of the subgroups that would be sampled within households at different rates, where the weights used to construct the MOS were proportional to the expected overall sampling rates to be applied for each subgroup. The use of this composite MOS was designed to (1) give relatively higher probabilities of selection to PSUs with higher proportions of the key subgroups of young adults, Blacks (African Americans),[1] and tobacco users; (2) to improve the chances that sufficient numbers of the various sampling subgroups would be included in the sample; and (3) to produce more balanced workloads for field interviewers.

Thirty-five PSUs were selected with certainty. These PSUs in effect serve as their own strata, and there is no PSU-to-PSU variability from the certainty PSUs. The remaining PSUs were grouped into strata based on known PSU characteristics that were related to the PATH Study variables of interest. This improves the precision of estimates of variables of interest (e.g., tobacco usage, perceptions, health, and possible changes over time in those characteristics) because PSUs within the same strata are in general more homogeneous than PSUs in the population as a whole. The variables used for stratification included census region and division, urban/rural designation, Core Based Statistical Area status and size, percent of adults age 25 and older with at least a bachelor's degree, percent of population with family income below 200 percent of the poverty level, percent of population who were Black, and percent of population who were Hispanic. These variables were available at the county level from the 2010 decennial census and the 2006-2010 5-year American Community Survey (ACS) data. Fifty-seven strata were formed using these variables. From each of these 57 strata, two, or occasionally three, PSUs were sampled systematically with probability proportional to the MOS, resulting in a total of 121 non-self-representing PSUs.

[1] Blacks were oversampled as a proxy for menthol cigarette users. For this purpose, Black was defined as Black alone or in combination with other race(s), whether Hispanic or non-Hispanic.

2.2.2 Segment Sampling

Segments were formed within the sampled PSUs, with the goal of minimizing travel time for the field staff while yielding the desired sample sizes for the segments. The segments were based on census-defined blocks and had, in general, a minimum of 100 occupied housing units.

Each segment formed was assigned a MOS to be used in drawing a PPS sample of segments. The segment composite MOS was computed in the same manner as the PSU composite MOS, except that the expected counts of adults in each of the eight age/race/tobacco-use subgroups were computed at the segment level, rather than at the PSU level.

Implicit stratification was achieved by using socio-demographic data from the 2010 census to group and sort the segments on the frame prior to drawing a PPS systematic sample from each PSU. The demographic factors used for the implicit stratification were percent Black,[2] percent Hispanic, and percent of occupied housing units that were owner-occupied. Segments that were similar to each other on these three factors were grouped together in the same stratum by using a clustering procedure. The implicit stratification ensured that a representative sample of segments was taken from each PSU, and helped reduce the variability within strata, so that the resulting sample of segments would produce estimates with smaller variance (Golder and Yeomans, 1973; Judkins and Singh, 1981). A systematic PPS sample of about 40 segments was drawn within each noncertainty PSU,[3] with more segments drawn in the larger certainty PSUs, for a total of 6,049 segments.

2.2.3 Address Sampling

At the third stage of sampling, addresses within each segment were ordered by census block, and a sample was selected by systematic sampling. The goal for address selection was to minimize the variation in weights across the first three stages of sampling and to produce a relatively even workload for field interviewers in each segment. The sampling rate within each segment was determined so that each address selected for the sample would have approximately the same unconditional probability of selection over all three stages of sampling.[4] Due to the form of the

[2] Black was defined as Black alone or in combination with other race(s), whether Hispanic or non-Hispanic.

[3] A few small PSUs had fewer than 40 segments; in these, all segments were selected with certainty.

[4] The sampling rate was adjusted for some segments to yield an expected number of addresses between 10 and 40. In addition, in some cases, dwelling units found to belong to multi-unit structures that received their mail at a single mail "drop" were subsampled.

composite MOS used for PPS sampling at the first two stages of the design, the allocated number of sampled addresses varied by segment. However, the sampling scheme was intended to produce approximately equal segment workloads in terms of the number of adult interviews conducted. A systematic sample of addresses with the desired sampling rate was chosen from each segment.

2.2.4 Within-Household Sampling

The fourth stage of sampling selected persons from the sampled households. During the household screener, one adult household member (referred to as the screener respondent) was asked to list members of the household and provide demographic, and, for adults, tobacco-use information about each for use in sampling three main groups of interest:

- Adults (up to two adults per household were sampled);
- Children ages 12 to 17 (referred to as "youth," generally up to two per household were sampled); and
- Children ages 9 to 11 (referred to as "shadow youth," generally up to two per household were sampled) to be enrolled in the youth cohort on reaching 12 years of age.

The sections below provide more detail regarding the selection of adults, youth, and shadow youth.

Adults

The sampling procedure for selecting adults within a household had two phases. Phase 1 sampling depended on the information provided by the household screener respondent. For the Phase 1 sampling, adults in the household were classified into one of eight subgroups defined by the cross-classification of age (ages 18 to 24, ages 25 and older), race (Black,[5] all others), and tobacco usage[6] (user, not user), as reported by the screener respondent. Adults were sampled with predetermined sampling rates to participate in Phase 2 of the screening process that led to the final sampling for the

[5] Black was defined as Black alone or in combination with other race(s), whether Hispanic or non-Hispanic.

[6] Because of the PATH Study's interest in persons who are experimenting with tobacco products or are likely to become users in the future, a broad definition of tobacco use was implemented when classifying adults into Phase 1 sampling domains. This definition classified an adult as a tobacco user if it was reported that he or she smoked a cigarette, cigar, or pipe, or used smokeless tobacco every day or some days; and/or had ever used an e-cigarette, snus, or dissolvable tobacco, or had ever smoked tobacco in a hookah.

adult interview, subject to the constraint that at most two adults were sampled from each household. Phase 2 sampling was included to address classification errors in the responses of household screener respondents, in particular the misclassification of a sampled adult as a nonuser of tobacco when the self-report would indicate the person was a user. The Phase 2 sampling was based on the sampled individual's self-reported information and thus considered more accurate in general.[7] The sampling rates for the two phases were designed to achieve predetermined minimal sample sizes for key subgroups, such as young adults (ages 18 to 24) and adult tobacco users of all ages.

At Phase 1, the sampling rates for nonusers were kept within reasonable bounds, compared to the rates for users, in order to ensure that the weights of any adults sampled at Phase 1 as nonusers who then reported themselves at Phase 2 to be users would be similar to the weights of those who were correctly classified as users at Phase 1. Misclassification in the other direction—with the screener respondent reporting the adult as a user when the person self-reported as a nonuser—was handled by deselecting some members of this group so that those retained would have sampling rates similar to those of other nonusers.

The population proportions of adults within the eight age/race/tobacco-use subgroups were unknown at the initial design stage. The within-household adult sampling rates were adjusted during the course of Wave 1 data collection as more accurate information accrued in order to achieve the desired sample sizes in the eight domains.

The two-phase sampling procedure and disproportionate sampling of younger adults, Black adults, and tobacco users have two main effects on the sampling errors. First, the procedure resulted in larger sample sizes in the oversampled subgroups than would occur with equal probability sampling, which reduces the sampling errors for estimates calculated for those subgroups. However, the increased precision for those subgroups came at the cost of increased weight variation for the adult sample as a whole. The average selection probabilities for subgroups ranged from approximately 0.10 (for non-Black nonusers ages 25 and older) to 0.80 (for Black users ages 18 to 24), which leads to weight factors between 10 and 1.25. Estimates calculated for all adults have larger standard errors under this design than they would if all adults had been sampled at the same rate.

[7] Because of the PATH Study's interest in persons who are experimenting with tobacco products or are likely to become users in the future, a "wide net" definition of tobacco use was implemented when classifying adults into Phase 2 sampling domains. This definition classified an adult as a tobacco user if he or she had smoked a cigarette, cigar, or pipe, or had used smokeless tobacco in the past 30 days; and/or had ever used an e-cigarette, snus, or dissolvable tobacco, or had ever smoked tobacco in a hookah.

Youth and Shadow Youth

The PATH Study youth sample has two components: selection of youth ages 12 to 17 for the Wave 1 and follow-up interviews, and selection of a "shadow sample" of youth ages 9 to 11 for inclusion in future waves. Youth in the shadow sample were selected at Wave 1 for the purpose of replenishing the 12- to 17-year-old youth sample in later waves; they were not interviewed at Wave 1. The sampling of youth within sampled households was independent of adult sampling and did not involve any oversampling by race, ethnicity, gender, or tobacco use. The number of persons in the age range 12-17 was tallied for the household. If only one or two children in that age range were in the household, those children were included in the youth sample. If there were more than two children in that age range, two were randomly selected for the youth sample.

The shadow youth selection procedure was identical to that for the youth ages 12 to 17 and was carried out independently of the youth and adult sampling within sampled households. The number of 9- to 11-year-old children in the household was tallied. If only one or two children in that age range were in the household, those children were included in the shadow sample. If there were more than two children in that age range, two were randomly selected for the shadow sample.

Given a special analytic interest in multiple-birth youth in the youth cohort, the shadow and youth cohort sampling procedures were modified when households containing multiple-birth youths were encountered so that the multiple-birth youths would have relatively higher probabilities of selection. This resulted in some households with more than two children selected for the youth and/or shadow youth samples.

3 Data Collection and Response Rates

3.1 Overview

The PATH Study collects interview data through in-person CAPI and ACASI methods. Through computer-assisted interviews, the PATH Study collects baseline and follow-up information on tobacco-use patterns, trends in risk perceptions and attitudes regarding harmful constituents, new and emerging tobacco products, and tobacco initiation, cessation, and relapse behaviors among youth ages 12 to 17 and adults ages 18 and older. Parents who provide permission for their child to complete a youth interview are asked to complete a brief parent interview that contains questions about parental supervision, school performance, and tobacco use by youth. The Wave 1 sample size is 32,320 adults (ages 18 and older) and 13,651 youths (ages 12-17).

The PATH Study also collects biospecimens from consenting adult respondents to assess markers of nicotine exposure, to detect and compare intermediate endpoints and incident health outcomes associated with the use of tobacco products and related disease processes, and to validate self-reported behavioral and health data. Respondents completing the adult interview for the first time are asked to provide samples of urine and blood (and buccal cells, until this collection was discontinued in May 2014); this includes new or "aged-up" adults (youths who have turned 18 years old and agree to continue in the PATH Study as adults) in waves subsequent to Wave 1. Additionally, a subsample of continuing adults who provided urine in a previous wave will be asked to continue providing a urine specimen in subsequent waves, with the goal of collecting a specimen from the same individuals over several waves.

Data collection involves four main components: (1) a CAPI household screening instrument, (2) ACASI instruments (separate instruments for youth and adults), (3) a CAPI parent instrument, and (4) collection of urine, blood, and through May 2014, buccal cell specimens from consenting adults.

Before the start of data collection, the PATH Study received approval from the Westat Institutional Review Board. The Study also obtained a Certificate of Confidentiality from the National Institutes of Health.

3.2 Wave 1 Data Collection

3.2.1 Advance Mail and Prepaid Incentives

Advance letters and brochures were mailed to all sampled addresses one to two weeks prior to the field interviewer's Wave 1 visit to inform the target respondents of the PATH Study sponsor, the nature and uses of the data collected, legal authorities, the voluntary nature of participation, and protection of the information. The advance letter contained a $2 bill as an attention-getter. Before administration of the household screener, the field interviewer asked if the household received the introductory letter and brochure. Those respondents who did not recall receiving or reading them were provided with copies, and were given time to read the information and ask any questions.

3.2.2 Household Screener

Up to two adults and two youths per eligible household were randomly selected using a CAPI screening instrument. One or more **additional** youth could also be selected among households with multiple-birth youths (e.g., twins). The screener respondent was an adult household member age 18 or older; he/she provided oral consent for the screener. The screener used a full household enumeration process to collect information on age for each reported household member, as well as information on race, active military service status, ability to speak in English or Spanish, and tobacco use for each adult household member. The relationship of all household members to the screener respondent was also collected. In addition to household enumeration information, the household screener respondent's and each sampled person's telephone numbers were collected to allow re-contact of the household for quality control purposes or to set appointments for interviews if any of the sampled persons or parents were unavailable at the time of the screening. The household's mailing address was also collected for purposes of re-contacting the sampled person(s).

The sampling algorithm for selecting up to two adults and two youths (except in the case of multiple-birth youths) per household was programmed within the CAPI screener software and tested extensively before data collection began.

3.2.3 Interview

The data collection procedures differed for adults and youth. For adults, some questions were asked at the beginning of the interview to serve the purpose of Phase 2 sampling, as discussed in Section 2.2.4.

Adult Data Collection

Among other purposes, the household screener (which is also the Phase 1 screener for adults) collected a minimum amount of high-level information about each adult's tobacco usage in order to classify him/her sufficiently for potential selection based on the PATH Study sampling algorithm. The household screener obtained tobacco-use information about all adults from the household screener respondent, and this information could be inaccurate. To obtain more complete and accurate information from an adult sampled through the household screener, a Phase 2 screener was administered to the sampled adult directly to ask a more extensive panel of questions about tobacco usage. The Phase 2 questions were also asked in a private setting using ACASI rather than CAPI.

Following the administration of the Phase 1 screener, if the sampled adult was available and had an adequate amount of time to complete the interview, the field interviewer (1) obtained informed consent; (2) administered the Phase 2 screener and adult interview, which included gathering additional contact information about the adult; (3) obtained consent for the biospecimen collection; (4) collected the urine sample (and buccal cell biospecimens through May 2014); (5) arranged a follow-up appointment for a phlebotomist to collect a blood sample; and (6) at the completion of the first home visit, paid the incentive to the respondent. If a sampled adult was unavailable or unable to complete the interview at that time, the field interviewer attempted to schedule an appointment for a return visit or, at a minimum, determined the best time for a return visit.

