Smoking and Mortality: New Evidence from a Long Panel*

Michael Darden[1], Donna B. Gilleskie[2], and Koleman Strumpf[3]

[1]*Department of Economics, Tulane University*
[2]*Department of Economics, University of North Carolina at Chapel Hill*
[3]*School of Business, University of Kansas*

March 2017

Abstract

Many public health policies continue to be rooted in the findings from medical and epidemiological studies that fail to account for behavioral influences. Using longitudinal data over nearly 50 years from men in the Framingham Heart Study, we provide estimates of the expected longevity consequences of different lifetime smoking patterns by jointly modeling individual smoking behavior and health outcomes and by allowing for correlated unobserved heterogeneity. Using simulations from our estimated empirical model, we compare the resulting mortality differences to the epidemiological literature that treats smoking behavior as random. The unconditional difference-in-means in age of death between lifelong smokers and nonsmokers is 9.3 years in our research sample, while simulations from our estimated dynamic model suggest the difference is only 4.3 years.

*The Framingham Heart Study (FHS) is conducted and supported by the NHLBI in collaboration with the FHS Study Investigators. This manuscript was prepared using a limited access dataset obtained from the NHLBI and does not necessarily reflect the opinions or views of the FHS or the NHLBI. We appreciate the comments of Robert Kaestner, Don Kenkel, Edward Norton, Tom Mroz, and seminar participants and discussants at Emory, Georgia State, Indiana, Lafayette, Lund, Notre Dame, SUNY-Stony Brook, Tulane, York, the Health Econometrics Symposium at Leeds University, the 5th (Spain) and 10th (Dublin) iHEA World Congress, the Econometric Society Summer Meetings, the Southeastern Health Economics Study Group, and the Triangle Health Economics Workshop. Financial support from the National Institutes for Health (grant #1RO1HD42256-01) is gratefully acknowledged. Correspondence: mdarden1@tulane.edu; donna_gilleskie@unc.edu; cigar@ku.edu.

1 Introduction

Many public health policies are rooted in findings from medical and epidemiological studies that often fail to account for health behaviors. We focus on smoking and mortality to demonstrate the importance of modeling behavioral contributions. Smoking is currently considered the leading preventable cause of death in the United States. According to the Centers for Disease Control, smoking causes 480,000 deaths each year and 8.6 million people have at least one serious illness due to smoking. Cigarette smoking is the primary causal factor in lung cancer and is a key risk factor in coronary heart disease.[1] In addition to the obvious negative health consequences of smoking, the medical and epidemiological literatures contend that quitting smoking has significant benefits. For example, ten years after quitting, an individual faces a cancer risk one-third to one-half as large as if he had continued smoking (Doll *et al.*, 2004). If a smoker quits smoking before age 40, this is associated with a 90 percent reduction in the excess mortality associated with smoking (Pirie *et al.*, 2013; Jha *et al.*, 2013). While we do not dispute that smoking causes significant excess morbidity and mortality, our research suggests that the accepted morbidity and mortality improvements accompanying smoking cessation may be overstated by as much as 50 percent.

There is ample biological evidence linking smoking to deleterious health outcomes. Yet, some puzzling aggregate trends demonstrate our concern with the literature's calculations of these impacts. Over the last twenty-five years adult smoking rates for both genders have fallen steadily to about half their initial levels, but the incidence of lung and bronchus cancer has doubled for women while declining for men. This variation could stem from heterogeneity in individual characteristics among the fifty million former smokers. While quitting smoking might suspend additional contributions to poor health, the precise nature of one's smoking history still predisposes that individual to cancers, heart disease, and other diseases.[2] Our research models smoking behavior and health outcomes over one's adult lifetime (ages 30 to 100) while controlling for the importance of variation in endogenous individual smoking and health histories.

Making accurate assessments of the longevity losses from cigarette smoking and the longevity gains from smoking cessation has proved difficult because smoking behavior is a choice. Ideally, the gain/loss predictions should be calculated from observed mortality differences following random assignment of lifetime smoking behavior. Because random

[1] For a good review of national trends in cigarette smoking and a summary of smoking-attributable diseases, see United States Surgeon General (2014).

[2] An additional explanation for the gender differences in lung cancer rates could be competing risks from other diseases (Honore and Lleras-Muney, 2006), yet mortality rates from cardiovascular diseases also fell more steeply for men during this period.

variation of this kind does not exist, researchers must rely on observational data to measure the effects of smoking on morbidity and mortality. When using non-experimental data, however, identification of the causal effect of smoking on mortality is difficult precisely because observed smoking behavior over one's lifetime is *not* random: individuals initiate smoking, may choose to quit smoking, and sometimes fail at quitting (i.e., relapse). These endogenous behaviors, which may occur at any age, produce very different lifetime smoking patterns. Therefore, the first contribution of this paper is determination of conditional impacts of the varying histories of smoking through joint estimation of smoking behaviors and health outcomes at frequent intervals over the life cycle (i.e., smoking histories are not exogenous).

