

ONLINE DATA APPENDIX

A-1. Current Population Survey – Tobacco Use Supplements (TUS-CPS)

One of the most commonly used datasets in the study of tobacco regulation is the TUS-CPS.¹ The CPS is a nationally representative household-based survey of the non-institutionalized population that is run by the US Bureau of Labor Statistics. Households are surveyed every month, and in certain months supplemental questionnaires are given that focus on certain topics. The tobacco use supplements were conducted in 1992-1993, 1995-1996, 1998-1999, 2000, 2001-2002, 2003, 2006-2007, 2010-2011, and 2014-2015.² State identifiers are available in the data, and the sample sizes are large enough to generate representative estimates at the state level. The survey usually includes over 160,000 households per cycle, although the exact size varies across cycles.³

The main benefit of the TUS-CPS is that it can be linked to the general CPS survey. This linkage provides detailed information about respondent demographics, geographic location, and labor supply. Table A-1 shows the main questions used by researchers to determine smoking status and the number of cigarettes smoked by smokers. These questions are asked separately for every-day and some-day smokers (who also are asked how many times they smoked in the past 30 days), and there is a battery of questions for former smokers and about recent quit attempts. Beginning in 2003, smokers also were asked about whether they usually purchased cartons or packs, what the price is for the last carton/pack purchased, and in what state the last carton/pack was purchased. Respondents were asked as well about smoking menthol cigarettes beginning in 2003.

A-2. National Health Interview Survey (NHIS)

The National Health Interview Survey is a nationally representative dataset of the non-institutionalized population that has surveyed individuals since 1957. The survey is run by the Centers for Disease Control and Prevention (CDC). The goal of the survey is to monitor the health of the nation, and since 1965 the survey has collected information on adult smoking.⁴ The NHIS provides nationally representative patterns on adult smoking. State identifiers are available

¹ These data can be accessed through the NBER Data Repository: <https://www.nber.org/data/current-population-survey-data.html>.

² The May 2010 and the May 2011 Follow-Up form a special Longitudinal Cohort TUS-CPS.

³ A cycle of the TUS-CPS typically is comprised of 3 waves across two years.

⁴ These data are available at <https://www.cdc.gov/nchs/nhis/tobacco.htm>.

in the data, but the sample sizes are too small to generate representative estimates at the state level. The survey includes about 100,000 individuals per year, although the exact size varies each year.

The specific language of the smoking questions has varied over time, but the questions listed in Table A-1 are consistently asked. These questions also align closely with the questions in the TUS-CPS, making comparisons among these datasets straightforward. Many of the NHIS surveys also ask about prior smoking behavior, when respondents initiated smoking, and past quit attempts. In 2014, NHIS began asking about e-cigarettes as well. Limited demographic information is included in the dataset, such as age and race/ethnicity.

While the NHIS does not include the wealth of demographic, household composition, and labor market outcomes that are available in the CPS, it includes rich information about other health behaviors and outcomes. Some of the major topics include physical and mental health, chronic conditions (e.g., diabetes), use of healthcare, health insurance coverage, type of health insurance, health-related behaviors (e.g., smoking, drinking, and physical activity), and disability. These data thus are particularly useful in understanding how smoking relates to other health behaviors and outcomes.

A-3. National Health and Nutrition Examination Survey (NHANES)

The National Health and Nutrition Examination Survey (NHANES) is a dataset designed to provide information about the health and nutritional status of adults in the US. It is administered by the CDC, and it is distinguished from NHIS in that it combines survey responses and physical examinations. Because the physical examinations are intensive, the sample sizes are rather small. State identification information is available, but the sample sizes are sufficiently small that the estimates are unlikely to be representative at the state level. NHANES surveys have been conducted somewhat regularly: 1971-1974 (NHANES I), 1976-1980 (NHANES II), 1988-1994 (NHANES III), and continuously in two-year cycles since 1999.