To begin the interview, the field interviewer provided a brief automated tutorial to the adult on using ACASI. The first part of the interview was the individual (Phase 2) screener, which may confirm or contradict the information provided in the Phase 1 (household) screener by the household screener respondent. Depending on the individual's self-reports (e.g., on tobacco usage), the computerized sampling algorithm may prompt the field interviewer to de-select the respondent and ask him or her not to complete the remainder of the extended interview. Throughout the interview, the field interviewer provided aid to the sampled person if needed on the use of ACASI or related questions. At the end of the extended interview, the field interviewer collected additional

contact information for that person and asked the respondent to consent to providing biospecimens.

The sampled adult who completed the interview or was de-selected based on his/her responses to the Phase 2 screener received $35 (the adult interview incentive) in appreciation for his/her time for completing the interview, as well as a thank-you letter. Sampled adults who initially declined to participate or were difficult to contact were sent a follow-up refusal conversion letter approximately 3 weeks later

Youth Data Collection

Following the administration of the household screener, if the parent or guardian of the selected youth was available and had time, the field interviewer (1) requested parent permission for the youth to participate, (2) requested consent for the short parent interview, and (3) administered the CAPI parent interview, which included collecting additional contact information from the parent. If a parent of a sampled youth was unavailable or unable to participate at that time, the field interviewer attempted to schedule an appointment for a return visit or, at a minimum, determined the best time for a return visit. The youth interview was not conducted until parental permission was obtained. The parent who completed a parent interview for the youth received $10 as a token of appreciation for completing the interview.

For a selected youth with parental permission, if the youth was available and had an adequate amount of time to complete the interview, the field interviewer requested youth assent. If a sampled youth was unavailable or unable to complete the interview at that time, the field interviewer attempted to schedule an appointment for a return visit or, at a minimum, determined the best time for a return visit.

After obtaining assent from the selected youth, the field interviewer provided a brief automated tutorial on using ACASI and launched the ACASI interview. Throughout the interview, the field interviewer provided aid to the sampled youth in using the ACASI instrument if necessary.

Sampled youth who completed the extended interview received $25 (the youth interview incentive) as a token of appreciation for completing the interview. The parent of the youth respondent received a thank-you letter. The parents of sampled youth who were difficult to contact or initially

declined permission for their youth to participate were sent a refusal conversion letter approximately 3 weeks later.

3.2.4 Biospecimen Collection

The field interviewer asked each adult interview respondent to consent to provide biospecimens as part of the PATH Study. However, providing biospecimens was voluntary and not a condition of participation. Completion of the Wave 1 adult interview was required from all respondents in order for them to provide biospecimens.

Buccal Cells and Urine Collection

For adults who consented to provide buccal cells and urine, the field interviewer collected the specimens following the completion of the interview. The field interviewer provided written and oral instructions to the respondent for collection of buccal cells and the urine specimen. The field interviewer packed the specimen(s) and shipped the package to the PATH Study biorepository.

If a sampled adult was unavailable or unable to continue with specimen collection immediately following completion of the interview, the field interviewer attempted to schedule an appointment for a return visit or, at minimum, determined the best time for a return visit.

The sampled adult who provided biospecimens during a field interviewer's visit received $25 as a token of appreciation for participating in the buccal cell and/or urine sample component of the Study.
Midway through Wave 1 in May 2014, buccal cell biospecimen collection was stopped. However, at some point data may be released from buccal cell laboratory analysis.

Blood Collection

For adults who consented to provide blood, the field interviewer scheduled the appointment for the visit by a phlebotomist to obtain the blood specimen. Upon visiting the respondent's home, the phlebotomist administered blood suitability exclusion questions for blood collection (CAPI instrument) and asked the respondent to answer items about his/her recent use of tobacco products

(ACASI instrument). The phlebotomist then collected the blood specimen, and packed and shipped it to the PATH Study biorepository.

The sampled adult who provided a blood specimen during this later home visit received $25 as a token of appreciation for participating in the blood sample component of the Study.

3.3 Wave 1 Response Rates

This section summarizes the Wave 1 response rates for the household screener, the adult interview, and the youth interview. The response rate calculations were based on the formula provided by the Office of Management and Budget in its "Standards and Guidelines for Statistical Surveys" (2006). This formula calls for calculating unit response rates as the ratio of the number of completed cases to the number of in-scope sample cases.[8]

The household screener unit response rate, denoted as RRU_{HH}, was calculated using equation (3.3.1):

$$RRU_{HH} = C_{HH}/(C_{HH} + R_{HH} + NC_{HH} + O_{HH} + e * U_{HH}) \quad (3.3.1)$$

where

C_{HH} = number of completed cases;
R_{HH} = number of refused cases;
NC_{HH} = number of noncontacted sample units known to be eligible;
O_{HH} = number of eligible sample units not responding for reasons other than refusal;
U_{HH} = number of sample units of unknown eligibility, not completed; and
e = estimated proportion of sample units of unknown eligibility that are eligible.

Wave 1 response rates for adults and youth depended on completion of the Phase 1 household screener.[9] The adult interview unit response rate (conditioning on the completion of Phase 1 household screener), denoted as RRU_A, was calculated as the product of (1) the Phase 2 screener response rate, and (2) the proportion of adults who completed the adult interview among those who

[8] The PATH Study sample for Wave 1 did not have any partial completes.

[9] At the adult and youth level, all cases involved in the response rate calculations were eligible.

completed the Phase 2 screener and were selected for the adult interview, as shown in equation (3.3.2):

$$RRU_A = (C_{P2}/(C_{P2} + NR_{P2})) \times (C_A/(C_A + NR_A)) \quad (3.3.2)$$

where

C_{P2} = number of completed cases for the Phase 2 screener;
NR_{P2} = number of finalized nonresponse cases for the Phase 2 screener;
C_A = number of completed cases for the adult interview; and
NR_A = number of finalized nonresponse cases for the adult interview.

The youth interview unit response rate (conditioning on the completion of household screener), denoted as RRU_Y, is calculated using equation (3.3.3):

$$RRU_Y = C_Y/(C_Y + NR_Y) \quad (3.3.3)$$

where

C_Y = number of completed cases for the youth interview; and
NR_Y = number of finalized nonresponse cases for the youth interview.

Both unweighted and weighted unit response rates were calculated. For the unweighted response rates, the numbers of cases used in the calculations were the actual case counts. For the weighted response rates, the numbers of cases used in the calculation were the sums of the basic design weights. At household level, the basic design weight is the household IPS weight described in equation 4.1.1 in Chapter 4. At the adult level and youth level, the basic design weight is the product of the household IPS weight and the inverse of the within-household probability of selection. The unweighted response rate measures the success of field operations in obtaining responses from the selected sample. The weighted response rate estimates the proportion of the population that would have responded if the entire population was asked to participate in the PATH Study, and provides a measure of the anticipated impact of nonresponse on the quality of estimates calculated using the weights.

Tables 1-3 show the unweighted and weighted response rates for the household screener, adult interview, and youth interview. The weighted response rates are 54.0 percent, 74.0 percent, and 78.4

percent for household screener, adult interview, and youth interview, respectively. Response propensities differed across demographic groups defined by age, gender, race/ethnicity, education, and income level. Differential response rates across demographic groups were compensated for by using nonresponse adjustments in the weighting methods, as described in Chapter 4.

Table 1. Household screener response rate

	A Completed (n)	B Finalized nonresponse[a] (n)	C Unknown eligible estimated to be eligible[b] (n)	Unweighted response rate[c] (%)	Weighted response rate (%)
Household screener	79,198	62,332	4,760	54.1	54.0

[a] Finalized nonresponse includes refused cases, uncontacted cases known to be eligible, and other eligible cases not responding for reasons other than refusal.

[b] Product of cases of unknown eligibility and estimated proportion of cases of unknown eligibility that are eligible.

[c] Response rate = A/(A+B+C).

Table 2. Adult interview response rate

	Phase 2 Screener		Adult Interview			
	A P2 screener, completed (n)	B P2 screener, finalized nonresponse[a] (n)	C Adult interview, completed (n)	D Adult interview, finalized nonresponse[a] (n)	Unweighted response rate[b] (%)	Weighted response rate (%)
Adult interview	44,303	14,785	32,320	80	74.8	74.0

[a] Finalized nonresponse includes refused cases and other eligible cases not responding for reasons other than refusal.

[b] Response rate = (A/(A+B))*(C/(C+D)).

Table 3. Youth interview response rate

	A Completed (n)	B Finalized nonresponse[a] (n)	Unweighted response rate[b] (%)	Weighted response rate (%)
Youth interview	13,651	3,800	78.2	78.4

[a] Finalized nonresponse includes refused cases and other eligible cases not responding for reasons other than refusal.

[b] Response rate = A/(A+B).

4 Weights and Imputation

Analysis of data from complex sample designs, such as the PATH Study design, requires the use of weights to compensate for variable probabilities of selection, differential nonresponse rates, and possible deficiencies in the sampling frame (e.g., undercoverage of certain population groups). It is also necessary to implement variance estimation procedures that appropriately account for sampling design factors (such as the stratification and sampling of PSUs and area segments, and the use of oversampling) and nonresponse adjustment factors. Section 4.1 describes the procedures used to calculate the adult and youth weights. Section 4.2 describes methods that should be used to estimate variances, with sample code given for the SAS software package. The weights and variance estimation methods will give correct inferences for analyses of the PATH Study adult data file and youth/parent data file. Section 4.3 provides a description of the imputed variables included in the adult data file and youth/parent data file.

4.1 Wave 1 Adult and Youth Weights

The sections below describe the computation of the household weights and the adult and youth weights included in the adult data file and youth/parent data file, respectively.

Household Weights

The initial household weights, denoted as HHIPSWT, were calculated for all households sampled (responding households and nonresponding households) as the inverse of the probability of selection (IPS) as shown in equation 4.1.1.

$$HHIPSWT_{ijk} = \frac{1}{P_{ijk}}, \quad (4.1.1)$$

where P_{ijk} is the probability that household k in segment j of PSU i was selected to be in the sample. However, some sampled addresses could not be located/accessed, others were found to be ineligible (e.g., vacant lots and group quarters), and some eligible households did not complete the

household screener. Adjustments were therefore made to the IPS weights of responding households to compensate for the estimated number of nonresponding households that were eligible for the PATH Study based on all the addresses in the sample for which eligibility status was determined. This eligibility adjustment was done separately for each census region.

Further adjustments were made within weighting classes based on information available for both responding and nonresponding households, including census 2010 and ACS 5-year (2009-2013) data pertaining to the segments, tracts, and blocks in which they were located. Census 2010 data were used to calculate the percent of occupied housing units that were owner-occupied, the percent of the population who were Black,[10] the percent of the population who were Asian,[11] the percent of the population who were Hispanic, and the percent of the population 25 years and older in the census block containing the address. The 5-year ACS data were used to estimate the median monthly housing unit costs in the census tract containing the address. Census region, the urbanicity of the PSU, and the urbanicity of the segment were also used when forming the weighting classes.

Then, within a weighting class, the IPS weights for the responding households were inflated proportionately so that they produced the same sum as the sum of the IPS weights of the responding and nonresponding households combined. The nonresponse-adjusted household weight (denoted as HHNRWT) for responding household k of PSU i and segment j is calculated as shown in equation 4.1.2.

$$HHNRWT_{ijk} = HHIPSWT_{ijk} \times \frac{\text{sum of HHIPSWT for eligible sampled households in weighting class}}{\text{sum of HHIPSWT for responding households in weighting class}}. \quad (4.1.2)$$

The nonresponse-adjusted weights were raked to the 1-year 2013 ACS household counts by census region and household composition. Household composition was defined by the number of non-adult persons in the household (0, 1, or 2+) and the number of adult household members (1, 2, 3+). For raking purposes, the household composition was imputed for households missing this information using logical imputation.[12] The final raked household weight is calculated as shown in equation 4.1.3.

$$HHRKWT_{ijk} = HHNRWT_{ijk} \times HHR_{ijk}, \quad (4.1.3)$$

[10] Black was defined as Black alone or in combination with other race(s), whether Hispanic or non-Hispanic.

[11] Asian was defined as Asian alone.

[12] See Lohr (2010) for a brief description of raking and imputation methods.

where HHR_{ijk} is the household raking adjustment factor for household k of PSU i and segment j. The household weights are not included on the data files.

Adult Weights

The raked household-level weight is used as the foundation for calculating the adult weight. The adult base weight, denoted as AP1BWT, was computed as the product of the final household weight HHRKWT and the reciprocal of the within-household probability of selection for adult l within household k of PSU i and segment j, as shown in equation 4.1.4.

$$AP1BWT_{ijkl} = HHRKWT_{ijk} \times \frac{1}{\text{Probability adult } l \text{ selected at Phase 1 from household } (ijk)}. \quad (4.1.4)$$

The final weights for adults were computed in three steps.

