HHS Public Access

Author manuscript

Tob Regul Sci. Author manuscript; available in PMC 2021 April 14.

Published in final edited form as:

Tob Regul Sci. 2021 January ; 7(1): 3–16. doi:10.18001/trs.7.1.1.

Longitudinal Uses of the Population Assessment of Tobacco and Health Study

Andrea Piesse, PhD [Senior Statistician],
Statistics and Evaluation Sciences, Westat, Rockville, MD, United States.

Jean Opsomer, PhD [Vice President],
Statistics and Evaluation Sciences, Westat, Rockville, MD, United States.

Sylvia Dohrmann, MS [Associate Director],
Statistics and Evaluation Sciences, Westat, Rockville, MD, United States.

Ralph DiGaetano, MA [Senior Statistician],
Statistics and Evaluation Sciences, Westat, Rockville, MD, United States.

David Morganstein, MA [Nonresident Senior Statistical Fellow],
Statistics and Evaluation Sciences, Westat, Rockville, MD, United States.

Kristie Taylor, PhD [Associate Director],
Behavioral Health and Health Policy, Westat, Rockville, MD, United States.

Charles Carusi, PhD [Vice President],
Behavioral Health and Health Policy, Westat, Rockville, MD, United States.

Andrew Hyland, PhD [Chair]
Department of Health Behavior, Roswell Park Comprehensive Cancer Center, Buffalo, NY, United States.

Abstract

Objectives: The Population Assessment of Tobacco and Health (PATH) Study is a nationally representative study of the US population on tobacco use and its effects on health, with 3 waves of data collection between 2013 and 2016. Prior work described the methods of the first wave. In this paper, we describe the methods of the subsequent 2 waves and provide recommendations for how to conduct longitudinal analyses of PATH Study data.

Methods: We use standard survey quality metrics to evaluate the results of the follow-up waves of the PATH Study. The recommendations and examples of longitudinal and cross-sectional analyses of PATH Study data follow a design-based statistical inference framework.

Results: The quality metrics indicate that the PATH Study sample of approximately 40,000 continuing respondents remains representative of its target population. Depending on the intended analysis, different survey weights may be appropriate.

Correspondence Dr Piesse; andreapiesse@westat.com.

Conflict of Interest Disclosure Statement
The authors have no conflicts of interest, financial or otherwise.

Conclusion: The PATH Study data are a valuable resource for regulatory scientists interested in longitudinal analysis of tobacco use and its effects on health. The availability of multiple sets of specialized survey weights enables researchers to target a wide range of tobacco-related analytic questions.

Keywords

longitudinal study; nonresponse bias; response rates; survey data; survey weights

The Population Assessment of Tobacco and Health (PATH) Study is a nationally representative, longitudinal cohort study of the US population on tobacco use and its effects on health. Data collected in the PATH Study help inform the Food and Drug Administration (FDA) tobacco regulatory activities, including understanding the impact of its actions. Data collection started in 2013. Interviews are conducted with tobacco users and nonusers ages 12 and older, and provide data about a wide range of tobacco products. Biomarkers of exposure and potential harm are measured in blood and urine specimens from a subsample of adult respondents.

The PATH Study is based on the Host, Agent, Vector, Environment (HAVE) conceptual model described by Hyland et al.[1] This model conceptualizes the connections among various factors (eg, variables relating to individuals, tobacco products, tobacco manufacturers, and the broader environment) and health outcomes. Hyland et al also describe the design and implementation of Wave 1 of the study, which involved the recruitment of the initial cohort in 2013-2014. Since that time, additional waves of data were collected, including Wave 2 in 2014-2015 and Wave 3 in 2015-2016.

Now that multiple waves of interviews have been conducted, the study is uniquely positioned to assess patterns of tobacco product use, tobacco exposures, health, and risks of disease over time. However, the availability of these longitudinal data also raises new issues with regard to weighting and interpretation of response rates over time, both of which are critical for conducting statistical analyses of the study data and interpreting results. In the current paper, we describe the response characteristics of Waves 2 and 3 and provide examples and recommendations for how to use PATH Study weights.

Changes to the sample design of the study were introduced in Wave 4 (2016-2017). These involved the selection of a supplementary probability sample of adults, youth, and “shadow youth” ages 10 to 11 from the US civilian, noninstitutionalized population (CNP) at the time of Wave 4. This “replenishment sample” supplemented the sample of study participants selected at Wave 1, to address attrition over time and allow for the inclusion of new entrants to the population of inference since the time of Wave 1. The additional issues raised in the computation and interpretation of response characteristics for Wave 4 and subsequent waves, and in the construction and use of survey weights, are sufficiently complex to warrant separate attention. They will be addressed in future work.

METHODS

PATH Study Sample Design Features

The PATH Study is a prospective study of tobacco use and associated health outcomes. Prospective studies collect baseline exposure data on all participants, and then measure outcomes at a later time. A retrospective study examines exposures in relation to an already established outcome. Prospective studies usually have fewer potential sources of bias and improved ability to control for confounding variables, compared to retrospective studies. The ability to control for confounders is especially important for causal inference, for example, concerning the effects of tobacco use on health or the effects of environmental and regulatory factors on tobacco-use behavior. A disadvantage of prospective studies is that it is not possible to control the sample sizes of cases of high interest as precisely as in retrospective studies. These advantages and disadvantages are widely accepted in the medical and epidemiological research communities.[2,3]

We briefly review the design and sampling results of Wave 1, because Waves 2 and 3 are follow-up efforts to collect data from those who responded at Wave 1. The target population (ie, the population to which the results are intended to generalize) consists of individuals age 9 or older in the CNP at the time of Wave 1. A nationally representative household sample was selected through a multi-stage stratified area probability design. At the first stage, a stratified sample of 156 geographical primary sampling units (PSUs) was selected, with each PSU comprised of one or several adjacent counties. Within the selected PSUs, smaller geographical areas, referred to as segments, were sampled at the second stage. Later stages sampled mailing addresses in the selected segments, and individuals from households identified at these addresses through a brief interview, known as the household screener. A 2-phase selection procedure was used at the final stage for adults to ensure sufficient representation of tobacco users in the sample. More details on the Wave 1 PATH Study design can be found in Hyland et al[1] and in the *PATH Study Restricted Use Files User Guide*[4] (available at the National Addiction and HIV Data Archive Program [NAHDAP] website at http://doi.org/10.3886/ICPSR36231.userguide).