As shown in Table A-1, the cigarette smoking questions are similar to those in CPS and NHIS but differ slightly in wording. The specific wording also has changed over time across NHANES cohorts. NHANES additionally contains questions about the timing of smoking initiation and past quit attempts. The main advantage of the NHANES data is information on cotinine levels from blood tests. Cotinine is a chemical left in the bloodstream from smoking, which provides a clinical measure of smoking intensity. Combined with traditional self-reported

questions on smoking, cotinine measures provide a check on survey-based measurement error. This measurement error can come from misreporting of smoking on the intensive and extensive margins. As argued by Adda and Cornaglia (2006), cotinine picks up intensity of how much each cigarette is smoked; they argue tax increases lead consumers to smoke more of each cigarette, which cannot be detected by sales or traditional consumption measures. The NHANES data include a rich set of clinical and survey information on health outcomes and behaviors, so like the NHIS it can be used to identify how smoking relates to other dimensions of health.

A-4. National Survey on Drug Use and Health (NSDUH)

The NSDUH is a nationally representative survey on 12+ year olds that is designed to measure use of alcohol, tobacco, and illicit drugs. The survey is run by the Substance Abuse and Mental Health Services Administration (SAMHSA) and currently is administered by RTI International. Data are available online starting in 1979. From 1979-2002 the survey was called the National Household Survey on Drug Abuse and was conducted irregularly until 1991, after which it became an annual survey. State, county, CBSA, and Census tract identifiers are available, but researchers must obtain a restricted-access license to use these geocodes. The smoking questions in NSDUH are very similar to those used in the other national surveys that ask about tobacco use, as shown in Table A-1. This dataset can be used to link tobacco use with other substance use in a large, nationally representative sample.

A-5. Behavioral Risk Factor Surveillance System (BRFSS)

The BRFSS dataset comes from a telephone-based survey conducted by the CDC that measures health outcomes and behaviors for a representative sample of US households each year.⁵ Data collection began in 15 states in 1984 and increased to cover all 50 states by 1993. Because it is telephone-based, it is a large survey that is designed to be representative at the state level. State identifiers are publicly available, as are county, metropolitan, and micropolitan statistical areas prior to 2013 when those areas are sufficiently large. In 2011 the survey started to include cell phones, and the proportion of those answering by cell phone has increased substantially over time.

The BRFSS questions on cigarette consumption match those in CPS and NHIS very closely (Table A-1). However, since 2001 the BRFSS no longer asks current smokers about how many cigarettes they smoke each day. The BRFSS includes questions related to recent quit

⁵ The BRFSS data are available here: <https://www.cdc.gov/brfss/index.html>.

attempts. The main advantage of the BRFSS data is the large sample size, while allows accurate estimates at the state and sub-state level. The data contain information on a wide range of health behaviors and outcomes as well, similar to NHANES and NHIS.

A-6. Youth Risk Behavior Surveillance System (YRBSS)

The YRBSS dataset is a national school-based survey conducted by the CDC designed to measure health-related behaviors among high school students. Surveys typically are conducted every two years in the spring semester. The national survey is representative of the public and private high school population in the US. The YRBSS also includes state, territorial, tribal government, and local school-based surveys of representative samples of high school students. However, the national surveys often differ from state and sub-state surveys. The survey began in 1991 and has surveyed student biannually since that time. State identifiers are available for the national survey and in the state surveys.

The YRBSS asks similar smoking questions to the other surveys discussed above, such as “Have you ever tried cigarette smoking, even one or two puffs?”, “During the past 30 days, on how many days did you smoke cigarettes?”, and “During the past 30 days, on the days you smoked, how many cigarettes did you smoke per day?” The dataset also includes a wide range of questions regarding alcohol and drug use, sexual behaviors, physical activity, and dietary behaviors. Like the BRFSS, a core strength of this dataset is the ability to link tobacco use with other health behaviors.

A-7. Monitoring the Future (MTF)

The Monitoring the Future dataset measures attitudes towards and use of tobacco, drugs, and alcohol among adolescents. The survey is funded by the National Institute on Drug Abuse and is administered by the University of Michigan Survey Research Center. Respondents report current, past-year, and lifetime usage of each substance, which allows researchers to measure time patterns of usage among adolescents. The survey has been run annually beginning in 1975. It is a school-based survey that includes nationally representative samples of 8th, 10th and 12th grade students. As shown in Table A-1, the cigarette smoking questions align closely with those in other national surveys. State, county, and zip code geocodes are available for researchers who obtain a restricted-access data license.