First, a nonresponse adjustment was performed to account for nonresponse to the adult interview using a combination of census 2010 and 5-year (2009-2013) ACS data (used for the household nonresponse adjustment) and person-level data collected during the household screening interview. Weighting classes were formed based on census region and urbanicity of the PSU and segment; the percent of occupied housing units that were owner-occupied, the percent of the population who were Black, the percent of the population who were Asian, the percent of the population who were Hispanic, and the percent of the population 25 years and older in the census block containing the address; the median housing unit costs in the census tract containing the address; the age and gender of the household screener respondent, and the number of adults in the household (capped at 5); and age, race/ethnicity, gender, and tobacco-use status of the adult.[13]

The resulting adult weight (denoted as AP1NRWT), adjusted for nonresponse between Phases 1 and 2 of the adult sampling procedure, for respondents to the Phase 2 Screener, is calculated as shown in equation 4.1.5.

$$AP1NRWT_{ijkl} = AP1BWT_{ijkl} \times \frac{\text{sum of AP1BWT for adults sampled at Phase 1 in weighting class}}{\text{sum of AP1BWT for adults responding to Phase 2 Screener in weighting class}}. \quad (4.1.5)$$

[13] Block-level population percentages were created from census 2010 data. Tract-level median housing unit costs were extracted from the 5-year (2009-2013) American Community Survey (ACS) data.

Second, the probability of selection at Phase 2 was used to find the Phase 2 weight, denoted as AP2WT, as shown in equation 4.1.6.

$$AP2WT_{ijkl} = AP1NRWT_{ijkl} \times \frac{1}{\text{Probability adult } l \text{ from household } (ijk) \text{ selected at Phase 2}}. \quad (4.1.6)$$

Finally, raking and trimming were performed in an iterative process. The Phase 2 adult weights were raked to independent population totals based on data from the 1-year 2013 ACS. The raking was done using combinations of census region, age, race/ethnicity, gender, and educational attainment. These variables were imputed if they were missing. (See Section 4.3 for more information about this imputation.)

However, the raking algorithm did not place any restrictions on the highest and lowest values of the raked weights, and a few of the raked weights became extremely large in the process of matching the population totals from the ACS. To reduce any extreme weights generated in the raking process, a trimming step was performed to bring any extreme weights down to the median weight plus four times the interquartile range[14] within groups defined by the eight age, race, and tobacco-use status categories used to select the adults at Phase 2. After trimming, the weights no longer matched the control constraints, so the raking process was repeated. The trimming and raking steps were iterated until the resulting weights summed to the 2013 ACS totals for the raking dimensions and were within the bounds defined by the interquartile range criterion.

After the iterative raking and trimming process, the final adult weight, denoted as ARKWT, was calculated as shown in equation 4.1.7.

$$ARKWT_{ijkl} = AP2WT_{ijkl} \times ART_{ijkl}\,, \quad (4.1.7)$$

where ART_{ijkl} is the combined raking and trimming adjustment for adult l within household k of PSU i and segment j. These final weights can be found in the adult data file in the variable named R01_A_PWGT, indicating that these are person-level, adult weights for Wave 1 of the PATH Study.

[14] The weight trimming procedure reduces the mean squared error by reducing the variation among the weights. See Battaglia et al. (2013) and Chowdhury et al. (2007).

Youth Weights

Similar to the adult weighting, the raked household-level weight is also used as the foundation for calculating the youth weight. The youth base weight, denoted as YBWT, was computed as the product of the final household weight HHRKWT and the reciprocal of the within-household probability of selection for youth l within household k of PSU i and segment j, as shown in the equation 4.1.8.

$$YBWT_{ijkl} = HHRKWT_{ijk} \times \frac{1}{\text{Probability youth } l \text{ selected from household } (ijk)}. \quad (4.1.8)$$

Similarly to the adjustment for household screener nonresponse, a nonresponse adjustment was performed to account for nonresponse to the youth interview using a combination of census 2010 and 2009-2013 ACS data (used for the household nonresponse adjustment) and person-level data collected during the household screening interview. Weighting classes were formed based on census region and urbanicity of the PSU and segment; the percent of occupied housing units that were owner-occupied, the percent of the population who were Black, the percent of the population who were Asian, the percent of the population who were Hispanic, and the percent of the population 25 years and older in the census block containing the address; the median housing unit costs in the census tract containing the address; the age and gender of the household screener respondent; the number of adults in the household (capped at 5); and the age, gender, and race/ethnicity of the selected youth.[15]

Then, within a weighting class, the base weights (YBWT) for the responding youth were inflated proportionately so that they produced the same sum as the sum of the base weights of the responding and nonresponding youth combined. The nonresponse-adjusted weight for responding youth is

$$YNRWT_{ijkl} = YBWT_{ijkl} \times \frac{\text{sum of YBWT for sampled youth in weighting class}}{\text{sum of YBWT for responding youth in weighting class}}. \quad (4.1.9)$$

For youth, the nonresponse adjusted weights (YNRWT) were raked to population totals from the 1-year 2013 ACS using the same iterative process as used in the adult weighting described above. The raking was done using census region, age, race/ethnicity, and gender as raking variables. These variables were imputed if they were missing. (See Section 4.3 for more information about this

[15] Block-level population percentages were created from census 2010 data. Tract-level median housing unit costs were extracted from the 5-year (2009-2013) American Community Survey (ACS) data.

imputation.) The trimming threshold was set to the median weight plus four times the interquartile range within groups defined by whether or not the youth was selected with certainty.

After the iterative raking and trimming process, the final youth weight, denoted as YRKWT, was calculated as shown in equation 4.1.10.

$$YRKWT_{ijkl} = YNRWT_{ijkl} \times YRT_{ijkl}\,, \tag{4.1.10}$$

where YRT_{ijkl} is the combined trimming and raking adjustment for youth l within household k of PSU i and segment j. These final weights can be found on the youth/parent data file in the variable named R01_Y_PWGT, indicating that these are person-level, youth weights for Wave 1 of the PATH Study.

4.2 Variance Estimation

The adult data file and youth/parent data file include replicate weights that may be used to estimate variances. The replication variance estimation approach is the preferred method for estimating variances from PATH Study data.

Replication variance methods are increasingly used to provide analysts with a method for calculating standard errors of statistics. They produce consistent estimators of the variance for statistics that are smooth functions of estimated totals (Krewski and Rao, 1981); these include most commonly used statistics such as means, ratios, linear/logistic/Poisson regression coefficients, correlation coefficients, and many measures of association for categorical data. In most complex designs, such as the multistage sample design used in the PATH Study, the variance is estimated by assuming that the first-stage sampling is performed with replacement (Wolter, 2007).

The basic idea behind replication is to select subsamples repeatedly from the whole sample, calculate the statistic of interest for each subsample, and then use these subsamples or replicate statistics to estimate the variance of the full-sample statistic. Different ways of creating subsamples from the full sample result in different replication methods. The subsamples are called replicates and the statistics calculated from these replicates are called replicate estimates.

One major advantage of replication methods is that they can produce variance estimates for statistics that might not be available in standard software. Another advantage is that the replication variance

estimation provides a simple way to account for adjustments that are made in weighting. As described above, the PATH Study full-sample weights are adjusted for nonresponse, trimming, and raking to control totals. By separately computing the weighting adjustments for each replicate, it is possible to reflect the effects of weight adjustments in the estimates of variance, which frequently results in a smaller variance because certain demographic estimates have been calibrated to known control totals (Valliant, 1993). Taylor series (linearization) methods for estimating the variance do not account for the variance reduction resulting from raking, and consequently often (but not always) provide variance estimates that are too large (Valliant, 2004; Chowdhury, 2013).

There are several ways of forming replicate weights, including balanced repeated replication (BRR), jackknife (JK-1, JK-2, and JK-n), and bootstrap. The choice of what kind of replicate weights to create is determined by the type of sampling design that was used to collect the data; in particular, whether or not stratification was used and how many PSUs were selected in each stratum. The replication method selected for the PATH Study is BRR (McCarthy, 1969).

The BRR method was selected for the PATH Study because (1) it allows calculation of the variance with fewer replicate weight variables than would be needed for the bootstrap for this dataset (this is possible because the subsets of PSUs for the replicates are carefully selected according to an orthogonal experimental design) and (2) BRR, unlike jackknife, produces consistent variance estimates for nonsmooth statistics such as quantiles (Rao and Shao, 1993, 1999). For the PATH Study, it was thought that quantiles might be of interest for some of the biological quantities studied; for example, it may be of interest to study the 10th percentile or the median of cotinine level. The BRR method allows correct estimation of standard errors for analyses involving quantiles and quantile regressions. The PATH Study uses a variant of BRR known as Fay's method (Judkins, 1990). Fay's method weights the inclusion of PSUs in the replicates by a predetermined value ϵ (set to 0.3 here), and produces more stability for the variance estimates for quantities in domains with small sample sizes. Using Fay's method, one half of the sample is weighted down by a factor ϵ and the remaining half is weighted up by a compensating factor $2 - \epsilon$.

Creation of Variables for Variance Estimation

The first step in constructing the replicate weights was to create variables for pseudo-strata and pseudo-PSUs that reflect the variance structure. These were created using methods described in Korn and Graubard (2011, p. 206) and are called VARSTRAT and VARPSU on the adult and youth/parent data files.

There are a total of 92 strata and 156 PSUs in the design. Fifty of the strata have two PSUs sampled; these were left as is for the variance estimation. Seven strata have three PSUs sampled. Because the BRR method assumes that two PSUs are selected from each stratum, two pseudo-PSUs were created for each of these seven strata by randomly selecting two of the three PSUs in each stratum and combining them into one pseudo-PSU. The remaining PSU then became the second pseudo-PSU in the stratum.

Thirty-five of the strata have one PSU that is selected with certainty and called self-representing (SR). In these strata, the variability comes from the secondary sampling units (segments), which were selected using an implicit stratification scheme that resulted in similar segments being close together in the ordered list. The following procedure was used to create pseudo-strata and pseudo-PSUs within the SR PSUs for variance estimation purposes.

1. Each SR PSU with fewer than 60 segments was treated as one pseudo-stratum with two pseudo-PSUs. The segments in the SR PSU were assigned to the two pseudo-PSUs so that the odd-numbered segments in the sorted list were assigned to one pseudo-PSU and the even-numbered segments in the sorted list were assigned to the other pseudo-PSU. In this way, adjacent segments in the sorted list were assigned to different pseudo-PSUs, so that the pseudo-PSU formation could take advantage of the implicit stratification used for sampling segments.

2. Four of the SR PSUs had large numbers of segments selected, and these were divided into pseudo-strata for variance estimation purposes. Because of the implicit stratification in the sort order, the pseudo-strata were formed by cutting the ordered list of segments into the desired number of pseudo-strata. Then, the segments within the pseudo-strata were assigned to pseudo-PSUs as described in (1).

Applying the BRR method to the PATH Study's pseudo-strata and the initial household weights yields 100 initial replicate weights. These 100 initial replicate weights were adjusted using the steps described in Section 4.1 to arrive at the final set of replicate weights for variance estimation.

The variance estimation variables for the adult and youth/parent data can be found in the adult data file and youth/parent data file, respectively. The 100 replicate weights are in the variables named R01_A_PWGT1 through R01_ A_PWGT100 for adults and R01_Y_PWGT1 through R01_ Y_PWGT100 for youth.

Software Options

The data files are provided in several formats. When the data analysis software package allows, the replicate weights should be used for all variance calculations to reflect the impact of the complex sample design and the various weighting adjustments on standard errors. Note that, although variables for pseudo-strata (VARSTRAT) and pseudo-PSUs (VARPSU) are included on the data files, variance estimates calculated using these variables with linearization (for example, by using the STRATA and CLUSTER statements in SAS) do not reflect the impact of the weighting adjustments and may result in incorrect inferences. Some example SAS, SUDAAN, STATA, R, and SPSS program code for generating popular statistics is provided in Appendix A.

The BRR replication method of variance estimation is available in both the SAS and Stata software packages. SPSS Complex Samples™ currently does not offer BRR or other forms of replication-based estimation. Note that the open-source R statistical software language does not have core programs that do survey data analysis. There are various contributed packages to R (such as the survey package) that analyze data from complex surveys and handle replication methods for variance estimation; however, these contributed packages are not peer-reviewed or subject to quality standards, and their features are subject to change in future versions of R.

Variance estimates for small domains may be unstable because some PSUs may contain no observations belonging to the domain. For small domains, the variance estimates produced by statistical software packages may differ because they use different methods to adjust for strata in which only one PSU contains domain members. For more information, see Graubard and Korn (1996), Lohr (2010, p. 570), or Lewis (2013).

4.3 Wave 1 Imputation

Demographic characteristics of adults and youth selected for the PATH Study were used in the creation of the Wave 1 sample weights. These included variables indicating gender, age, education level (for adults only), race, and ethnicity of the sampled persons. However, because some of this information may be missing for some sampled adults and youth, imputation methods were used to assign values when self-reported information was not available. For both adults and youth, imputation was performed by first considering information provided in the household screener and then by using statistical imputation methods. The imputation methods were performed for

respondents and nonrespondents[16] since data for both were needed for weighting, but only the values for respondents are included on the data files.

The sections below provide the methods used for imputing gender (Section 4.3.1), age (Section 4.3.2), education (Section 4.3.3), race (Section 4.3.4), and ethnicity (Section 4.3.5). Section 5.4 provides a description of the variable naming convention used for the adult data file and youth/parent data file.

4.3.1 Gender

Gender was assigned for adults and youth from information provided in the extended interview (questionnaire item R01_AM0004 for adults and item R01_YM0004 for youth). If this was not available, either because the sampled person refused to provide it or because the sampled person did not respond to the interview, gender was assigned based on the information provided in the household screener.

After reviewing these sources, however, gender remained unavailable for a small number of sampled persons. Common demographic variables (for example, Census region, age, education, race) are not indicators of gender and in most cases, this information was also unavailable if information on gender could not obtained from the extended interview and household screener, then gender was randomly assigned so that approximately one half of the cases with missing gender were assigned as female and the other half were assigned as male.