At Wave 1, 45,971 persons (including 32,320 adults ages 18 and older and 13,651 youth ages 12 to 17) were interviewed. In addition, a “shadow sample” of 7207 youth ages 9 to 11 was established. The purpose of these shadow youth was to ensure that there were youth ages 12 and older in the next 3 waves without having to draw a new sample from the population at every wave.

The Wave 2 and Wave 3 follow-up interviews built on the information collected from adults and youth during their baseline interview at Wave 1. Information such as demographic characteristics (eg, sex and race) was collected only at baseline. Similarly, information on lifetime use of tobacco products up to the time of the baseline interview was not requested again. The follow-up interviews updated information on the use of tobacco products and health outcomes in the past 12 months and asked about the use of new products. Once the Wave 1 shadow youth turned age 12, they were asked to complete a baseline interview. This interview asked demographic and lifetime health and tobacco-use questions similar to those asked of youth in the Wave 1 interview. Similarly, when youth turned age 18 and completed

an adult interview, they were asked additional lifetime health and tobacco-use questions not covered in the youth interview.

The data were collected in person, using audio computer-assisted self-interviewing (ACASI) instruments (separate for youth and adults) because this mode of interview administration has been shown to improve response to sensitive questions,[5] and a computer-assisted personal-interviewing (CAPI) parent instrument. Table 1 summarizes data collection for the first 3 waves.

There was no additional sampling for Wave 2 or Wave 3. All Wave 1 respondents were eligible for Wave 2 if they continued to live in the US and were not incarcerated. They were eligible for Wave 3 if they lived in the US and were not incarcerated, even if they were ineligible or did not participate in the study at Wave 2.

Issues of eligibility over time require careful consideration in the longitudinal context and have repercussions on the target populations of inference. Starting with Wave 1 respondents, statistically representative of the Wave 1 target population, only those eligible at the time of the subsequent Wave 2 or Wave 3 interview could participate in that wave. Conceptually, this means that *the (longitudinal) target population at a follow-up wave consists of those persons who are eligible at that wave: in the CNP age 9 or older at the time of Wave 1 and living in the US and not incarcerated at the time of the corresponding follow-up wave*. In the next subsection, we describe the impact of this target population definition on weighting and estimation.

Weighting and Estimation

Three years of PATH Study data, covering the period 2013-2016, are available for most study participants at the completion of Wave 3 (Table 1). Some Wave 1 respondents did not respond in Wave 2 or Wave 3, and shadow youth interviewed for the first time in these waves do not have data in years prior. If we ignore individuals missing data for one or more waves, the longitudinal data are statistically representative of the behavior over this period of the target population, which is comprised of the CNP ages 9 and older at the time of Wave 1 and adjusted for eligibility changes at later waves (as noted above).

Because study participants were selected in Wave 1 through a complex sampling design and are subject to nonresponse, valid statistical analyses of the data need to account for these selection effects. The most common and generally recommended approach to do this is to employ survey weights in estimation and a design-based variance estimation method. This is explained by Hyland et al[1] in the context of the Wave 1 data, but here we describe how this is implemented for data from Waves 1-3.

Absent additional nonresponse and eligibility changes in Waves 2 and 3, the creation of estimation weights for analysis of longitudinal data would be straightforward, because the Wave 1 weight could be used in the later waves. However, in the presence of nonresponse and eligibility changes across waves, determining suitable longitudinal estimation weights for the PATH Study is more complicated, because it is dependent on analytic goals. The simplest approach, which is also the appropriate one if the analysis requires data from all 3

waves, is to remove study participants who exhibited nonresponse or were ineligible at any wave. After also removing the Wave 1 shadow youth, the sample size is 34,716 (the number of participants with interview data for every wave), as Table 2 shows. Note that shadow youth from Wave 1 began providing data at follow-up waves and contribute to the number of interview respondents in Waves 2 and 3.

For the PATH Study, a weight was created for estimating characteristics across 3 waves. This weight is referred to as the Wave 3 "all-waves" weight. For example, the 34,716 respondents with data at all 3 waves have a Wave 3 all-waves weight. The target population for an analysis of these respondents is the CNP age 12 or older at the time of Wave 1, still living in the US and not incarcerated at the time of Wave 2 nor at the time of Wave 3. In addition, the shadow youth who turned age 12 in Wave 2 and who responded at both Waves 2 and 3 also receive an all-waves weight, bringing the number to 36,663 as shown in the last row of Table 2. Even though they do not have data at each wave, these respondents can be used in analyses involving only Waves 2 and 3, for which the all-waves weight also applies. The target population for an analysis of these 36,663 respondents is subtly different from that for the 34,716 all-waves respondents: the Wave 1 CNP age 12 or older *at the time of Wave 2*, still living in the US and not incarcerated at the time of Wave 2 nor at the time of Wave 3. (The total number of participants with a Wave 3 all-waves weight is 38,561, which includes those shadow youth at Waves 1 and 2 whose parents updated contact information with the study. This scenario is not covered in Table 2.)

Other weighting scenarios are possible. For instance, if one is interested in changes in the characteristics or behavior of individuals between Wave 1 and Wave 3, then a dataset that includes only participants who responded and were eligible in all 3 waves needlessly removes those whose only nonresponse or ineligibility was in Wave 2. Therefore, 2 additional sets of weights were created for analyses that include only 2 waves: one for all participants who were interview respondents and eligible in Waves 1 and 2 (38,443 respondents) and one for all who were interview respondents and eligible in Waves 1 and 3 (35,969 respondents). These sets of weights are referred to as the Wave 2 and Wave 3 "single-wave" weights, respectively, because they compare a single follow-up wave with Wave 1.