An interrelated concern leading to difficulty in assessment of smoking's impact is that morbidity and mortality, while certainly not random, may be attributable to observed and unobserved non-smoking factors. It may be the case that failure to control for heterogeneity that explains correlation in smoking behavior and other behaviors that adversely influence health (e.g., excessive alcohol consumption, drug use, poor nutrition, etc.) leads to an overstatement of the health effects of smoking. Similarly, ignoring the factors that explain, for example, the inverse smoking/obesity correlation may understate the influence of smoking. While these correlations may be explained by observed individual variation, it is quite likely that unobserved characteristics such as risk-aversion, time preference or self-esteem and unobserved stress and health influence observed smoking and health patterns over the life cycle.[3] The second contribution of this paper is its generous inclusion of theoretically-justified controls for both observed (in our data) and unobserved (yet econometrically relevant) individual heterogeneity in order to uncover the conditional relationship between cigarette smoking and morbidity and mortality outcomes (i.e., confounding factors may influence health outcomes). Importantly, we allow the unobserved heterogeneity controls to be flexibly correlated across smoking and morbidity and mortality equations.

To achieve these empirical contributions we leverage a panel dataset of smoking behavior and health outcomes obtained at frequent intervals (via medical exams and survey questions) throughout much of the respondent's adult lifetime. Since the early 1950s, the Framingham Heart Study (FHS) has followed three generations of participants in order to identify contributors to heart disease.[4] Since 1948, most of the 5209

[3]There is evidence that differences in the health of certain regions of the brain influence the propensity to quit, and that these neural differences also influence other behaviors (Naqvi *et al.*, 2007).

[4]The original objective of the FHS, directed by National Heart Institute (now known as the National Heart, Lung, and Blood Institute or NHLBI), was to identify the common factors or characteristics that contribute to cardiovascular disease (CVD). Additional cohorts — offspring of the original cohort (1971), a more diverse sample (1994), and a third generation (2002) — have been recruited and are

subjects of the original cohort have returned to the study every two years (if alive) for a detailed medical history, physical examination, and laboratory tests. The non-death attrition rate of only three percent mitigates a typical source of selection bias.

This long-term, ongoing study consists of contemporaneous responses; it relies on recall of participants only to identify age of smoking initiation. We use 46 years of longitudinal observations (23 waves) on the male participants of the original cohort to construct detailed smoking histories (including duration, quits, and relapse). Among men, an important smoking transition is quitting (or attempts to quit), as most initiation occurs during adolescence. Often, depictions of individual smoking histories rely on less accurate, retrospective data gathered at a few disperse intervals making accurate identification of quits almost impossible. Our detailed, reliable histories and the modeling of behavior every two years allow us to simulate a range of quitting behaviors — quits at different ages, quits after different smoking durations, and quits with different cessation lengths — in order to evaluate the resulting impact on lifespan and cause of death. Additionally, much of the available health information in FHS is gathered during frequent (about every two years) detailed medical exams. Health events (e.g., diagnoses of heart disease, cancer, and diabetes and cardiovascular disease events including stroke) are dated and measures of risk factors (e.g., weight, blood pressure, cholesterol) are documented. With such detailed smoking and health data over time, we can model the dynamic effects of smoking histories on health as well as the dynamic effects of health histories on smoking behaviors. In fact, most studies rely on repeated cross-sections or panel data of only a few years in length, and so are unable to model dynamic behavior or to adequately include individual heterogeneity.

To estimate the marginal impact of different lifetime smoking patterns on health, we have four sources of identification. First, we collect new data on historical cigarette prices and advertising for over a century that we interact with age to get variation across individuals and time. We provide quasi-experimental evidence that these supply shifters are causally related to smoking levels. These theoretically-justified variables enter our behavioral equations (i.e., smoking) but are excluded from our health outcome equations (as in the simultaneous equation literature). Second, we use variation in the history of all exogenous explanatory variables captured by our dynamic equation specification (as in the dynamic panel data literature). Third, we include additional exogenous variables that explain the jointly-estimated initial condition equations (as in the literature that accounts for endogenous initial conditions). Finally, we leverage the functional forms of the non-linear estimators as well as covariance restrictions on the error structure across

being followed. (`www.framinghamheartstudy.org`).

equations and over time (as in the structural econometrics literature). We estimate the parameters of our multi-equation dynamic empirical model via full information maximum likelihood (FIML).