A-8. National Youth Tobacco Survey (NYTS)

The National Youth Tobacco Survey is run by the CDC and measures tobacco use and attitudes among school-age children. It is a school-based, nationally representative survey that was conducted in 1999, 2000, 2002, 2004, 2006, 2009, 2011 and annually thereafter. Middle school (grades 6-8) and high school (grades 9-12) students are the focus of the survey. The questions on smoking are similar to those employed in other surveys, most notably YRBSS. The NYTS is distinguished by focusing on tobacco use and asking more in-depth questions about multiple forms of tobacco use and students' attitudes towards tobacco use. State identifiers are not available.

A-9. National Longitudinal Surveys of Youth (NLSY)

The National Longitudinal Surveys of Youth (1979 and 1997) are ongoing longitudinal surveys conducted by the Bureau of Labor Statistics. The NLSY79 cohort is comprised of 12,686 individuals who were 14-22 in 1979, while the 1997 cohort is comprised of 9,000 respondents who were 12-16 years old in 1997. The children of the 1979 NLSY cohort (CNLSY) was started in 1986 and includes all children born to NLSY79 female respondents.

The NLSY79 surveys have asked questions on smoking initiation and participation off and on over time. The 1984 survey asks questions on age at first use, most recent use, and smoking in the past 30 days. In 1992, 1994, 1998 and 2008-2014, respondents were asked about the number of cigarettes smoked per day and the number of months/years since they had last smoked daily. Several survey waves also ask female respondents about smoking while pregnant. The NLSY97 cohort was asked in the initial interview about whether (and at what age) they had smoked a cigarette. In each subsequent round, respondents are asked about new smoking initiation as well as among smokers the number of days they smoked during the 30 days prior to the interview and the number of cigarettes they usually had on the days they smoked during those 30 days. A similar set of questions were asked of the CNLSY sample to ascertain age of smoking initiation, recent smoking prevalence, and the amount smoked among smokers.

The defining characteristic of the NLSY data relative to other datasets is its longitudinal nature. Because the NLSY surveys focus on initiation and ask whether individuals currently smoke in each follow-up, one can trace smoking behaviors across the lifecycle. Since most smoking initiation begins prior to the age of 20, and much of it begins prior to the age of 18, these data are useful in understanding how tobacco control policies affect initiation behavior as well as quit attempts and smoking intensity among smokers. The NLSY data also contain

detailed information on childhood background, schooling, labor market outcomes, family formation, and cognitive and non-cognitive measures. The ability to link such information to smoking behavior is unique to the NLSY surveys. The data include state identifiers as well, but they require access to the restricted-use version of the data.

A-10. Nielsen/IRI Homescan Data

Homescan data are provided by AC Nielsen or IRI and contain transaction-level purchases for a demographically representative set of households in the US. For each product purchased, the data include the Universal Product Code (UPC), the price paid, and the time and store of purchase. The data thus provide unique insight into the exact cigarette products being purchased and prices paid by consumers. Additionally, the data contain household demographics that include labor supply, income, race/ethnicity, and the makeup of the household (overall size and age composition).

Households are provided with a scanner and are asked to scan all items following a purchase. The scanner scans the barcode and records the UPC code, and then households record the amount purchased and the price paid. If a purchase is made at certain stores at which Nielsen or IRI monitors store-level prices, they will use the price based on store records for the given week.⁶ Households are paid for their participation by earning points for data uploads. The points can be used in exchange for products similar to the way credit card points are used for merchandise. One concern is that households only report data from their main shopping trip. Harding, Leibtag, and Lovenheim (2012) provide evidence that households report cigarette-only purchases as well.

The Homescan data are useful because they are arguably less subject to measurement error from recall or from the stigma of smoking. They also contain detailed purchase information about other goods (e.g., food) as well as price information for every cigarette pack/carton purchased and the location of purchase. This type of detailed price and product information linked to household characteristics is not available in other datasets. Another advantage is that the Homescan data are longitudinal. Although the panel covers a shorter time period than the NLSY, the Homescan data provided much higher-frequency observations (e.g. weekly

⁶ Einav, Leibtag, and Nevo (2010) assess measurement error in these data by matching Homescan data with data provided by a large retailer. They find evidence of measurement error, but these errors are unlikely to be related to changes in tobacco control policies.

purchases). The central drawback of the Homescan data is that they are household-based, so one cannot observe who smokes the purchased cigarettes in multi-person households. Unlike the other datasets discussed above, these data also must be purchased.⁷

A-11. State Cigarette Taxed Sales Data

Cigarette sales at the state-year level are available from the Orzechowski and Walker publication, *The Tax Burden on Tobacco*. All cigarettes that are legally sold in a state are reported to the state tax authorities in order to monitor excise tax compliance. These data are useful in providing administrative records on all cigarettes sold in a state. The main drawback is that one does not know who the consumers are who purchase the cigarettes. In particular, cross-state or Native American Reservation purchasing behavior leads state sales to be a problematic measure of cigarette consumption among state residents. See Section 5.1 for a more in-depth discussion of this issue.