The single-wave weights cannot be used for longitudinal analyses of Wave 1 shadow youth because these participants did not complete an interview at Wave 1. The Wave 1 shadow youth were nevertheless assigned a single-wave weight for follow-up waves at which they completed an interview because these weights can be used for cross-sectional estimation for Waves 2 and 3 by including Wave 1 shadow youth (for a total of 40,534 Wave 2 interview respondents and 39,962 Wave 3 interview respondents). We return to cross-sectional estimation for Waves 2 and 3 in our later discussion.

The Wave 1 weights account for the PATH Study sampling design and Wave 1 nonresponse. For example, the Wave 1 weights adjust for the oversampling of adult tobacco users. Oversampling is used to enrich the sample with cases having characteristics that are of particular interest, and this is especially useful for longitudinal analyses. For the PATH Study, the oversampling increases sample size for analyses of adult tobacco users. An

unweighted analysis of PATH Study data will produce estimates of tobacco use that reflect the disproportionately large number of tobacco users in the responding sample, rather than the underlying population, and so, can be misleading. Therefore, the use of the weights is essential for inference about the target population. The Wave 1 weights were calibrated based on socio-demographic population characteristics available from the US Census Bureau to improve the precision of the PATH Study estimates. Specifically, the 2013 one-year Public Use Microdata Sample data from the American Community Survey (ACS) were used for this purpose. Intermediary weights at the household level were calibrated by census region and household composition (number of adults and number of non-adult persons), and used as the starting point for the person-level weights. The adult weights were calibrated using combinations of census region, age, race/ethnicity, sex, and educational attainment. For youth (and shadow youth), the weights were calibrated using combinations of census region, single year of age, race/ethnicity, and sex.

The creation of the 3 sets of Wave 2 and 3 weights started from these Wave 1 weights, then adjusted for additional nonresponse in the relevant wave(s) and applied longitudinal calibration adjustments to account for changes in the responding sample composition across waves. The weights for Waves 2 and 3 were calibrated using the population-based estimates used to calibrate the Wave 1 weights, as well as Wave 1 sample-based estimates of tobacco and e-cigarette use. The sample-based calibration improves estimation of changes in tobacco use and related outcomes over time. More details on the nonresponse and calibration adjustments can be found in the *PATH Study Restricted Use Files User Guide*.[4]

For each PATH Study weight, 100 associated replicate weights are provided for inference. Replicate variance estimation is a common approach in large-scale surveys. A number of methods are described in the statistical literature.[6] Variance estimation with replicate weights is performed by calculating estimates of interest using each of the replicate weights and taking the average of the squared deviations between the estimates with the full-sample weights and those with each of the replicate weights. For many types of estimates and analyses, this variance calculation is performed automatically by standard survey software packages. The balanced repeated replication (BRR) method was selected for the PATH Study, reflecting the highly stratified selection of PSUs. The study uses a variant of BRR known as Fay's method.[7]

Also included with the PATH Study data are variables for pseudo-strata and pseudo-PSUs that reflect the variance structure for analysts without access to software packages that support replication techniques, or who prefer to use the Taylor series (linearization) method. However, variance estimates created using linearization do not fully reflect the impact of the weighting adjustments and may result in inaccurate inferences.

RESULTS

In this section, we discuss response rates and nonresponse bias analyses for Waves 1-3. Table 3 presents weighted response rates for collections in these waves. In accordance with American Association for Public Opinion Research guidance,[8] we computed weighted response rates using the inverse-of-probability-of-selection (IPS) weights. The parental

consent response rate for shadow youth is not directly relevant to the Wave 1 released data (because no data are available for shadow youth) but becomes so once the Wave 1 shadow youth turn age 12 in later waves and begin to provide interview data. The Wave 1 youth interview response rate reflects nonresponse due to lack of parental consent as well as nonresponse on behalf of the youth.

The response rates for Wave 2 are conditional on Wave 1 response, ie, on Wave 1 interview completion for those selected as youth or adults, or Wave 1 parental consent for those selected as shadow youth. Therefore, the Wave 2 response rates reflect attrition between Waves 1 and 2, not cumulative attrition from the time of sampling. Because Wave 2 nonrespondents were fielded for Wave 3, the Wave 3 response rates similarly are conditional on Wave 1 response and reflect attrition across the first 2 follow-up waves since Wave 1. Further details on the response rate calculations for Waves 1-3 appear in the *PATH Study Restricted Use Files User Guide*.[4]

Although nonresponse bias can be a concern in household surveys, participant nonresponse does not necessarily induce nonresponse bias in survey estimates. Some estimates may be unaffected by nonresponse, whereas others can be subject to large biases.[9] The effect of nonresponse biases on estimates can be reduced or even eliminated by incorporating nonresponse adjustments as part of the weighting procedures. The PATH Study's nonresponse bias analysis reports (https://www.icpsr.umich.edu/icpsrweb/NAHDAP/studies/36231/datadocumentation) assess the effects of nonresponse on selected socio-demographic and tobacco-use estimates for Waves 1-3. We briefly review the main findings below. Two types of analyses were performed, one using PATH Study data only, and the other comparing estimates of cigarette smoking with estimates from other national studies.

For the Wave 1 household screener and adult interview, the demographic and socio-economic characteristics of the respondents mostly aligned with estimates from the 2013 ACS when using the IPS weights for the PATH Study. However, there were exceptions for single-person households, sex, education, and ethnicity. In particular, men were underrepresented and persons of Hispanic ethnicity were overrepresented among the Wave 1 adult interview respondents. When the estimates were adjusted for nonresponse using the Wave 1 final weights, they more closely approximated the ACS estimates. For the Wave 1 youth interview, most demographic characteristics of respondents were consistent with the estimates from the 2013 ACS, when using the IPS weights. However, persons of Hispanic ethnicity also were overrepresented among the youth respondents. When adjusted for youth nonresponse, the Wave 1 estimates more closely approximated the 2013 ACS estimates.