Figure 1 illustrates the advantages of our methods and these data. Figure 1a depicts the survival curves, by lifetime smoking pattern, generated using the observed age of death and smoking history of original cohort participants of the FHS. Figure 1b depicts survival rates calculated using data simulated from an estimated dynamic model of smoking behavior and health outcomes that includes the heterogeneity discussed above. While both figures indicate that smokers have, on average, higher mortality rates than non-smokers at all ages, the differences between the two groups are noticeably smaller when we account for non-random selection and confounding (Figure 1b) than when we simply examine the raw data (Figure 1a).[5] We depict similar comparisons in Figures 1c (observed data) and 1d (simulations from our estimated model) for smokers who quit by age 50 and never smokers. We discuss endogenous quits in more detail as the paper proceeds.

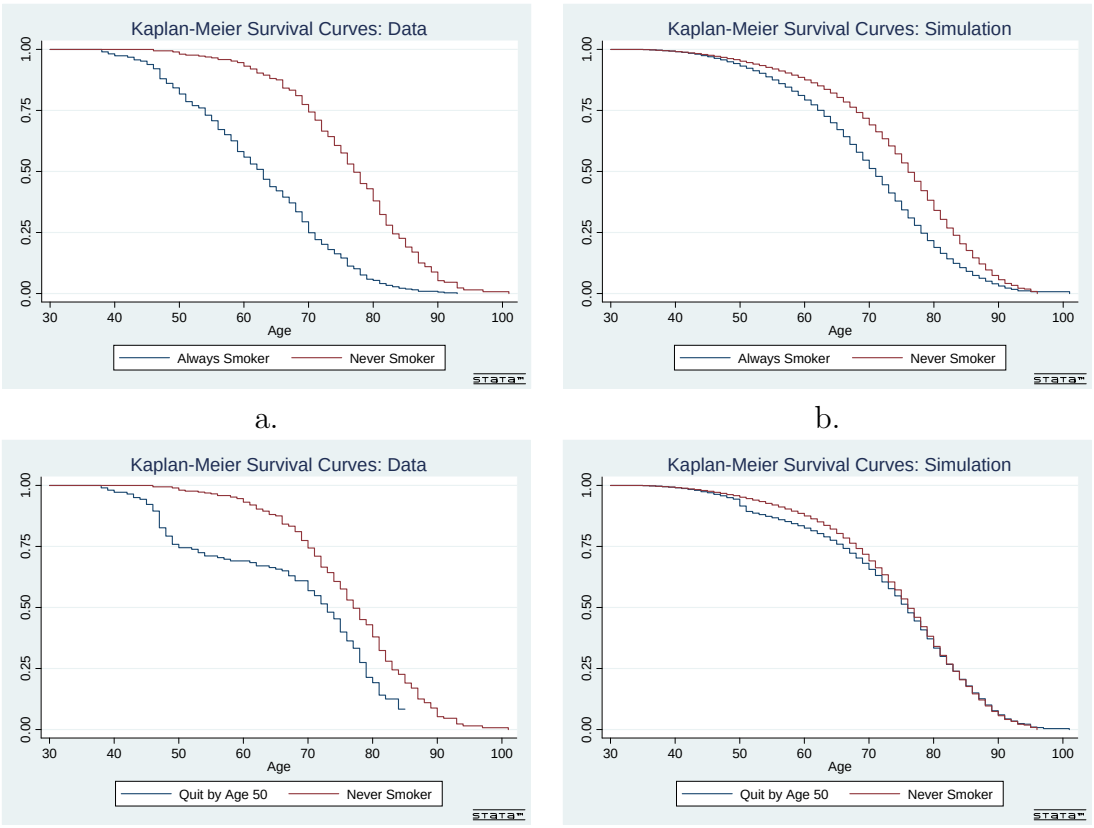
As discussed above, we contend that differences between Figures 1a and 1b are explained by the non-random nature of lifetime smoking patterns and by confounding factors in the production of health. Our results suggest that our dynamic specification has sizable impacts on the mortality effects of smoking and smoking cessation. As a basis for comparison, using the data on smoking and mortality alone, differences in the distribution of age of death by lifetime smoking behavior suggest that continuing to smoke reduces mean life expectancy by 9.3 years relative to those who never smoke. However, simulations from our estimated model show a reduction of just 4.3 years. We find a similar difference when we condition on cause of death. For example, for those who die of cardiovascular disease (CVD), the mean age of death is 9.3 years older for never smokers compared with continual smokers. Our simulations suggest that this difference is only 4.6 years. We also show that unobserved heterogeneity (UH) matters.[6] We find large differences in simulated smoking patterns when we condition on different unobserved "types". Overall, our results suggest that smoking does indeed reduce expected longevity; however, we find that failing to control for heterogeneity in smoking histories, health histories, and individual unobservables overstates the magnitude of these reductions.