The values of R01_AM0004 and R01_YM0004 are provided in the variables R01R_A_SEX for adults and R01R_Y_SEX for youth, with the responses of "Refused" or "Don't Know" set to missing values. Note that 'A' in the variable name indicates adults and 'Y' in the variable indicates youth; for purposes of this report, *x* = A (adult) or Y (youth). All values, imputed and unimputed, are contained in the variable R01R_*x*_SEX_IMP; the variable R01R_*x*_SEX_IMPFLAG indicates which values are from the interview, i.e., not imputed (R01R_*x*_SEX_IMPFLAG = 0), from the household screener (R01R_*x*_SEX_IMPFLAG = 1), or the result of the random imputation method described above (R01R_*x*_SEX_IMPFLAG = 2).

[16] These are adults and youth belonging to responding households who did not respond to the interview.

4.3.2 Age

A single year of age variable was created for both adults and youth. For adults, age was first calculated based on the date of birth provided during the interview. If date of birth was not provided in the interview, the age in years was used if available.[17] For youth, age was first calculated based on the date of birth provided by the parent during the consent process. If date of birth was not provided, the age in years (also requested during the consent process) was used. The variable R01R_*x*_AGE contains the age in years assigned using these processes, where *x* = A (adult) or Y (youth).

If R01R_*x*_AGE was missing (because the responding adult refused to provide their date of birth or age, the responding youth was an emancipated minor and no parental consent was required, or the sampled person did not respond to the extended interview), age in years provided in the household screener was used. All values for respondents, whether imputed or not, are contained in the variable R01R_*x*_AGE_IMP; the variable R01R_*x*_AGE_IMPFLAG indicates which values are from the interview (R01R_*x*_AGE_IMPFLAG = 0) or from the household screener (R01R_*x*_AGE_IMPFLAG = 1), or for which no further information is available (adults only) (R01R_*x*_AGE_IMPFLAG = 3).

For nonresponding youth, if R01R_Y_AGE_IMP was missing after the above processing, a value was statistically imputed since age-in-years was used in the youth weighting process. Given that the only information available about these youth is from the household screener and that information (such as race, ethnicity or gender) is not indicative of a child's age, the age in years was randomly assigned using the youth population distribution according to the 2013 Public Use Microdata Sample (PUMS) data from the American Community Survey.

For adults, a four-level age category (18-24, 25-44, 45-64, and 65 and older) was used for the weighting process, but was not included in the analytic data file. If R01R_A_AGE_IMP was missing after the above processing, but the sampled adult indicated in the Phase 2 screener that they were in one of the age categories under 30 years old, that age category was used.[18] If that information was not available, the age category asked in the household screener was used.

[17] Only adults who did not provide their date of birth were asked to provide their age in years.

[18] Questionnaire item R01_AM0003 requesting an age category was asked in the interview only if the sampled adult refused to provide a date of birth or age in years. The highest age category was "30 or older."

Similar to the youth, if none of this information was available for an adult, the age category was imputed using the population distribution according to the 2013 PUMS data. Because a higher proportion of adults in the older age categories are female, the random assignment was conducted separately by gender. Age range information, provided in the Phase 2 screener, was used for some adults who indicated they were 30 years or older. For these adults, the age category was randomly assigned using the population distribution by gender according to the 2013 PUMS data for that age group. If it was unknown whether the sampled adult was 30 or above, the age category was randomly assigned using the population distribution by gender for those 18 and older.

4.3.3 Education Category (Adults Only)

Adult respondents were asked questionnaire item R01_AM0018 regarding the highest level of education they attained. A five-level education variable was used for the adult weighting and created by collapsing the 11 response categories: less than high school or GED; high school graduate; some college, but no degree, or associates degree; bachelor's degree; advanced degree. Missing values were imputed using hot deck imputation because demographic characteristics that are indicative of educational attainment were available.[19] The imputation cells were formed by cross-classifying categories of census region, age category, and gender, using imputed values (calculated as described above) as appropriate.

The variable R01R_A_EDUC contains the five education categories before imputation. All values, imputed and unimputed, are contained in the variable R01R_A_EDUC_IMP. The variable R01R_A_EDUC_IMPFLAG indicates which values are from the interview (R01R_A_EDUC_IMPFLAG = 0) or the result of the imputation method described above (R01R_A_EDUC_IMPFLAG = 2).

4.3.4 Race

A four-level race variable was used for both the adult and youth weighting. In the extended interview, respondents were asked to indicate which of 14 race categories applied to them (questionnaire items with the prefix R01_AM0006 for adults and items with the prefix R01_YM0006

[19] Hot deck imputation is a method for handling missing data in which each missing value is replaced with an observed response from a "similar" unit in the imputation cell.

for youth). The responses were combined into a single variable with four categories: white alone, black alone, Asian alone (including multiple Asian categories), and other (including multi-racial). The variables R01R_A_RACE and R01R_Y_RACE contain the combined race responses from the extended interview for adults and youth, respectively.

If the sampled adult or youth did not respond to the race question, or did not respond to the extended interview, the race information was assigned according to the information provided in the household screener.[20] The race information gathered for the sampled person in the household screener was considered first; if that was not available, the race information that the household screener respondent reported about himself or herself was used as a proxy for the sampled person.

If none of this information was available, race was imputed using hot deck imputation because demographic characteristics that are indicative of race were available. For adults, imputation cells were formed by cross-classifying categories of census region, age category, and education category, using imputed values (calculated as described in the preceding sections) as appropriate; for youth, the hot deck imputation cells were the four categories of census region.

All values, imputed and unimputed, are contained in the variable R01R_*x*_RACE_IMP, where *x* = A (adult) or Y (youth). The variables R01R_A_RACE_IMPFLAG and R01R_Y_RACE_IMPFLAG indicate which values are from the interview (R01R_*x*_RACE_IMPFLAG = 0), from the household screener (R01R_*x*_RACE_IMPFLAG = 1), or the result of the imputation method described above (R01R_*x*_RACE_IMPFLAG = 2).

4.3.5 Ethnicity

A two-level ethnicity variable indicating whether the sampled person is of Hispanic origin was used for both the adult and youth weighting. This variable was initially created from the Hispanic origin question asked in the extended interview (questionnaire item R01_AM0005_01 for adults and item R01_YM0005_01 for youth). The variables R01R_A_HISP and R01R_Y_HISP indicate whether the respondent is of Hispanic origin based on these variables for adults and youth, respectively.

[20] There were only five race category responses in the household screener, but they may be uniquely coded into the four race categories used in weighting.

If the sampled adult or youth did not respond to this question, or did not respond to the extended interview, the ethnicity information was assigned according to the information provided in the household screener: The ethnicity information gathered for the sampled person in the household screener was considered first; if that was not available, the ethnicity information that the household screener respondent reported about himself or herself was used as a proxy for the sampled person.

If none of this information was available, ethnicity was imputed using hot deck imputation because demographic characteristics that are indicative of ethnicity were available. For adults, imputation cells were formed by cross-classifying categories of census region, age category, and education category, using imputed values (calculated as described in the preceding sections) as appropriate; for youth, imputation cells were formed by the four categories of region. In Wave 1, while the ethnicity variable was imputed for adults, imputation was not necessary for the youth.

All values for respondents, imputed or not, are contained in the variable R01R_*x*_HISP_IMP, where *x* = A (adult) or Y (youth). The variables R01R_A_HISP_IMPFLAG and R01R_Y_HISP_IMPFLAG indicate which values are from the interview (R01R_*x*_HISP_IMPFLAG = 0), from the household screener (R01R_*x*_HISP_IMPFLAG = 1), or from use of the hot deck imputation method described above (R01R_*x*_HISP_IMPFLAG = 2).

Data Files 5

5.1 Data Structure

As discussed in Section 1.2, the adult data file and youth/parent data file have similar content. They both include the questionnaire items from the interview, some derived variables such as tobacco-use definitions, and other variables such as census geographic information, an indicator whether a sampling area is urban, and some census block-level demographic characteristics.

Adult: The adult data file contains one record for every completed adult interview.

Youth/Parent: The youth/parent data file contains one record for every completed youth interview and also contains responses to the brief parent interview about the youth who completed the youth interview. Every youth interview may not have a corresponding parent interview because a parent/guardian can refuse to complete the parent interview but give permission for his/her youth to complete the youth interview. One parent could have completed multiple parent interviews for multiple youths within a household. There are certain questions in the parent interview that are asked only once if the same parent completed interviews for multiple youths within a household. The data for the variables that were not asked in additional interviews with the same parent have been copied from the data record of the first interview completed by the parent. For reference, the parent questions that are not repeated are identified in the parent annotated instrument (combined in the same file as the youth annotated instrument). For Wave 1, these are identified in Boxes R01_PTR01 and R01_PNR01.

5.2 Record Identifier

The variable named PERSONID serves as the unique identifier for each record in the adult and youth/parent data files. In the adult data file, each PERSONID represents a unique adult respondent. In the youth/parent data file, each PERSONID represents a unique youth respondent. Each PERSONID begins with a "P" and is followed by a randomly assigned nine-digit number that does not include any direct or indirect references to personally identifiable information or geographic location.

5.3 Variables Excluded

Some specific questionnaire items are not provided in the data files. This may be for one of two reasons:

1. Some items are excluded to reduce the risk of disclosing a respondent's identity. All direct identifiers such as names, addresses, and phone numbers are excluded from the files, but other specific questionnaire items are also excluded for this reason.

2. Some items were collected or created in the instrument for operational purposes only. These include items derived throughout the instrument for routing purposes, including items that indicate types and levels of tobacco usage. (See the appendices of the annotated instruments for the full list of tobacco-use variables used for routing purposes.) Variables indicating types and levels of tobacco usage that are better suited for analysis were derived and provided with the data files.

Other items require more processing before they are ready to be used for analysis. For Wave 1, this includes the variables containing the UPC codes resulting from the product bar code scanning during the adult interview (the AX0217 series) and an occupation code derived from item AM0016. It is anticipated that these variables will be included in a later update to the restricted use data files. Table B-1 in Appendix B contains the full list of questionnaire items excluded from the initial Wave 1 data files.

5.4 Variable Names

The PATH Study variable naming convention was designed with a goal to help provide context for each variable based on the content being represented by the item. The intention of the naming convention is to help make it easier for data users to identify and classify items for analyses. The PATH Study is longitudinal, and therefore the variable naming convention includes an indicator to identify the data collection wave.

The naming convention applies to all variables on the data files, with the exception of record identifiers and the variables that reflect the variance structure. It is anticipated that these items will retain the same variable name throughout the course of the PATH Study.

Variables on the PATH Study data files are named using multiple components, each of which is separated by an underscore "_" character for readability. The naming convention is based on a syntax methodology that begins with broader modifiers, moving to more specific modifiers. The basic structure of the naming convention is given in Figure 1. This structure creates variables that are platform independent (no unusual characters) and easy to read, and allows variables to be visually identifiable as related. The structure also prevents a variable on one file that happens to have the same name as, but different contents from, a variable on a second file from overwriting the first variable when records from the two files are merged (e.g., when merging adult and youth/parent data files). Users may stack (append) datasets for analysis including both adults and youth, after renaming the variables of interest including the weight and replicate weight variables.

Figure 1. PATH Study variable name components

The components of the variable names are:

- **Wave Indicator:** Each variable name begins with a wave indicator, represented by three characters. The first character is "R", which is followed by a two-digit zero-filled number indicating the data collection wave, beginning with "01" for Wave 1. As such, all variables in the Wave 1 data files other than PERSONID, VARSTRAT and VARPSU will have the prefix "R01".

  All derived and imputed variables include the additional character "R" (indicating that the variable is a recode) after the wave indicator. As such, all derived and imputed variables on the Wave 1 data files will have the prefix "R01R". Similarly, all variables created from an external source include the additional character "X" so that all such variables on the Wave 1 data files will have the prefix "R01X".

- **Modifiers**

  - **PATH ID:** A PATH ID represents a unique question in an instrument. All interview variables are named using the PATH IDs that were assigned to questions in the instrument specifications. PATH IDs are represented in the format: @$####, where @ represents the instrument type (A for Adult, P for Parent, and Y for Youth), $ represents the item content, and #### is a numeric key that is unique within each instrument. The numbers in the PATH IDs do not correspond to any specific ordering of questions; they were simply assigned during the instrument design process based on availability. Some variable names include suffixes to the PATH ID to differentiate between tobacco product types, when the same question was repeated for different tobacco products. For example, variable R01_AG1005TC corresponds to PATH ID AG1005, when it

was asked about traditional cigars in the adult interview. In some instances, there are questions with similar content that are asked in different sections or in different instruments. In such cases, the number identifier in the PATH ID is the same across both sections/instruments whereas the instrument/content indicator would be different. For example, in Wave 1, question AC1005 (lifetime use of cigarettes) in the cigarette section is also asked of e-cigarettes in question AE1005 (lifetime use of e-cigarettes).

– **Interview Identifier:** Where appropriate, administrative, management, processing, derived, and imputed variables include an indicator of the interview to which the information pertains: "A" (adult interview), "Y" (youth interview), "E" (emancipated youth) or "P" (parent interview). For example, the month and year of the parent interview is in the variable named R01R_P_INTERVIEW_MMYR.

– **Mnemonic:** All administrative, management, processing, derived, external and imputed variables are named using mnemonics. Mnemonics can be full words, phrases, or abbreviations to represent the content of the variable, and can range from the very simple to the complex. For example, the variable indicating the youth's height in inches at Wave 1 is R01R_Y_HTINCH.