Similar to how the response rates were computed for the follow-up waves, the potential for nonresponse bias in Wave 2 and Wave 3 estimates from the PATH Study was evaluated conditional on Wave 1 response. Prior to adjusting for Wave 2 nonresponse, the nonresponse bias analysis showed that many characteristics of Wave 2 interview respondents aligned with those of Wave 2 nonrespondents. However, the exceptions included underrepresentation of men and of adult current established tobacco users at Wave 1 among the Wave 2 adult respondents. After adjusting for nonresponse using the Wave 2 weights, these discrepancies were essentially eliminated.

In a similar analysis for Wave 3, men were notably underrepresented among Wave 3 adult respondents and those recruited as shadow youth at Wave 1 were notably underrepresented among Wave 3 youth respondents, prior to adjusting for nonresponse. After Wave 3 weighting adjustments, the differences between respondents and those eligible for interview were negligible.

The findings of the nonresponse bias analyses are not surprising given that men tend to exhibit lower response propensity than women in most household surveys[10,11] and smokers may be more likely to drop out over the course of a panel survey.[12-14] However, the weighting adjustments at Waves 1-3 proved successful in addressing differences found due to nonresponse.

Weighted estimates of cigarette-smoking behavior for each of the first 3 waves of the PATH Study were compared to weighted estimates from the Tobacco Use Supplement to the Current Population Survey (TUS-CPS), the National Health Interview Survey (NHIS), the National Health and Nutrition Examination Survey (NHANES), the National Survey on Drug Use and Health (NSDUH), and the National Youth Tobacco Survey (NYTS). For PATH Study estimates pertaining to Wave 3, we used the Wave 3 single-wave weight. The estimates from the other national studies were based on weighted data or published figures that most closely corresponded to the time frame of the respective PATH Study wave and were publicly available at the time of the nonresponse bias assessment. The specific citations for these other studies are in the PATH Study's nonresponse bias analysis reports. These reports also provide details about differences between the studies, including mode of administration, question context, proxy responses, and target populations, which may lead to differences in estimates even if none of them is affected by nonresponse bias.

The PATH Study estimates of ever cigarette use for youth in Waves 1-3 were generally lower than the estimates from NSDUH, NHANES, and NYTS. The 95% confidence intervals overlapped between the PATH Study, NHANES, and NSDUH for some of these estimates but the NYTS estimates were consistently the highest. To illustrate some of the differences between these studies, Table 4 summarizes the questionnaire items and cigarette-smoking definitions relevant to each study for the youth estimates. The PATH Study, NHANES, and NSDUH use ACASI for the questions about tobacco use by youth, and these are administered individually in a household or mobile examination center setting. The NYTS is a pencil-and-paper survey that is self-administered in the classroom. Although the estimates for each study were restricted to youth ages 12 to 17, the NYTS includes only public and private school students enrolled in regular middle schools and high schools in grades 6 through 12. The higher estimates observed for NYTS are in line with other research noting higher smoking rates in school-based surveys.[15]

Estimates of adult current cigarette-smoking rates were generally lowest in TUS-CPS and highest in NSDUH. The estimates from NHIS and NHANES were similar to those from the PATH Study, but generally lower and higher, respectively. Table 5 summarizes the questionnaire items and cigarette-smoking definitions relevant to each study for the adult estimates. The PATH Study question used to establish whether an adult has smoked at least 100 cigarettes in their lifetime asks the respondent to choose among ranges specifying the

number of cigarettes smoked, whereas the TUS-CPS, NHIS, and NHANES questions call for a yes/no response with respect to the threshold of 100 cigarettes. As noted above, the PATH Study and NSDUH both use ACASI administration for the tobacco-use questions. By contrast, TUS-CPS, NHIS, and (for adults) NHANES have direct questioning by an interviewer. Moreover, the TUS-CPS allows proxy responses. The cigarette-smoking questions are near the beginning of the PATH Study adult questionnaire. In TUS-CPS, the smoking questions are near the beginning of the adult questionnaire on tobacco, but the survey is administered as part of the CPS. In NHIS, the smoking questions follow a long series of questions on health problems. These and other factors may be associated with differences in responses. However, there is no evidence of nonresponse bias in Waves 1-3 of the PATH Study with respect to current cigarette-smoking behavior among adults, in the sense that the study's estimates are within the range of estimates from other national studies.

Overall, the PATH Study findings were consistent with those of other studies, with differences likely reflecting methodological differences between the studies (such as mode of administration, and question order and context). Assuming that the Wave 1 demographic, socio-economic, and tobacco-use characteristics examined are correlated with key tobacco and health-related outcome measures in Waves 2 and 3 of the PATH Study, these results indicate little if any nonresponse bias in the adult and youth interview estimates due to attrition since Wave 1.

DISCUSSION

Different types of analyses can be performed using the PATH Study data, but care is needed in selecting the appropriate weights and understanding the population of inference. In this section, we provide guidelines and examples of several analyses using the PATH Study Restricted Use Files.

The simplest type of analysis is the estimation of cross-sectional characteristics using data from a single wave of the study. For example, what is the percentage of 12-to-17-year-olds who used cigarettes in the past 30 days? This can be readily computed for any of the 3 waves, but it is important to keep the target population in mind when interpreting the estimates. If computed for Wave 1, this is an estimate of the percentage of past 30-day cigarette users at the time of Wave 1, among individuals who were ages 12 to 17 and in the CNP at the time of Wave 1. At Wave 2, this is an estimate of the percentage of past 30-day cigarette users at the time of Wave 2 among individuals who were in the CNP at the time of Wave 1, and eligible and ages 12 to 17 at the time of Wave 2. This estimate is subtly different from the percentage of individuals in what is likely to be the desired population of interest (ie, the CNP at the time of Wave 2), due to eligibility changes as well as changes in the underlying population, such as recent immigrants. However, because the time between Waves 1 and 2 is approximately one year, the difference between the target population and the desired population of interest is unlikely to have more than a negligible impact on the estimates. Therefore, this type of "pseudo-cross-sectional" interpretation of PATH Study estimates is reasonable, especially if the (minor) population discrepancy is explicitly noted. If such an estimate is computed at Wave 3, the difference between the desired and actual target populations can be expected to increase modestly, but still be negligible. In both cases,

the appropriate weight to use for computing these estimates is the single-wave weight for the relevant wave, as previously described in this paper.