⁷ The Nielsen Homescan data are available to academic researchers through the Kilts Center for Marketing at the University of Chicago Booth School of Business. <https://www.chicagobooth.edu/research/kilts/datasets/nielsen>.

Table A-1. Datasets Commonly Used to Study Cigarette Consumption

Dataset	Ages Covered	Years Covered	Main Cigarette Questions
Current Population Survey – Tobacco Use Supplement (TUS-CPS)	18+	1992-2017 (intermittent)	<ul style="list-style-type: none"> • Have you smoked at least 100 cigarettes in your entire life? • Current Cigarette Smoking Status • On average, about how many cigarettes do you smoke a day?
National Health Interview Survey (NHIS)	18+	1957-2017	<ul style="list-style-type: none"> • Have you smoked at least 100 cigarettes in your entire life? • Do you now smoke cigarettes every day, some days or not at all? • On the average, how many cigarettes do you now smoke a day?
NHANES	18+	1971-2018 (intermittent until 1999 and continuous thereafter)	<ul style="list-style-type: none"> • Ever tried cigarette smoking? (early surveys) • Smoked at least 100 cigarettes in life (later surveys) • # days smoked cigarettes in past month (30 days) • # cigarettes per day during past month • Cotinine levels from blood samples
National Survey on Drug Use and Health (NSDUH)	12+	1971-present (1979-present available online)	<ul style="list-style-type: none"> • Have you smoked at least 100 cigarettes in your life? • When was the most recent time you had a cigarette? • How many days smoked cigarettes in past 30 days? • Average number of cigarettes smoked per day on days smoked
Behavioral Risk Factor Surveillance System (BRFSS)	18+	1984-2018 (full state coverage from 1993); questions on cigarettes per day were discontinued in 2001.	<ul style="list-style-type: none"> • Have you smoked at least 100 cigarettes in your entire life? • Do you now smoke cigarettes every day, some days, or not at all? • On average, about how many cigarettes do you smoke a day? (not included 2001-2018)
Youth Risk Behavior Surveillance System (YRBSS)	High School Students	1991-2017 (biennial)	<ul style="list-style-type: none"> • Have you ever tried cigarette smoking, even one or two puffs? • During the past 30 days, on how many days did you smoke cigarettes? • During the past 30 days, on the days you smoked, how many cigarettes did you smoke per day?

Monitoring the Future (MTF)	8 th , 10 th , 12 th grade students	1975-present	<ul style="list-style-type: none"> • Have you ever smoked cigarettes? • How frequently have you smoked cigarettes during the past 30 days? • To be more precise, during the past 30 days about how many cigarettes have you smoked per day?
National Youth Tobacco Survey	6 th -12 th grade students	1999-present (intermittent)	<ul style="list-style-type: none"> • Have you ever tried cigarette smoking, even one or two puffs? • About how many cigarettes have you smoked in your entire life? • During the past 30 days, on how many days did you smoke cigarettes? • During the past 30 days, on the days you smoked, how many cigarettes did you smoke per day?
National Longitudinal Surveys of Youth (1979, 1997, and Children of 1979)	10+	1979-present (1979 cohort); 1997-present (1997 cohort); 1984-present (CNLSY sample)	<ul style="list-style-type: none"> • Questions vary across surveys but focus on age of initiation, current smoking status, recent smoking prevalence, and the number of cigarettes smoked per day on average.
Nielsen/IRI Homescan Data	All (Household Based)	Variable, depends on dataset	<ul style="list-style-type: none"> • Record of every cigarette purchased, if scanned, including UPC code & price. • Location of purchase • Day of purchase
State Taxed Cigarette Sales	N/A	1950-present	<ul style="list-style-type: none"> • Sales of all cigarettes legally sold in each state and year