The tobacco-use variables have components in the mnemonic that indicate the status, intensity, timeframe, and product. For example, the variable indicating whether an adult was a former established past 12 months cigarette user at Wave 1 is R01R_A_FMR_ESTD_P12M_CIGS. Table 4 includes a list of the tobacco-use mnemonic components.

Table 4. Tobacco-use variable name mnemonic components[21]

Type/Attribute	Abbreviation	Description
Status	CUR	Current
	EDY	Every Day
	SDY	Some Days
	NVR	Never
	FMR	Former
	INT	Initiator
	NLL	Not at All
	EVR	Ever
Intensity	ESTD	Established User
	EXPR	Experimental User
	NEST	Not Established

[21] The maximum length for a variable name in the PATH Study data files is 32 characters. Current and recent versions of most statistical software packages are able to accommodate variable names with 32 (or more) characters.

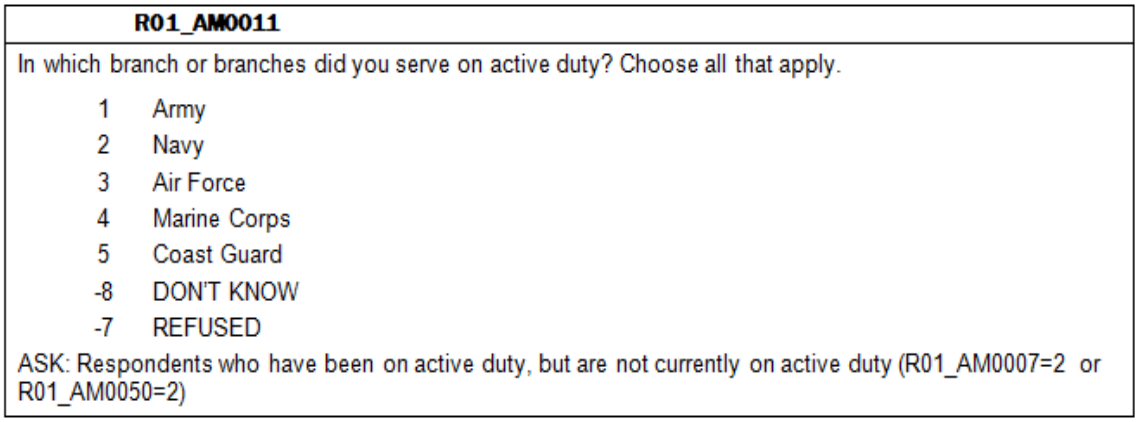
Table 4. Tobacco-use variable name mnemonic components (continued)

Type/Attribute	Abbreviation	Description
Timeframe	12MA	12 Months Ago (year ago)
	P12M	In Past 12 Months
	6MA	6 Months Ago
	P6M	In Past 6 Months
	30DA	30 Days Ago
	P30D	In Past 30 Days
	7DA	7 Days Ago
	P7D	In Past 7 Days
Product	CIGS	Cigarettes
	CIGSMFG	Cigarettes, Manufactured
	CIGSRYO	Cigarettes, Roll Your Own
	DISSBL	Dissolvables
	ECIG	E-Cigarettes
	GFILTR	Filtered Cigars
	GRILLO	Cigarillos
	GTRAD	Traditional Cigars
	HOOK	Hookah
	NRT	Nicotine Replacement Therapy
	PIPE	Pipe
	RX	Prescription Therapy
	SMKLS	Smokeless
	SNUS	SNUS
	BIDI	Bidis
	KRETEK	Kreteks
	TOB	All Tobacco Products

- **Source:** As additional sources of external data are added, a modifier for the source will be included. For example, consider item R01X_CB_URBAN_PSU, which represents the urban PSU indicator; CB identifies the information source as the Census Bureau.

- **Sub-item Identifier:** Some interview questions have multiple components for a single PATH ID, such as choose-all-that-apply items or dates where responses are recorded in multiple variables. In such cases, the variable name includes a sub-item identifier to differentiate components. For example, consider item R01_AM0011, which asks about branch of military service when on active duty. The question and response options are shown in Figure 2 for reference.

Figure 2. Example of question and response options



R01_AM0011

In which branch or branches did you serve on active duty? Choose all that apply.

1	Army
2	Navy
3	Air Force
4	Marine Corps
5	Coast Guard
-8	DON'T KNOW
-7	REFUSED

ASK: Respondents who have been on active duty, but are not currently on active duty (R01_AM0007=2 or R01_AM0050=2)

The respondent could choose multiple responses for this item. Because there are five sub-items for R01_AM0011, there are five variables (R01_AM0011_01 through R01_AM0011_05), each of which corresponds to one of these sub-items. The variables for the sub-items correspond to the order of the responses presented in the question.[22] Therefore, if a respondent chose "Army" and "Coast Guard" as answers to R01_AM0011, only the variables R01_AM0011_01 and R01_AM0011_05 would contain the value that means the respondent chose that response option (typically this value is 1).

5.5 Variable Labels

All variables on the Wave 1 data files include a label that briefly explains the content of the item. For clarity, consistency, and usability, labels are assigned to each variable using a standard convention. The components of the variable labeling convention are shown in Figure 3.

Figure 3. PATH Study variable labeling convention

For all interview, administrative, and management variables, the labels begin with the variable name followed by a colon and a brief description of the content. For derived, imputed, or imputation flag variables, the labels begin with the variable name followed by a colon, the respective phrase "DERIVED", "IMPUTED", or "IMPUTATION FLAG", a spaced hyphen and then a brief description of the content.

[22] An exception to this is the set of sub-items relating to adult race (R01_AM0006), which has alpha character sub-item identifiers (such as "WH" indicating "White") rather than numeric ones.

5.6 Value Labels

The labels associated with variable values are provided. All value labels include both the data value as well as the description of the value. For example, a variable for an item with yes/no responses would have "1 = Yes" and "2 = No" as value labels.

Interview Variables

The value labels for all interview variables match the text of the response option as seen by the respondent.

Recoded Values

Certain variables have new recoded values that were not presented as response options during the interview. Recoded values are assigned to items with Other-Specify response options. The label for any recoded value is prefixed with the term "RECODE:" for explicit identification.

Values for Derived, Imputed, and External Variables

The labels for all created values for derived, imputed, and external variables include the coded value and a description of the value. If a value could not be assigned, it is coded as system missing. Please see table 5 for a description of the system missing value.

5.7 Missing Values

The data files include missing values, as currently defined in Table 5.

Table 5. PATH Study missing value codes and their descriptions

Definition		Applicable Data Types	Applicable Variable Types	Description
-1	Inapplicable	Numeric, Character	Interview	This value is coded to identify items that were validly skipped on a route specific to a respondent.
-5	Improbable response removed	Numeric	Interview	This value is coded to represent improbable numeric answers provided by respondents.
-7	Refused	Numeric, Character	Interview	This value is coded when respondents answered a question as "Refused."
-8	Don't know	Numeric, Character	Interview	This value is coded when respondents answered a question as "Don't know."
-9	Missing – Not ascertained	Numeric, Character	Interview	This value is coded when data were supposed to be collected on a route specific to a respondent but the collection did not occur due to a skip error in the instrument.
.	System missing	Numeric	Derived, Imputed	This value is coded in derived variables when none of the algorithmic conditions specified to assign a non-missing value were satisfied. It is also coded in imputed variables when no value was imputed and in date/time variables.

5.8 Outlier Values

The PATH Study adult and youth instruments were self-administered, and although there were checks throughout the instrument to query potentially illogical or unlikely responses, the number of such checks was limited in order to minimize frustration on the part of the respondent. Moreover, these checks were designed to warn the respondent if a response was unlikely, but the respondent could continue through the instrument without changing the response. In order to maintain the confidentiality afforded by the ACASI instrument, it was important to minimize the number of instances where the respondent might need to ask the field interviewer for assistance in overcoming a check that required correction of data before continuing. As a result, the PATH Study questionnaire data include outliers, or values that lie outside the expected range of data. The two types of outliers were handled in two different ways.

First, there are outliers that are logically impossible. For example, a few respondents reported doing something at an age that was older than their current age (such as a respondent who was age 37 reporting initiation of cigarillo use at age 39). For these clear instances where a response was logically impossible, the outlier was removed from the dataset, replaced with a value of -5, and

labeled as "Improbable response removed." Please see Appendix C for a full list of variables with outlier values coded.

Second, there are outliers that are unlikely but logically possible. For example, a few respondents reported sending more than 1,000 text messages a day, which is unlikely but not impossible. In the absence of established upper or lower limits for such unlikely but logically possible values, they were retained in the dataset. Therefore, analysts are advised to review carefully the distributions of the variables used in all analyses, and to consider the impact of these values on any findings.

6 Linking Files

The PATH Study restricted use files include a master linkage data file that for Wave 1 indicates in which file(s) a respondent has data: the adult data file, the youth/parent data file, and/or specific biospecimen data files as they become available. For subsequent waves, this file will indicate whether a Wave 1 respondent completed an interview and, if so, identify the type of interview completed. This linkage file will be extended for each subsequent wave. The linkage file can help analysts identify which files contain data for a particular respondent (or a set of respondents). This file is described in more detail in Section 6.1.

Records in the data files within a wave may be linked together using the unique household identifier provided with the adult and youth interview data. Also, some parents responding to the parent interview were sampled for the adult cohort. If the parent responded to the adult interview as well, it is possible to link the youth/parent record to the adult record (and potentially to the adult interview record of the parent's spouse or partner). These types of linkages are described further in Section 6.2.

Section 6.3 provides some general file merging notes and recommendations.

6.1 Master Linkage Data File Structure

The master linkage file includes a unique record for each PERSONID for all Wave 1 respondents with a complete adult or youth interview. The master linkage data file also includes records for respondents identified as shadow youth at Wave 1, even though no data exist for shadow youth in the PATH Study Wave 1 data files. The records for shadow youth are included in the master linkage data file to account for respondents aging up to youth in Wave 2 and subsequent waves.

The linkage data file will be maintained as a standalone dataset that will be updated with each data file release. The number of records in the linkage data file is defined by the adult, youth, and shadow samples established in Wave 1 and should not change across waves unless or until the PATH Study design calls for replenishment sample to be added at some future wave. The number of variables on the linkage data file will increase as each new file with respondent-level data is released. For example,

as biospecimen laboratory analysis data are delivered, new variables will be added to the linkage data file to identify each biospecimen data file in which a respondent could have a data record. The number of variables on the linkage data file will also increase as data for future waves become available.

The first iteration of the linkage data file will only include two variables: PERSONID and WAVE1_INTERVIEW. The PERSONID represents the unique identifier for each PATH Study respondent. The variable WAVE1_INTERVIEW identifies whether the respondent completed an adult or youth interview or was a shadow youth at Wave 1.

In subsequent waves, variables specific to the wave will be added to the linkage data file. For example, in Wave 2 the variable WAVE2_INTERVIEW will identify whether the person completed an adult or youth interview, remained a shadow youth, or was a nonrespondent.

6.2 Linking Records Within and Across the Adult and Youth/Parent Data Files

The adult data file and the youth/parent data file contain an identifier for each household in which the adults and youth lived at the time of the interview. For the Wave 1 files, this variable is named R01_HHID. Each HHID begins with an "H" and is followed by a randomly assigned 11-digit number that does not include any direct or indirect references to personally identifiable information or geographic location. This identifier is included as a way to provide contextual information about respondents linked to the same household, and not as a promotion of analysis at the household level. Because household-level weights are not provided, inferences at the household level cannot be made.

Two other variables included on the youth/parent data file represent the unique PERSONIDs of the parent/guardian asked to complete the parent interview (for Wave 1, R01_PARENT_PERSONID) and his or her spouse or partner (for Wave 1, R01_PT0046_PERSONID), if one was specified. These are available because all persons in screened households were assigned a PERSONID. If the parent or spouse identified in the parent interview was sampled and responded to the adult interview, as indicated in the Wave 1 variables R01R_P_PARENT_ADULTCOMP and R01R_P_PT0046_ADULTCOMP, then these IDs

provide a link between the youth and parent questionnaire responses and the adult questionnaire responses.

This linkage ties these adult and youth respondents together. However, it is important to note that it was not a stipulation of the PATH Study that youth respondents also have a parent respond to the parent interview or that the parent or his/her spouse/partner respond to the adult interview. As such, if there is no associated adult interview data for a parent and/or spouse, it is because the person was not selected for the adult interview or was selected but did not respond to that interview.

Any information gained from linking a parent's adult interview data to the youth record should be viewed as a characteristic of the youth. Analyses of such data should use the youth weight and inferences should be described as characteristics of the youth.

Some example program code illustrating the linking of the Wave 1 adult and youth/parent data using R01_HHID, R01_PARENT_PERSONID, and R01_PT0046_PERSONID is provided in Appendix D.

6.3 File Merging Notes and Recommendations

Because the adult and youth/parent data files have a large number of variables, merges that retain all variables from multiple files should be avoided. Although the process of merging the data files with large numbers of variables is not resource intensive, the resulting (merged) file could be quite large, which could significantly increase the amount of time to run analyses. As such, when merging data files, it is recommended to keep only the variables needed from each file to run the analyses. This practice keeps the merged file(s) as small as possible, which helps analytic processes run faster.

The records included in the file(s) that result from merging two or more files can vary depending on how the merge is specified. As such, please carefully consider the set of cases needed for any specific analyses and define the merge accordingly.

6.4 Case Identification Number

ICPSR creates a variable for each of its datasets named CASEID. This variable numbers the cases sequentially so that the file can easily be returned to its original order at any time. This variable should not be confused with any other identification variable previously mentioned. It should not be used to merge files or link records together.