In the example above, the age range 12 to 17 is for the target wave of interest. Thus, most shadow youth recruited at Wave 1 when they were age 11 are included in the Wave 2 estimate, and most youth age 17 in Wave 1 have aged out of the target age range. Similarly, the Wave 3 estimate includes most shadow youth who were age 10 or 11 in Wave 1 and excludes most youth who were age 16 or 17 in Wave 1. Note that not every respondent is exactly one year older at the next wave, because the realities of data collection sometimes prevent contacting individuals exactly one year after their previous contact.

Another common type of analysis involves using multiple waves of data to compute the cross-sectional change over time in the target population. For example, we consider estimating the difference between the percentage of 12-to-17-year-olds who used cigarettes in the past 30 days in Wave 1 versus in Wave 2. PATH Study data can be used for this purpose, in this example using the Wave 1 weight and the Wave 2 single-wave weight, respectively. The estimated percentages of 12-to-17-year-olds who used cigarettes in the past 30 days, with their respective standard errors (calculated using the replicate weights) in parentheses, are 4.58% (0.20) for Wave 1 and 3.95% (0.22) for Wave 2. (Cases with a missing value for Wave 1 age or past 30-day cigarette use were excluded from the Wave 1 estimates; cases with a missing value for Wave 1 age, Wave 2 age, or Wave 2 past 30-day cigarette use were excluded from the Wave 2 estimates.) The estimated change in the percentages between the waves is 0.63% with associated standard error of 0.22%. The estimates for the 2 waves are correlated, due to the overall sampling design (in particular, its multi-stage structure) and the fact that some respondents in the target age range can contribute to both estimates (eg, a 12-year-old in Wave 1 is expected to be a 13-year-old in Wave 2, and thus, is included in both estimates). Correct inference for this type of measure of change needs to account for this correlation, which in this case leads to a reduction in the standard error compared to what would occur if the sets of Wave 1 and Wave 2 respondents were independent. This correlation is accounted for when using the BRR replicate weights provided with the study datasets.

Cross-sectional differences between waves, as illustrated in the previous paragraph, are useful measures of change over time in the target population. A key feature of longitudinal studies is that it is also possible to measure change over time *within the same population group*. Returning to the example, suppose we are interested in past 30-day cigarette use of 12-to-17-year-olds at Wave 1 compared to the behavior of these same individuals at Wave 2 (who are now a year older on average). We can estimate the following 4 quantities: (1) percentage of individuals reporting the behavior in both Wave 1 and Wave 2, (2) percentage of individuals reporting the behavior in Wave 1 but not in Wave 2, (3) percentage of individuals not reporting the behavior in Wave 1 but doing so in Wave 2, and (4) percentage of individuals not reporting the behavior in either Wave 1 or Wave 2. Table 6 shows the weighted estimates of these quantities, using the Wave 2 single-wave weight.

Based on these results, the estimated percentage of 12-to-17-year-olds who used cigarettes in the past 30 days at Wave 1 is 4.58. A year later, when this group is generally age 13 to 18,

the estimated percentage is 7.29, so that the estimated net change is 2.71% over this period. This net change can be decomposed into its 2 gross change components, with an estimated 4.11% initiating past 30-day cigarette use by Wave 2, and an estimated 1.40% ceasing this behavior by Wave 2. Similar analyses could be performed comparing Wave 1 and Wave 3 behavior, or Wave 2 and Wave 3 behavior. In the former case, the Wave 3 single-wave weight would be used, whereas the Wave 3 all-waves weight would be used for the latter.

More sophisticated analyses are possible with PATH Study data, including fitting of regression models to investigate determinants and consequences of tobacco-use behavior over time while adjusting for potential confounders. The generally recommended approach for fitting regression models to data from complex sample designs is to use the survey weights and replicate weights. The reasons for doing so are to protect against possible biases in the model parameter estimates due to informativeness of the sampling design and nonresponse, and to represent the randomness of the sampling design. Bias in the estimates can occur when the true model of interest is “masked” by the selection mechanism that resulted in the observed set of sample respondents. Pfeffermann and Sverchkov[16] use the terminology *sample model* and *population model* to emphasize the effect of the selection; the sample model is the model that describes the characteristics of the respondents in the specific study under consideration, whereas the population model holds for individuals regardless of whether they have been selected or not. The analytic goals are almost always related to the population model, not the sample model.

Differences between the sample model and the population model can be both obvious and subtle. For instance, the observed sample for the PATH Study has, by design, a disproportionately high number of tobacco users, young adults, and black or African-American adults. Additionally, PATH Study data are collected from clusters consisting of relatively nearby housing units to reduce data collection costs, with these clusters randomly but unevenly spread over the country according to the sampling design. This clustering and the oversampling of subgroups are both incorporated into model fitting through the use of the survey weights and appropriate variance estimation methods, so that the analysis correctly targets the desired population model. Korn and Graubard[17] provide an extensive discussion of the advantages and disadvantages of performing weighted and unweighted analyses for data from a complex sample design.

When performing weighted fitting of regression models, the choice of weight depends on which waves are included in the data to be analyzed. As an example, suppose we are interested in determining the probability that a person used cigarettes in the past 30 days at Wave 2, based on their opinions, prior use, and demographic information. A logistic regression model can be specified to address this question and would be fitted using all Wave 2 interview respondents who also have interview data at Wave 1, and the Wave 2 single-wave weight. Table 7 shows the parameter estimates and 95% confidence intervals for such a model. If the analysis time frame were shifted one year to estimate the probability of Wave 3 use based on Wave 2 characteristics, then Wave 1 shadow youth who completed interviews in Waves 2 and 3 could be included, and the Wave 3 all-waves weight would be used in the analysis.

Finally, we note that when models involve individual waves or transitions between pairs of waves, it is possible to increase the available sample size modestly by including all respondents with data in the relevant waves or pairs of waves (as opposed to only those respondents who have data in all 3 waves along with the Wave 3 all-waves weight). In this case, weighting and analysis can be accomplished by concatenating or "stacking" data files from the relevant waves to form a single file. For models involving individual waves, the stacked data file would contain one record per respondent per wave of interest in which they provided data; the corresponding analysis weight would be the Wave 1 weight (on Wave 1 records) and the Wave 2 or Wave 3 single-wave weight (on Wave 2 or Wave 3 records, respectively). For models involving pairs of waves, the stacked data file would contain one record per respondent per pair of waves of interest in which they provided data; the corresponding analysis weight would be the Wave 2 or Wave 3 single-wave weight (on records pertaining to Waves 1 and 2 or to Waves 1 and 3, respectively), and the Wave 3 all-waves weight (on records pertaining to Waves 2 and 3). By using these stacked data, weights, and corresponding replicate weights in model fitting and inference, the design effects are correctly incorporated in the analysis. However, the use of different weights within a given analysis does introduce subtle complexities in the interpretation of the population of inference. The exact format of the dataset and method of incorporating the replicate weights or variance structure may depend on the software used to perform the analysis.

With the introduction of the replenishment sample at Wave 4, there are further possibilities for analysis and multiple weights available to support them. Many of the considerations outlined in this paper, such as understanding the importance of weights and target populations for longitudinal analyses, continue to apply.

IMPLICATIONS FOR TOBACCO REGULATION

The FDA Center for Tobacco Product's goal is to "reduce the harm from all regulated tobacco products across the entire population" (https://www.fda.gov/tobacco-products/about-center-tobacco-products-ctp/center-tobacco-products-overview). Tobacco regulatory science research provides a scientific foundation to support the FDA in its mission. The PATH Study is a rich and important data source that provides tobacco-use data on over 10 distinct tobacco products (including cigarettes, traditional and filtered cigars, cigarillos, electronic nicotine delivery systems, hookahs, pipes, dissolvable tobacco, as well as bidis and kreteks for youth), allowing researchers to track tobacco product initiation and cessation at a granular level. The study also measures numerous downstream health outcomes related to tobacco use, including different types of cancer, respiratory disease, cardiac disease, oral health, and pregnancy outcomes. This paper, by explaining the study design and weighting considerations for longitudinal data analysis, supports researchers in making nationally representative inferences and providing key scientific findings that contribute to tobacco regulatory activities.

Human Subjects Approval Statement

The PATH Study has been conducted by Westat and the study procedures and materials were approved by the Westat Institutional Review Board. Westat is an employee-owned research company that conducts tobacco, health, and social science research studies for a variety of clients, including federal and state governments. Westat has never done any work with, for, or funded by the tobacco industry. All PATH Study respondents ages 18 and older provided informed consent, with youth respondents ages 12 to 17 providing assent while each one's parent/legal guardian provided consent.

Acknowledgements

This manuscript was supported with federal funds from the National Institute on Drug Abuse (NIDA), National Institutes of Health (NIH), and the Center for Tobacco Products, Food and Drug Administration, Department of Health and Human Services, under contract to Westat (Contract Nos. HHSN271201100027C and HHSN271201600001C) and through an interagency agreement between the FDA Center for Tobacco Products and the US Centers for Disease Control and Prevention. We acknowledge our colleagues at the FDA and NIDA who facilitated the funding, clearance, and execution of this research as well as our colleagues at Westat who contributed to the analyses. The findings and conclusions in this paper are those of the authors and do not necessarily represent the official position of the US Department of Health and Human Services or any of its affiliated institutions or agencies.

References

1. Hyland A, Ambrose BK, Conway KP, et al. Design and methods of the Population Assessment of Tobacco and Health (PATH) Study. Tob Control. 2017;26(4):371–378. doi:10.1136/tobaccocontrol-2016-052934 [PubMed: 27507901]
2. Ho PM, Peterson PN, Masoudi FA. Evaluating the evidence: is there a rigid hierarchy? Circulation. 2008;118:1675–1684. doi:10.1161/circulationaha.107.721357 [PubMed: 18852378]
3. Euser AM, Zoccali C, Jager KJ, Dekker FW. Cohort studies: prospective versus retrospective. Nephron Clin Pract. 2009;113:c214–c217. doi:10.1159/000235241 [PubMed: 19690438]
4. US Department of Health and Human Services, National Institutes of Health, National Institute on Drug Abuse, Food and Drug Administration, Center for Tobacco Products. Population Assessment of Tobacco and Health (PATH) Study Restricted Use Files User Guide. 10.3886/ICPSR36231.userguide
5. Tourangeau R, Smith TW. Asking sensitive questions: the impact of data collection mode, question format, and question context. Public Opin Q. 1996;60:275–304. 10.1086/297751
6. Wolter KM. Introduction to Variance Estimation. 2nd ed. New York, NY: Springer; 2007.
7. Judkins DR. Fay's method for variance estimation. J Off Stat. 1990;6:223–239. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/fay39s-method-for-variance-estimation.pdf. Published 1990. Accessed December 17, 2020.
8. American Association of Public Opinion Research (AAPOR). Standard Definitions: Final Dispositions of Case Codes and Outcome Rates for Surveys. 8th ed. Oakbridge Terrace, IL: AAPOR; 2015.
9. Groves RM. Nonresponse rates and nonresponse bias in household surveys. Public Opin Q. 2006;70:646–675. 10.1093/poq/nfl033
10. Groves RM, Couper MP. Nonresponse in Household Interview Surveys. New York, NY: Wiley; 1998.
11. Stoop IA. The Hunt for the Last Respondent: Nonresponse in Sample Surveys. The Hague, Netherlands: Social and Cultural Planning Office of the Netherlands; 2005.
12. Cunradi CB, Moore R, Killoran M, Ames G. Survey nonresponse bias among young adults: the role of alcohol, tobacco, and drugs. Subst Use Misuse. 2005;40:171–185. doi:10.1081/ja-200048447 [PubMed: 15770883]