References

Battaglia, M.P., Hoaglin, D.C., and Frankel, M.R. (2013). Practical considerations in raking survey data. *Survey Practice*, 2(5).

Chowdhury, S.R. (2013). Effects of poststratification and raking on precision of MEPS estimates. Proceedings of the Federal Committee on Statistical Methodology Conference, https://fcsm.sites.usa.gov/files/2014/05/F1_Chowdhury_2013FCSM.pdf.

Chowdhury, S., Khare, M., and Wolter, K. (2007). Weight trimming in the National Immunization Survey. *Proceedings of the Survey Research Methods Section of the American Statistical Association*, 2651-2658.

Golder, P.A. and Yeomans, K.A. (1973). The use of cluster analysis for stratification. *Applied Statistics*, 22, 213-219.

Graubard, B.I., and Korn, E.L. (1996). Survey inference for subpopulations. *American Journal of Epidemiology*, 144, 102-106.

Judkins, D.R. (1990). Fay's method for variance estimation. *Journal of Official Statistics*, 6, 223-239.

Judkins, D.R. and Singh, R.P. (1981). Using clustering algorithms to stratify primary sampling units. *Proceedings of the American Statistical Association Survey Research Methods Section*, 279-284.

Korn, E., and Graubard, B. (2011). *Analysis of Health Surveys.* Hobart, NJ: Wiley.

Krewski, D. and Rao, J.N.K. (1981). Inference from stratified samples: Properties of the linearization, jackknife and balanced repeated replication methods. *The Annals of Statistics*, 9, 1010-1019.

Lewis, T. (2013). Considerations and techniques for analyzing domains of complex survey data. SAS Global Forum Proceedings, http://support.sas.com/resources/papers/proceedings13/449-2013.pdf.

Lohr, S. (2010). *Sampling: Design and Analysis, 2nd ed.* Boston: Brooks/Cole.

McCarthy, P.J. (1969). Pseudo-replication: Half samples. *Review of the International Statistical Institute*, 37, 239-264.

Rao, J.N.K., and Shao, J. (1993). On balanced half sample variance estimation in stratified sampling. *Journal of the American Statistical Association*, 91, 343-348.

Rao, J.N.K., and Shao, J. (1999). Modified balanced repeated replication for complex survey data. *Biometrika*, 86, 403-415.

SAS Institute Inc. (2011). *SAS/STAT® User's Guide.* Cary, NC: SAS Institute, Inc.

Valliant, R. (1993). Poststratification and conditional variance estimation. *Journal of the American Statistical Association*, 88, 89-96.

Valliant, R. (2004). The effect of multiple weighting steps on variance estimation. *Journal of Official Statistics*, 20, 1-18.

Wolter, K.M. (2007). *Introduction to Variance Estimation*, 2nd ed. New York, NY: Springer.

Example Program Code for Generating Popular Statistics

This appendix contains example SAS, SUDAAN, Stata, R, and SPSS program code for generating statistics using the PATH Study data. A couple of notes about the examples:

- Text in italics represents placeholders for actual dataset and variable names.
- These examples use the adult full-sample and replicate weights, but analyses of youth/parent data would be the same with the respective youth full-sample and replicate weights.
- In the examples using subdomain analysis, the domain of interest (e.g., Asian youth or adults ages 18-24) is identified by setting *domainvar* = 1 for those cases, and setting *domainvar* = 0 for all other cases.

These examples are provided primarily to illustrate the correct specifications for creating appropriate variance estimates. They are not meant to be exhaustive or to provide instruction for users unfamiliar with a particular software package.

SAS

The following code creates tables including the unweighted frequencies of categorical variables *var1, var2, var3* and *var4*, and weighted estimates of population totals and population proportions for each level of those variables (using the weight R01_A_PWGT) along with the standard errors of these estimates (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

```
proc surveyfreq data=analysis_dataset
varmethod=BRR (fay=0.3);
    tables var1 var2 var3 var4;
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
    title "PROC SURVEYFREQ using BRR-Fay Replication";
run;
```

The following code creates the weighted mean of continuous variable *var5* (using the weight R01_A_PWGT) along with the standard error of that estimate (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

```
proc surveymeans data=analysis_dataset
varmethod=BRR (fay=0.3);
    var var5;
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
    title "PROC SURVEYMEANS using BRR-Fay Replication";
run;
```

The following code fits a linear regression model using continuous variable ***respvar*** as the outcome variable and continuous variables ***cov1*** and ***cov2*** as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100).

```
proc surveyreg data=analysis_dataset
varmethod=BRR (fay=0.3);
    model respvar = cov1 cov2 / solution;
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
    title "PROC SURVEYREG using BRR-Fay Replication";
run;
```

The following code fits a logistic regression model using dichotomous variable ***respvar2*** (with values of 0 and 1) as the outcome variable and continuous variables ***cov1*** and ***cov2*** as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100).

```
proc surveylogistic data=analysis_dataset
varmethod=BRR (fay=0.3);
    model respvar2 (event='1')= cov1 cov2 ;
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
    title "PROC SURVEYLOGISTIC using BRR-Fay Replication";
run;
```

The following code provides examples of domain analyses in SAS. SAS will produce output for all levels of the domain variable. The Output Delivery System (ODS) statements may also be used to output the domain cases of interest for further processing.

```
proc surveyfreq data=analysis_dataset varmethod=BRR (fay=0.3);
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
```

```
    table var1;
    by domainvar;
    title 'Domain Analysis for PROC SURVEYFREQ using BRR-Fay
    Replication';
run;

proc surveymeans data=analysis_dataset varmethod=BRR (fay=0.3);
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
    var var5;
    by domainvar;
    title 'Domain Analysis for PROC SURVEYMEANS using BRR-Fay
    Replication';
run;

proc surveyreg data=analysis_dataset
varmethod=BRR (fay=0.3);
    model RespVar = cov1 cov2 / solution;
    by domainvar;
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
    title 'Domain Analysis for PROC SURVEYREG using BRR-Fay
    Replication';
run;

proc surveylogistic data=analysis_dataset
varmethod=BRR (fay=0.3);
    model respvar2 (event='1')= cov1 cov2 ;
    by domainvar;
    weight R01_A_PWGT;
    repweights R01_A_PWGT1 - R01_A_PWGT100;
    title 'Domain Analysis for PROC SURVEYLOGISTIC using BRR-
    Fay Replication";
run;
```

Note that the current SAS/STAT software documentation states that for a complex survey design, analyses for a subpopulation should specify the subpopulation through a `DOMAIN` statement in `proc surveymeans`, `proc surveyreg`, or `proc surveylogistic`, or as the first dimension in the TABLES statement in `proc surveyfreq` (see SAS Institute Inc. 2011, pp. 7417-7419). However, a simplification is possible for analysis of PATH Study data using the replicate weights. (Replication is the recommended method of variance estimation for the PATH Study). Specifically, the `DOMAIN` statement is not required for analyses using the replication method for variance estimation with the replicate weights provided with the data files. The appropriateness of this simpler approach has been confirmed with software developers at SAS Institute, Inc.

Thus, when using the replicate weights provided with the PATH Study data, the data may be subset to the subpopulation of interest outside the SAS/STAT software procedures or within the procedures using a `BY` or `WHERE` statement to produce the correct standard errors for confidence intervals and hypothesis tests even though, currently, a warning may appear in the program log stating that the method "does not provide a statistically valid subpopulation or domain analysis." SAS Institute, Inc. has communicated that they plan to make the necessary changes in the next release to eliminate that message for replication methods where the replicate weights have been created using the full data set. In addition, they will also update the documentation to reflect that the `DOMAIN` statement is needed only for the linearization method of variance estimation or when replicate weights are created as part of the same SAS survey procedure that requests the variance estimates. When the `REPWEIGHT` statement is used with the replicate weights provided for the PATH Study, an analysis with the `DOMAIN` statement and an analysis that subsets the data to the subpopulation of interest will produce the same, correct, standard errors. For this reason, the example SAS code provided in this User Guide does not include a `DOMAIN` statement.

SUDAAN

When using Fay's method of BRR in SUDAAN, the BRR-Fay factor must be calculated outside of the software as shown in equation A-1:

$$ADJFAY = 1/[(1-\mathrm{k})^2] \qquad \text{(A-1)}$$

When $k = 0.3$, as with the PATH Study, $ADJFAY =$ 2.040816.

The following code creates tables including the unweighted frequencies of categorical variables *var1*, *var2*, *var3* and *var4*, and weighted estimates of population totals and population proportions for each level of those variables (using the weight R01_A_PWGT) along with the standard errors of these estimates (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

```
proc crosstab data= analysis_dataset filetype=sas design=brr;
        weight R01_A_PWGT;
        repwgt R01_A_PWGT1 - R01_A_PWGT100/adjfay=2.040816;
        tables var1 var2 var3 var4;
        class  var1 var2 var3 var4;
        print /style=nchs tablecell=all  ;
        title 'SUDAAN proc crosstab using BRR-Fay Replication';
```

```
run;
```

The following code creates the weighted mean of continuous variable *var5* (using the weight R01_A_PWGT) along with the standard error of that estimate (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

```
proc descript data= analysis_dataset filetype=sas design=brr;
        weight R01_A_PWGT;
        repwgt R01_A_PWGT1- R01_A_PWGT 00/adjfay=2.040816;
        var  var5;
        print /style=nchs;
        title 'SUDAAN proc descript using BRR-Fay Replication';
run;
```

The following code fits a linear regression model using continuous variable *RespVar* as the outcome variable and continuous variables *cov1* and *cov1* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100).

```
proc regress data= analysis_dataset filetype=sas design=brr;
        weight R01_A_PWGT;
        repwgt R01_A_PWGT1- R01_A_PWGT100/adjfay=2.040816;
        model  respvar = cov1 cov2;
        title 'SUDAAN proc regress using BRR-Fay Replication';
run;
```

The following code fits a logistic regression model using dichotomous variable *respvar2* (with values of 0 and 1) as the outcome variable and continuous variables *cov1* and *cov2* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100).

```
proc rlogist data=test filetype=sas design=brr;
        weight R01_A_PWGT;
        repwgt R01_A_PWGT1- R01_A_PWGT100/adjfay=2.040816;
        model  respvar2 = cov1 cov2;
        title 'SUDAAN proc rlogist using BRR-Fay Replication';
run;
```

Domain analyses can be performed using the same examples above with the inclusion of a subpopn statement. For example, for frequencies *var1*, *var2*, *var3* and *var4* for just those respondents with *domainvar* = 1, the following code could be used:

```
proc crosstab data= analysis_dataset filetype=sas design=brr;
        subpopn domainvar = 1;
        weight R01_A_PWGT;
        repwgt R01_A_PWGT1 - R01_A_PWGT100/adjfay=2.040816;
        tables var1 var2 var3 var4;
        class  var1 var2 var3 var4;
        print /style=nchs tablecell=all;
        title 'SUDAAN proc crosstab using BRR-Fay Replication';
run;
```

Stata

The full-sample weight, variance estimation method, BRR-Fay replicate weights, and Fay's factor are relayed to the Stata using the `svyset` statement. The following statement should be used with the PATH Study data:

```
svyset [pweight= R01_A_PWGT], brr(R01_A_PWGT1 - R01_A_PWGT100)
vce(brr) mse fay(.3)
```

Assuming this `svyset` is used, the following code creates tables including the unweighted frequencies of categorical variables *var1*, *var2*, *var3* and *var4*, and weighted estimates of population totals and population proportions for each level of those variables (using the weight R01_A_PWGT) along with the standard errors of these estimates (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

For weighted frequencies/estimates of population totals

```
svy: tabulate  var1, count se
svy: tabulate  var2, count se
svy: tabulate  var3, count se
svy: tabulate  var4, count se
```

For weighted/population proportions

```
svy: tabulate  var1, se obs percent
svy: tabulate  var2, se obs percent
svy: tabulate  var3, se obs percent
svy: tabulate  var4, se obs percent
```

Note that if all four variables are used in a single tabulate statement, the result is a multi-dimensional table containing all four variables rather than four one-dimensional tables.

To create these estimates for respondents with *domainvar* = 1, the following code may be used for each variable:

```
svy: tabulate  var1, count se subpop(domainvar)
svy: tabulate  var1, se obs percent
     subpop(domainvar)
```

Note that in the example above, Stata assumes that the subdomain of interest has the value of "1" in the variable.

Assuming the `svyset` statement above is used, the following code creates the weighted mean of continuous variable *var5* (using the weight R01_A_PWGT) along with the standard error of that estimate (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

```
svy: mean var5
```

To create these estimates for each of the levels of *domainvar*, the following code may be used:

```
svy: mean var5, over(domainvar)
```

Assuming the `svyset` statement above is used, the following code fits a linear regression model using continuous variable *respvar* as the outcome variable and continuous variables *cov1* and *cov1* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100):

```
svy: regress  respvar cov1 cov2
```

To perform this regression for respondents with *domainvar* = 1, the following code may be used:

```
svy, subpop(if domainvar==1): regress respvar cov1 cov2
```

The following code fits a logistic regression model using dichotomous variable *respvar2* (with values of 0 and 1) as the outcome variable and continuous variables *cov1* and *cov2* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100).

For odds ratios

```
svy: logistic respvar2 cov1 cov2
```

For coefficient estimates

```
svy: logit respvar2 cov1 cov2
```

To perform this regression for respondents with *domainvar* = 1, the following code may be used:

```
svy, subpop(if domainvar ==1): logistic respvar2 cov1 cov2
svy, subpop(if domainvar==1): logit respvar2 cov1 cov2
```

R

The full-sample weight, variance estimation method, BRR-Fay replicate weights, and Fay's factor are relayed to the R survey package using the `svrepdesign` function. The following should be used with the PATH Study data:

```
options(survey.replicates.mse=TRUE)

y <-
  svrepdesign(
    id = ~PID,
    weights = ~R01_A_PWGT,
    repweights =  "R01_A_PWGT[1-9]+",
    type = "Fay",
    rho = 0.3,
    data = analysis_dataset
  )
```

Assuming the options and `svrepdesign` function above are executed, the following code creates the estimates of population totals and population proportions of the categorical variables *var1*, *var2*, *var3* and *var4* (using the weight R01_A_PWGT) along with the standard errors of these estimates (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

For weighted frequencies/estimates of population totals

```
svytotal(~factor(var1),design = y,na.rm=T)
svytotal(~factor(var2),design = y,na.rm=T)
svytotal(~factor(var3),design = y,na.rm=T)
svytotal(~factor(var4),design = y,na.rm=T)
```

For weighted/population proportions

```
svymean(~factor(var1),design = y,na.rm=T)
svymean(~factor(var2),design = y,na.rm=T)
```

```
svymean(~factor(var3),design = y,na.rm=T)
svymean(~factor(var4),design = y,na.rm=T)
```

Note that this code should be executed for one variable at a time or the function will output estimates using only those records that have non-missing values across all the variables specified.

To create these estimates for each of the levels of *domainvar*, the following code may be used:

```
svyby(~factor(var1), ~domainvar, svytotal, design=y, na.rm=T)
svyby(~factor(var1), ~domainvar, svymean, design=y, na.rm=T)
```

Assuming the options and `svrepdesign` function above are executed, the following code creates the weighted mean of continuous variable *var5* (using the weight R01_A_PWGT) along with the appropriate standard errors of that estimate (using the replicate weights R01_A_PWGT1 - R01_A_PWGT100):

```
svymean(~var5,design = y,na.rm=T)
```

To create these estimates for each of the levels of *domainvar*, the following code may be used:

```
svyby(~var5,~domainvar, svymean, design=y, na.rm=T)
```

Assuming the options and `svrepdesign` function above are executed, the following code fits a linear regression model using continuous variable *respvar* as the outcome variable and continuous variables *cov1* and *cov2* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100).

For coefficient estimates

```
svyglm(respvar ~ cov1 + cov2, design=y)
```

For standard errors and significance test of model coefficients

```
summary(svyglm(respvar ~ cov1 + cov2, design=y))
```

The following code fits a logistic regression model using dichotomous variable *respvar2* (with values of 0 and 1) as the outcome variable and continuous variables *cov1* and *cov2* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using the replicate weights (R01_A_PWGT1 - R01_A_PWGT100).

For coefficient estimates

```
svyglm(respvar2 ~ cov1 + cov2, design=y, family=binomial)
```

For standard errors and significance test of model coefficients

```
summary(svyglm(respvar2 ~ cov1 + cov2, design=y,
family=binomial))
```

To perform these regressions for respondents with *domainvar* = 1, the design statement should be changed as follows:

```
design=subset(y, domainvar =="1"))
```

SPSS

SPSS does not have the functionality to create variance estimates using the replication method, the preferred method for the PATH Study. As such, only the Taylor series linearization approach to variance estimation may be used with this software package. The appropriate design variables and full-sample weights are relayed to SPSS through a plan file created using the `csplan` function. The following code should be used to create a plan file for use with the PATH Study data:

```
csplan analysis
/plan file="c:\myspace\myplan.csaplan"
/planvars analysisweight=R01_A_PWGT
/print plan
/design strata=VARSTRAT cluster=VARPSU
/estimator type=wr.
```

Assuming the plan file is created as specified above, the following code creates tables including the frequencies of categorical variables *var1*, *var2*, *var3* and *var4*, and estimates of population totals and population proportions for each level of those variables (using the weight R01_A_PWGT) along with the standard errors of these estimates (created using linearization):

```
cstabulate
/plan file="c:\myspace\myplan.csaplan"
/tables variables = var1 var2 var3 var4
/cells popsize tablepct
/statistics count se.
```

Estimates for each of the levels of *domainvar* may be created by including the following statement:

```
/subpop table= domainvar
```

Assuming the plan file is created as specified above, the following code creates the weighted mean of continuous variable *var5* (using the weight R01_A_PWGT) along with the appropriate standard error of that estimate (using linearization):

```
csdescriptives
/plan file="c:\myspace\myplan.csaplan"
/summary variables var5
/mean
```

Estimates for each of the levels of *domainvar* may be created by including the following statement:

```
/subpop table= domainvar
```

Assuming the plan file is created as specified above, the following code fits a linear regression model using continuous variable *respvar* as the outcome variable and continuous variables *cov1* and *cov2* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using linearization.

```
csglm respvar with cov1 cov2
/plan file="c:\myspace\myplan.csaplan"
/model cov1 cov2
/statistics parameter se.
```

The following code fits a logistic regression model using dichotomous variable *respvar2* as the outcome variable and continuous variables *cov1* and *cov2* as the predictors. All parameter estimates are weighted (using the weight R01_A_PWGT) and the standard errors are calculated using linearization.

```
cslogistic respvar2 with cov1 cov2
/plan file="c:\myspace\myplan.csaplan"
/model cov1 cov2
/statistics parameter se.
```

To perform either of these regressions for respondents with *domainvar* = 1, the following statement may be inserted:

```
/domain variable domainvar (1)
```

Questionnaire Variables Excluded B

Table B-1. Questionnaire variables excluded from the Wave 1 data files

Variable name	Instrument	Variable description
R01_AM0001	Adult interview	Date of birth
R01_AM0002	Adult interview	Age (in years) (asked only if date of birth not provided)
R01_AM0016	Adult interview	Main job title or occupation
R01_AX0217_1A	Adult interview	Tobacco Product 1, bar code scan
R01_AX0217_2A	Adult interview	Tobacco Product 2, bar code scan
R01_AX0217_3A	Adult interview	Tobacco Product 3, bar code scan
R01_AX0217_4A	Adult interview	Tobacco Product 4, bar code scan
R01_AX0217_5A	Adult interview	Tobacco Product 5, bar code scan
R01_AX0217_6A	Adult interview	Tobacco Product 6, bar code scan
R01_AX0217_7A	Adult interview	Tobacco Product 7, bar code scan
R01_AX0217_8A	Adult interview	Tobacco Product 8, bar code scan
R01_AX0217_9A	Adult interview	Tobacco Product 9, bar code scan
R01_AX0217_10A	Adult interview	Tobacco Product 10, bar code scan
R01_CPT07	Parent interview	Youth date of birth
R01_PM0021	Parent interview	First name of sibling that youth is a twin of
R01_PM0030	Parent interview	First names of siblings in multiple birth that are identical to youth
R01_PM0035	Parent interview	First names of siblings that are in the multiple birth with youth

Variables with Outlier Values Coded

C

Table C-1. Wave 1 questionnaire variables with outlier values coded

Variable Name	Data File	Variable Description
R01_AC1006	Adult	Age when first smoked part or all of a cigarette
R01_AC1007	Adult	Age when first started smoking cigarettes fairly regularly
R01_AC1020	Adult	Age when first started smoking cigarettes every day
R01_AC1009_NN	Adult	How long since you completely quit smoking cigarettes - Number
R01_AC9002_YR	Adult	How long smoking / smoked cigarettes fairly regularly - Years
R01_AC9002_MO	Adult	How long smoking / smoked cigarettes fairly regularly - Months
R01_AC9003_NN	Adult	How many cigarettes smoked per day when you smoked fairly regularly - Number
R01_AC1021_NN	Adult	Average number of cigarettes now smoked each day - Number
R01_AC1022	Adult	Number of days smoked cigarettes in past 30 days
R01_AC1023_NN	Adult	Average number of cigarettes smoked per day on days smoked in past 30 days - Number
R01_AC9005_NN	Adult	Average number of cigarettes smoked per day when smoked fairly regularly in the past - Number
R01_AC1024_NN	Adult	Time to first cigarette after waking - Number
R01_AC9006_NN	Adult	Average number of cigarettes smoked per day 12 months ago - Number
R01_AC1051RY_NN	Adult	How long been smoking regular brand of roll-your-own cigarette tobacco - Number
R01_AC1041_D	Adult	Amount usually paid for a carton of cigarettes - Dollars
R01_AC1041_C	Adult	Amount usually paid for a carton of cigarettes - Cents
R01_AC1042_D	Adult	Amount usually paid for a pack of cigarettes - Dollars
R01_AC1042_C	Adult	Amount usually paid for a pack of cigarettes - Cents
R01_AC1043_D	Adult	Amount usually paid for a single cigarette - Dollars
R01_AC1043_C	Adult	Amount usually paid for a single cigarette - Cents
R01_AC1051MC_NN	Adult	How long smoked regular brand of cigarettes - Number
R01_AE1006	Adult	Age when first time used an e-cigarette, even one or two times
R01_AE1009_NN	Adult	How long since you took last drag from an e-cigarette - Number
R01_AE9029_NN	Adult	Time since you took last puff from an e-cigarette - Number
R01_AE1024_NN	Adult	Time to first e-cigarette puff after waking - Number
R01_AE1051_NN	Adult	How long used regular brand of e-cigarettes - Number
R01_AG1006TC	Adult	Age when first smoked part or all of a traditional cigar, even one or two puffs
R01_AG1007TC	Adult	Age when first started smoking traditional cigars fairly regularly
R01_AG1009TC_NN	Adult	How long since you last smoked a traditional cigar - Number
R01_AG1022TC	Adult	Number of days smoked traditional cigars in past 30 days
R01_AG9029TC_NN	Adult	Time since you last smoked traditional cigars - Number
R01_AG1024TC_NN	Adult	Time to first traditional cigar after waking on days smoked - Number
R01_AG1098TC	Adult	Number of times used a coupon to buy traditional cigars in past 30 days
R01_AG1051TC_NN	Adult	How long smoked regular brand of traditional cigar - Number

Table C-1. Wave 1 questionnaire variables with outlier values coded (continued)

Variable Name	Data File	Variable Description
R01_AG1006CG	Adult	Age when first time smoked part or all of a cigarillo, even one or two puffs
R01_AG1007CG	Adult	Age when first started smoking cigarillos fairly regularly
R01_AG1009CG_NN	Adult	How long since you last smoked a cigarillo - Number
R01_AG1022CG	Adult	Number of days smoked cigarillos in past 30 days
R01_AG9029CG_NN	Adult	Time since you last smoked cigarillos - Number
R01_AG1024CG_NN	Adult	Time to first cigarillo after waking on days smoked - Number
R01_AG1051CG_NN	Adult	How long smoked regular brand of cigarillo - Number
R01_AG1006FC	Adult	Age when first smoked part or all of a filtered cigar, even one or two puffs
R01_AG1007FC	Adult	Age when first started smoking filtered cigars fairly regularly
R01_AG1020FC	Adult	Age when first started smoking filtered cigars every day
R01_AG1009FC_NN	Adult	How long since you last smoked a filtered cigar - Number
R01_AG1022FC	Adult	Number of days smoked filtered cigars in past 30 days
R01_AG9029FC_NN	Adult	Time since you last smoked filtered cigars - Number
R01_AG1024FC_NN	Adult	Time to first filtered cigar after waking on days smoked - Number
R01_AG1051FC_NN	Adult	How long smoked regular brand of filtered cigar - Number
R01_AP1006	Adult	Age when first smoked part or all of a pipe filled with tobacco, even one or two puffs
R01_AP1007	Adult	Age when first started smoking a pipe filled with tobacco fairly regularly
R01_AP1009_NN	Adult	How long since you last smoked a pipe filled with tobacco - Number
R01_AP1022	Adult	Number of days smoked a pipe filled with tobacco in past 30 days
R01_AP9029_NN	Adult	Time since you last smoked a pipe filled with tobacco - Number
R01_AP1024_NN	Adult	Time to first pipe bowl after waking on days smoked - Number
R01_AH1006	Adult	Age when first smoked hookah, even one or two puffs
R01_AH1007	Adult	Age when first started smoking hookah fairly regularly
R01_AH1009_NN	Adult	How long since you last smoked a hookah - Number
R01_AH1024_NN	Adult	Time to first puff from hookah after waking on days smoked - Number
R01_AH1051_NN	Adult	How long used regular brand of hookah tobacco - Number
R01_AS1006SU	Adult	Age when first used snus pouches, even one or two times
R01_AS1009SU_NN	Adult	How long since you last used snus pouches - Number
R01_AS9029SU_NN	Adult	Time since you last used snus pouches - Number
R01_AS1024SU_NN	Adult	Time to first snus pouch after waking on days used - Number
R01_AS1051SU_NN	Adult	How long used regular brand of snus pouches - Number
R01_AS1006SM	Adult	Age when first used smokeless tobacco, even one or two times
R01_AS1007SM	Adult	Age when first started using smokeless tobacco fairly regularly
R01_AS1009SM_NN	Adult	How long since you last used smokeless tobacco - Number
R01_AS9029SM_NN	Adult	Time since you last used smokeless tobacco - Number
R01_AS1024SM_NN	Adult	Time to first use of smokeless tobacco on days used - Number
R01_AS1051SM_NN	Adult	How long used regular brand of smokeless tobacco - Number
R01_AD1006	Adult	Age when first used dissolvable tobacco, even one or two times
R01_AD1009_NN	Adult	How long since you last used dissolvable tobacco - Number
R01_AD9029_NN	Adult	Time since you last used dissolvable tobacco - Number
R01_AD1051_NN	Adult	How long used regular brand of dissolvable tobacco - Number
R01_AY0010_NN	Adult	Time to first use of any tobacco product after waking - Number