13. Young AF, Powers JR, Bell SL. Attrition in longitudinal studies: who do you lose? Aust N Z J Public Health. 2006;30:353–361. doi:10.1111/j.1467-842x.2006.tb00849.x [PubMed: 16956166]
14. Song Y Rotation group bias in smoking prevalence estimates using TUS-CPS. Paper presented at the Federal Committee on Statistical Methodology Research Conference. 2013 Washington DC. https://nces.ed.gov/FCSM/pdf/I3_Song_2013FCSM.pdf Accessed September 9, 2020.
15. Fowler FJ, Stringfellow VL. Learning from experience: estimating teen use of alcohol, cigarettes, and marijuana from three survey protocols. J Drug Issues. 2001;31:643–664. 10.1177/002204260103100304
16. Pfeffermann D, Sverchkov M. Parametric and semi-parametric estimation of regression models fitted to survey data. Sankhya. 1999;61:166–186.
17. Korn EL, Graubard BI. Analysis of Health Surveys. New York, NY: John Wiley & Sons; 1999.

Table 1

PATH Study Data Collection Summary, by Wave

	Data collection		Interviews conducted		
Wave	**Start date**	**End date**	**Adults**	**Youth**	**Parents of youth**
1	September 12, 2013	December 14, 2014	32,320	13,651	13,588
2	October 23, 2014	October 30, 2015	28,362	12,172	12,129
3	October 19, 2015	October 23, 2016	28,148	11,814	11,807

Table 2

Number of Interview Respondents at Each Wave and Combination of Waves

Completed an interview in …	Number of interview respondents
Wave 1	45,971
Waves 1 and 2	38,443
Waves 1, 2, and 3	34,716
Waves 1 and 3	35,969
Wave 2 (including Wave 1 shadow youth)	40,534
Wave 3 (including Wave 1 shadow youth)	39,962
Waves 2 and 3 (including Wave 1 shadow youth)	36,663

Table 3

PATH Study Response Rates across Waves

Wave	Data collection[ab]	Weighted response rate (%)
1	Household screener	54.0
	Adult interview	74.0
	Youth interview	78.4
	Shadow youth (parental consent)	80.2
2	Adult interview	83.2
	Youth interview	87.3
3	Adult interview	78.4
	Youth interview	83.3

Note.

[a]The Wave 1 interview and parental consent response rates condition on household screener response; the Wave 2 and Wave 3 interview response rates condition on Wave 1 interview response or parental consent for shadow youth participation.

[b]Study participants were counted in the adult or youth follow-up wave response rates based on their participant type (adult or youth) at the time of the respective wave. Age information from previous waves was used to determine the expected participant type at the follow-up wave for nonrespondents.

Table 4

Questions and Responses Used to Construct Estimates of Youth Cigarette Smoking for the PATH Study, NHANES, NSDUH, and NYTS

Study	Ever tried cigarette smoking	Smoked cigarettes in past 30 days
PATH Study	YES: "Have you ever tried cigarette smoking, even one or two puffs?" (first interview, yes) or for any follow-up interview "In the past 12 months, have you smoked a cigarette, even one or two puffs?" (yes) NO: "Have you ever tried cigarette smoking, even one or two puffs?" (first interview, no) and for every follow-up interview "In the past 12 months, have you smoked a cigarette, even one or two puffs?" (no)	YES: "When was the last time you smoked a cigarette, even one or two puffs?" (earlier today, not today but sometime in the past 7 days, not in the past 7 days but sometime in the past 30 days) NO: *{If first interview* "Have you ever tried cigarette smoking, even one or two puffs?" (no) *or if follow-up interview* ["In the past 12 months, have you smoked a cigarette, even one or two puffs?" (no) and *criterion for YES not met*[a]]} or "When was the last time you smoked a cigarette, even one or two puffs?" (not in the past 30 days but sometime in the past 6 months, not in the past 6 months but sometime in the past year, 1 to 4 years ago, 5 or more years ago)
NHANES	YES: "About how many cigarettes have you smoked in your entire life?" (1 or more puffs but never a whole cigarette, 1 cigarette, 2 to 5 cigarettes, 6 to 15 cigarettes, 16 to 25 cigarettes, 26 to 99 cigarettes, 100 or more cigarettes) NO: "About how many cigarettes have you smoked in your entire life?" (I have never smoked – not even a puff)	YES: "During the past 30 days, on how many of the past 30 days did you smoke cigarettes?" (1-30) NO: "During the past 30 days, on how many of the past 30 days did you smoke cigarettes?" (0) or "About how many cigarettes have you smoked in your entire life?" (I have never smoked – not even a puff, 1 or more puffs but never a whole cigarette[b])
NSDUH	YES: "Have you ever smoked part or all of a cigarette?" (yes) NO: "Have you ever smoked part or all of a cigarette?" (no)	YES: "Have you ever smoked part or all of a cigarette?" (yes) and ['Time since last smoked cigarettes' (within the past 30 days) or "During the past 30 days, that is since *DATEFILL*, on how many days did you smoke part or all of a cigarette?" (1-30) or "What is your best estimate of the number of days you smoked part or all of a cigarette during the past 30 days?" (1 or 2 days, 3 to 5 days, 6 to 9 days, 10 to 19 days, 20 to 29 days, all 30 days)] NO: "Have you ever smoked part or all of a cigarette?" (no) or 'Time since last smoked cigarettes' (more than 30 days ago but within the past 12 months, more than 12 months ago but within the past 3 years, more than 3 years ago, used more than 12 months ago, used more than 30 days ago, used more than 30 days ago but in past 3 years) or ["During the past 30 days, that is since *DATEFILL*, on how many days did you smoke part or all of a cigarette?" (never used cigarettes, did not use cigarettes in the past 30 days) and "What is your best estimate of the number of days you smoked part or all of a cigarette during the past 30 days?" (never used cigarettes, did not use cigarettes in the past 30 days)]
NYTS	YES: "Have you ever tried cigarette smoking, even one or two puffs?" (yes) NO: "Have you ever tried cigarette smoking, even one or two puffs?" (no)	YES: "During the past 30 days, on how many days did you smoke cigarettes?" (1 or 2 days, 3 to 5 days, 6 to 9 days, 10 to 19 days, 20 to 29 days, all 30 days) NO: "During the past 30 days, on how many days did you smoke cigarettes?" (0 days)

Note.

[a]In Waves 2 and 3, youth completing a follow-up interview who reported not smoking a cigarette in the past 12 months, but who had previously reported ever trying a cigarette, were asked when the last time was that they had smoked a cigarette. Some of these youth reported smoking in the past 30 days, and so were included in the "yes" category.

[b]For NHANES, youth who had taken one or more puffs but never smoked a whole cigarette were not asked on how many of the past 30 days they had smoked a cigarette, and so were included in the "no" category.

Table 5

Questions and Responses Used to Construct Estimates of Adult Current Cigarette Smoking for the PATH Study, TUS-CPS, NHIS, NHANES, and NSDUH

Study	Current cigarette smoking
PATH Study	YES: "Do you now smoke cigarettes every day, some days, or not at all?" (every day, some days) and for any completed youth or adult interview "How many cigarettes have you smoked in your entire life? A pack usually has 20 cigarettes in it" (100 or more cigarettes) NO: [*If first interview* "Have you ever smoked a cigarette, even one or two puffs?" (no) *or if follow-up interview* "In the past 12 months, have you smoked a cigarette, even one or two puffs" (no)] or "Do you now smoke cigarettes every day, some days, or not at all?" (not at all) or "How many cigarettes have you smoked in your entire life? A pack usually has 20 cigarettes in it" (1 or more puffs but never a whole cigarette, 1 to 10 cigarettes, 11 to 20 cigarettes, 21 to 50 cigarettes, 51 to 99 cigarettes)
TUS-CPS	YES: "Have you smoked at least 100 cigarettes in your entire life?" (yes) and "Do you now smoke cigarettes every day, some days, or not at all?" (every day, some days) NO: "Have you smoked at least 100 cigarettes in your entire life?" (no) or "Do you now smoke cigarettes every day, some days, or not at all?" (not at all)
NHIS	YES: "Have you smoked at least 100 cigarettes in your ENTIRE LIFE?" (yes) and "Do you NOW smoke cigarettes every day, some days or not at all?" (every day, some days) NO: "Have you smoked at least 100 cigarettes in your ENTIRE LIFE?" (no) or "Do you NOW smoke cigarettes every day, some days or not at all?" (not at all)
NHANES	YES: "{Have you/Has SP} smoked at least 100 cigarettes in {your/his/her} entire life?" (yes) and "{Do you/Does SP} now smoke cigarettes every day, some days, or not at all?" (every day, some days) NO: "{Have you/Has SP} smoked at least 100 cigarettes in {your/his/her} entire life?" (no) or "{Do you/Does SP} now smoke cigarettes every day, some days, or not at all?" (not at all)
NSDUH	YES: "Have you ever smoked part or all of a cigarette?" (yes) and 'Time since last smoked cigarettes' (within the past 30 days) and "Have you smoked at least 100 cigarettes in your entire life?" (yes) NO: "Have you ever smoked part or all of a cigarette?" (no) or 'Time since last smoked cigarettes' (more than 30 days ago but within the past 12 months, more than 12 months ago but within the past 3 years, more than 3 years ago, more than 12 months ago, more than 30 days ago, more than 30 days ago but in the past 3 years) or "Have you smoked at least 100 cigarettes in your entire life?" (no)

Table 6

Past 30-day Cigarette Use of 12-to-17-year-olds at Wave 1 Compared to their Behavior at Wave 2 (Weighted Percentages and Standard Errors)[a]

Wave 1	Wave 2		
	Used cigarettes in past 30 days	**Did not use cigarettes in past 30 days**	**All youth**
Used cigarettes in past 30 days	3.18 (0.16)	1.40 (0.12)	4.58 (0.20)
Did not use cigarettes in past 30 days	4.11 (0.23)	91.31 (0.30)	95.42 (0.20)
All youth	7.29 (0.28)	92.71 (0.28)	

Note.

[a]Cases with a missing value for age or past 30-day cigarette use for either Wave 1 or Wave 2 were excluded from all estimates.

Table 7

Weighted Parameter Estimates and 95% Confidence Intervals for a Logistic Regression Model of Past 30-day Cigarette Use at Wave 2 among 12-to-17-year-olds at Wave 1[a]

Parameter	Estimate	Standard error	Value of t	Pr > |t|	95% Confidence limits	
Intercept	−1.02	0.07	−14.97	< .001	−1.15	−0.88
Wave 1: Cigarettes cause a lot of harm[b]	−0.12	0.06	−1.95	.054	−0.24	0.00
Wave 2: Cigarettes cause a lot of harm[c]	−0.50	0.06	−7.95	< .001	−0.62	−0.37
Race is White (alone)	0.24	0.05	4.69	< .001	0.14	0.35
Wave 2 household income above $50,000	−0.31	0.05	−6.14	< .001	−0.41	−0.21
Wave 1 past 30-day cigarette use[d]	1.83	0.06	30.09	< .001	1.71	1.95

Note.

[a]Cases with a missing value for any of the variables in the model were excluded.

[b]Wave 1 value recoded from R01_YC1125 (How much do you think people harm themselves when they smoke cigarettes? A lot of harm vs some, a little, or no harm).

[c]Wave 2 value recoded from R02_YC1125 for youth (as done for R01_YC1125) and R02_AC9050 for those 18 or older at Wave 2 (How harmful do you think cigarettes are to health? Extremely or very harmful vs Somewhat, slightly, or not at all harmful).

[d]Wave 1 value recoded from R01R_Y_CUR_CIGS (smoked a cigarette in past 30 days). Wave 2 value (dependent variable) recoded from R02R_Y_CUR_CIGS for youth and R02R_A_P30D_CIGS for those 18 or older at Wave 2 (smoked a cigarette in past 30 days).