Table C-1. Wave 1 questionnaire variables with outlier values coded (continued)

Variable Name	Data File	Variable Description
R01_AN0130_NN	Adult	Length of time you stopped smoking / using tobacco product(s) because you were trying to quit, in the past 12 months - Number
R01_AN0251_NN	Adult	How long been using nicotine medication to help quit - Number
R01_AN0309	Adult	Number of nicotine patches, gum, inhaler, nasal spray, lozenges or pills used today / yesterday / the day before yesterday
R01_AN0175_NN	Adult	How long ago did you stop using nicotine medication - Number
R01_AN0252_NN	Adult	How long have you been using a prescription drug to help quit - Number
R01_AN0205_NN	Adult	How long ago you stopped using prescription medication - Number
R01_AN0130E_NN	Adult	Length of time you stopped smoking / using tobacco product(s) because you were trying to quit, in the past 12 months - Number
R01_AX0068	Adult	Number of hours in past 7 days that you were in close contact with others when they were smoking
R01_AX0155	Adult	Number of times you visited an emergency room or urgent care center for a health problem of your own in past 12 months
R01_AX0114	Adult	Age when you were first told you had high blood pressure
R01_AX0115	Adult	Age when you were first told you had high cholesterol
R01_AX0116	Adult	Age when you were first told you had congestive heart failure
R01_AX0117	Adult	Age when you were first told you had a stroke
R01_AX0112	Adult	Age when you were first told you had a heart attack
R01_AX0120	Adult	Age when you were first told you had COPD
R01_AX0121	Adult	Age when you were first told you had chronic bronchitis
R01_AX0123	Adult	Age when you were first told you had emphysema
R01_AX0124	Adult	Age when you were first told you had asthma
R01_AX0131	Adult	Age when you were first told you had gum disease
R01_AX0133	Adult	Age when you were first told you had pre-cancerous oral lesions
R01_AX0280	Adult	Age when you were first told you had diabetes, sugar diabetes, high blood sugar or borderline diabetes
R01_AX0143	Adult	Age when you were first told you had an ulcer
R01_AX0148	Adult	Age when you were first told you had stomach or gastro-intestinal bleeding
R01_AX0150	Adult	Age when you were first told you had osteoporosis
R01_AX0198	Adult	Age when you were first told you had a bone fracture because you have fragile bones
R01_AX0152	Adult	Age when you were first told you had a cataract or glaucoma
R01_AX0146_01	Adult	Age when bladder cancer was first diagnosed
R01_AX0146_04	Adult	Age when brain cancer was first diagnosed
R01_AX0146_06	Adult	Age when cervix (cervical) cancer was first diagnosed
R01_AX0146_07	Adult	Age when colon cancer was first diagnosed
R01_AX0146_09	Adult	Age when gallbladder cancer was first diagnosed
R01_AX0146_10	Adult	Age when kidney cancer was first diagnosed
R01_AX0146_16	Adult	Age when melanoma was first diagnosed
R01_AX0146_17	Adult	Age when mouth / tongue / lip cancer was first diagnosed
R01_AX0146_21	Adult	Age when prostate cancer was first diagnosed
R01_AX0146_24	Adult	Age when skin (unknown kind) cancer was first diagnosed
R01_AX0146_30	Adult	Age when uterus (uterine) cancer was first diagnosed
R01_AX0086	Adult	Age when first drank alcohol at all, counting small tastes or sips
R01_AX0074	Adult	Age when first alcoholic drink was consumed

Table C-1. Wave 1 questionnaire variables with outlier values coded (continued)

Variable Name	Data File	Variable Description
R01_AX0079	Adult	Age when first used marijuana, hash, THC or grass
R01_AX0082_01	Adult	Age when first started using: Ritalin or Adderall
R01_AX0082_02	Adult	Age when first started using: Painkillers, sedatives or tranquilizers
R01_AX0082_03	Adult	Age when first started using: Cocaine or crack
R01_AX0082_04	Adult	Age when first started using: Stimulants like methamphetamine or speed
R01_AX0082_05	Adult	Age when first started using: Any other drugs like heroin, inhalants, solvents, or hallucinogens
R01_AX0137_NN	Adult	Weeks / Months pregnant - Number
R01_AX0309	Adult	Calendar year of most recent pregnancy
R01_AN0145_NN	Adult	Longest time period for which you stopped smoking / using tobacco product(s) because you were trying to quit, in the past 12 months - Number
R01_AN0145E_NN	Adult	Longest time period for which you stopped using e-cigarettes because you were trying to quit, in the past 12 months - Number
R01_AN0165_NN	Adult	Length of time used nicotine medication during last tobacco product(s) quit attempt - Number
R01_AN0165E_NN	Adult	Length of time used nicotine medication during last e-cigarette quit attempt - Number
R01_AN0195_NN	Adult	Length of time used prescription medication during last tobacco product(s) quit attempt - Number
R01_AN0195E_NN	Adult	Length of time used prescription medication during last e-cigarette quit attempt - Number
R01_AN0135	Adult	Approximate end date of last tobacco product(s) quit attempt
R01_PT0007_FT	Youth	Youth's current height (feet)
R01_PT0007_IN	Youth	Youth's current height (inches)
R01_PT0008_LB	Youth	Youth's current weight (pounds)
R01_PT0043	Youth	Age youth was first told he/she has high cholesterol
R01_PT0038	Youth	Age youth was first told he/she has asthma
R01_PT0042	Youth	Age youth was first told by a doctor or other health professional that he/she has diabetes, sugar diabetes, high blood sugar or borderline diabetes
R01_YC1006	Youth	Age when first tried cigarette smoking, even one or two puffs
R01_YC1022	Youth	Number of days smoked cigarettes in past 30 days
R01_YC1051_NN	Youth	How long smoked regular brand of cigarettes - Number
R01_YE1006	Youth	Age when first tried an e-cigarette, even one or two times
R01_YE1022	Youth	Number of days used an e-cigarette in past 30 days
R01_YG1006TC	Youth	Age when first tried a traditional cigar, even one or two puffs
R01_YG1022TC	Youth	Number of days smoked a traditional cigar in past 30 days
R01_YG1006CL	Youth	Age when first tried a cigarillo, even one or two puffs
R01_YG1022CL	Youth	Number of days smoked a cigarillo in past 30 days
R01_YG1006FC	Youth	Age when first tried a filtered cigar, even one or two puffs
R01_YP1006	Youth	Age when first tried pipe tobacco, even one or two puffs
R01_YH1006	Youth	Age when first tried smoking a hookah, even one or two puffs
R01_YS1006SU	Youth	Age when first tried snus pouches, even one or two times
R01_YD1006	Youth	Age when first tried a dissolvable tobacco product, even one or two times
R01_YB1006BD	Youth	Age when first tried a bidi, even one or two puffs
R01_YB1006KK	Youth	Age when first tried a kretek, even one or two puffs

Table C-1. Wave 1 questionnaire variables with outlier values coded (continued)

Variable Name	Data File	Variable Description
R01_YX0086	Youth	Age when first drank alcohol at all
R01_YX0074	Youth	Age when consumed first alcoholic drink
R01_YX0079	Youth	Age when first used marijuana, hash, THC or grass
R01_YX0082_01	Youth	Age when first used: Ritalin or Adderall
R01_YX0082_02	Youth	Age when first used: Painkillers, sedatives or tranquilizers
R01_YX0082_03	Youth	Age when first used: Cocaine or crack
R01_YX0082_04	Youth	Age when first used: Stimulants like methamphetamine or speed
R01_YX0082_05	Youth	Age when first used: Any other drugs like heroin, inhalants, solvents or hallucinogens

Example Program Code for Linking Adult and Youth/Parent Data Files

D

This appendix contains example SAS, Stata, R, and SPSS program code for linking the adult and youth/parent data. Note that text in italics represents placeholders for actual dataset and variable names. These examples are provided as a means to further illustrate the data structure. They are not meant to be exhaustive or to provide instruction for users unfamiliar with a particular software package.

SAS

The example below illustrates how to link the adult and youth/parent data so to create the dataset *Parent_Adult_Resps* with the parent interview respondents who were also respondents to the adult interview. In the example, *Youthdata* is sorted by R01_PARENT_PERSONID and *Adultdata* is sorted by PERSONID.

```
data Parent_Adult_Resps;
merge Youthdata (rename=(R01_PARENT_PERSONID = PERSONID)
in=inyouth keep= R01_PARENT_PERSONID var1 var2 var3) Adultdata
(in=inadult keep= PERSONID var4 var5 var6);
by PERSONID;
if inyouth and inadult;
run;
```

Similar code replacing R01_PARENT_PERSONID with R01_PT0046_PERSONID will create a dataset with adult interview data for the spouses of the parent interview respondents.

The example below provides a dataset with adult interview data from respondents living in the same household as respondents to the youth interview. The dataset *Youth_HHs* is created first and contains a unique list of household IDs from the dataset *Youthdata* and *Adultdata* is sorted by R01_HHID.

```
proc sort data=Youthdata out=Youth_HHs (keep=R01_HHID) nodupkey;
by R01_HHID;
run;

data Adult_Resps_in_Youth_HHs;
merge Youth_HHs (in=inyouth) Adultdata (in=inadult keep=R01_HHID
PERSONID var1 var2 var3);
```

```
by R01_HHID;
if inyouth and inadult;
run;
```

Stata

The example below illustrates how to link the adult and youth/parent data so to create the dataset *Parent_Adult_Resps* with the parent interview respondents (in dataset *Youthdata*) who were also respondents to the adult interview. In the example, *Adultdata* is sorted by PERSONID.

```
use c:\myspace\Youthdata, clear
rename R01_PARENT_PERSONID personid
sort personid

save c:\myspace\Youthdata2

merge 1:1 personid using c:\myspace\Adultdata

save c:\myspace\Parent_Adult_Resps
```

Similar code replacing R01_PARENT_PERSONID with R01_PT0046_PERSONID will create a dataset with adult interview data for the spouses of the parent interview respondents.

The example below provides a dataset with adult interview data from respondents living in the same household as respondents to the youth interview. The dataset *Youth_HHs* is created first and contains a unique list of household IDs from the dataset *Youthdata* and *Adultdata* is sorted by R01_HHID.

```
use c:\myspace\Youthdata, clear
sort R01_HHID
quietly by R01_HHID: gen dup = cond(_N==1, 0, _n)
drop if dup>1

save c:\myspace\Youth_HHs

merge 1:1 R01_HHID using c:\myspace\Adultdata

save c:\myspace\Adult_Resps_in_Youth_HHs
```

R

The example below illustrates how to link the adult and youth/parent data so to create the dataset *Parent_Adult_Resps* with the parent interview respondents who were also respondents to the adult interview. In the example, *Youthdata* is sorted by R01_PARENT_PERSONID and *Adultdata* is sorted by PERSONID. Note that variables in italics are placeholders for any analysis variables of interest.

```
Youthtemp <- Youthdata[c("R01_PARENT_PERSONID", "var1", "var2",
"var3")]
names(Youthtemp)[names(Youthtemp) == 'R01_PARENT_PERSONID'] <-
'PERSONID'

Adulttemp <- Adultdata[c("PERSONID", "var4", "var5", "var6")]

Parent_Adult_Resps <- merge(Youthtemp, Adulttemp, by =
'PERSONID')
```

Similar code replacing R01_PARENT_PERSONID with R01_PT0046_PERSONID will create a dataset with adult interview data for the spouses of the parent interview respondents.

The example below provides a dataset with adult interview data from respondents living in the same household as respondents to the youth interview. The dataset *Youth_HHs* is created first and contains a unique list of household IDs from the dataset *Youthdata* and *Adultdata* is sorted by R01_HHID.

```
Youth_HHs <-
subset(Youthdata["R01_HHID"],!duplicated(Youthdata$R01_HHID))

Adulttemp <- Adultdata[c("R01_HHID", "PERSONID", "var1", "var2",
"var3")]

Adult_Resps_in_Youth_HHs <- merge(Youth_HHs, Adulttemp, by =
'R01_HHID')
```

SPSS

The example below illustrates how to link the adult and youth/parent data so to create the dataset *Parent_Adult_Resps* with the parent interview respondents who were also respondents to the adult interview. In the example, *Adultdata* is sorted by PERSONID.

```
get file = "c:\myspace\Youthdata.sav".
```

```
rename variables (R01_PARENT_PERSONID = personid).
sort cases by personid.

save outfile = "c:\myspace\Youthdata2.sav".

match files file="c:\myspace\Adultdata.sav" /in=inparent
/file="c:\myspace\Youthdata2.sav" /in=inyouth
/by personid.
select if inparent and inyouth.
save outfile="c:\myspace\Parent_Adult_Resps.sav".
```

Similar code replacing R01_PARENT_PERSONID with R01_PT0046_PERSONID will create a dataset with adult interview data for the spouses of the parent interview respondents.

The example below provides a dataset with adult interview data from respondents living in the same household as respondents to the youth interview. The dataset *Youth_HHs* is created first from the dataset *Youthdata* by sorting by household ID and *Adultdata* is sorted by R01_HHID.

```
get file = "c:\myspace\Youthdata.sav".
SORT CASES BY R01_HHID.

save outfile = "c:\myspace\Youth_HHs.sav".

match files file="c:\myspace\Adultdata.sav" /in=inadult
/file="c:\myspace\Youth_HHs.sav" /in=inyouth
/by R01_HHID
/FIRST=by1st.

select if inadult and inyouth and by1st=1.

save outfile="c:\myspace\Adult_Resps_in_Youth_HHs.sav".
```