Our main result – that the expected longevity effects of smoking are overstated

[5]The same comparisons among current and former smokers reveal qualitatively similar patterns.

[6]Throughout the paper we will distinguish between observed and unobserved individual heterogeneity, yet our emphasis is that "heterogeneity" matters. The variables that we observe are data source specific and, because no data are perfect, a researcher cannot possibly control for all relevant factors. For brevity, we abbreviate unobserved heterogeneity by UH henceforth.

Figure 1: Survival Curves by Lifetime Smoking Pattern



Notes: Figure 1a depicts the proportion of individuals in our estimation sample who remain alive at each age. Figure 1b depicts the survival rates of the same individuals as in Figure a. based on simulations from our preferred model when we impose the two lifetime smoking patterns. Smoker defines an individual who smoked from initiation until death; never smoker defines an individual who never smoked. Figures 1c and 1d present survival curves for those who quit by age 50 relative to never smokers from the data and from simulation, respectively.

by the medical literature – is strongly robust to a variety of modeling choices and robustness checks. For example, our basic result holds when we modify the morbidity outcomes being modeled, the number of points of support in the distribution of unobserved heterogeneity, and the modeling of the initial conditions. Furthermore, functional form alone is not driving our result: the effect of smoking on age of death in a simple regression drops dramatically after controlling for various time-invariant individual characteristics.

Our results echo conventional wisdom in some regards: twenty years of smoking experience has little impact on life expectancy if the individual quits by age 40. Additionally, we find that quits of five years followed by relapses have little benefit in terms

of life expectancy. The latter result suggests that cessation programs without follow-up support for former smokers will not be effective. Indeed, our results are an important contribution to public health discussions. For example, the Affordable Care Act allows exchange health insurance plans to charge a 50% surcharge for smokers; if the surcharge is calculated based on an overstated relationship between smoking and health, then the surcharge may prevent some smokers from purchasing health insurance. As another example, due to the associated health benefits from quitting, CBO (2012) projects that a $0.50 increase in the federal excise tax on cigarettes (fiscal year 2013) would reduce the federal budget deficit by $3 billion over the following decade.

The next section discusses the current state of knowledge on the health consequences of smoking, and shows how the approach and rich data that we use advances the literature. Section 3 describes how we constructed our research sample and details the structure of these data. We use Section 4 to introduce notation, to explain the empirical model, and to summarize the variables used in estimation. Section 5 provides results: parameter estimates, model fit, and simulations. We end with a discussion of the policy relevance of our findings.

2 Background

2.1 Medical Literature

Research from the epidemiological literature that seeks to understand the impact of smoking/quitting on mortality uses very limited, if any, empirical strategies to address bias associated with non-random selection or confounding behaviors. The heavily cited work of Doll *et al.* (1994) and Doll *et al.* (2004) employs panel data on British doctors over a forty and fifty year period, respectively. The authors compare mortality rates of physicians within 30-year birth cohorts by smoking history. They find that physicians who are current smokers of cigarettes (and who habitually smoked cigarettes) died on average ten years earlier than those who never smoked. Furthermore, quitting smoking at ages 60, 50, 40, and 30, relative to continuing to smoke, implies an increase in longevity of 3, 6, 9, and 10 years, respectively. The researchers condition on birth cohort and age only. There is no attempt to account for endogenous selection into different lifetime patterns of smoking. While the authors acknowledge that their findings may reflect a positive correlation between smoking and alcohol consumption (and hence the smoking effect may be biased upward), they claim that confounding factors reflected in other causes of death "are unlikely to have influenced greatly the absolute difference between the overall mortality rates of cigarette smokers and lifelong non-smokers" (Doll

et al., 2004).

Many of the United States Surgeon General's conclusions about the impact of smoking on health (United States Surgeon General, 2004) are based on the Peto *et al.* (2000) study that matches individuals diagnosed with lung cancer (cases) with observationally-similar individuals who did not have cancer (controls). The relative risk of cancer from smoking, compared to not smoking, is based on the ratio of smokers among the cases and controls. Their case-control approach only accounts for a small set of observed individual-level characteristics, such as gender and age. The limited observed variation and absent unobserved variation prevents a full understanding of the smoking-morbidity-mortality relationship. Furthermore, the findings rely on information collected at a single point in time (which means smoking histories are reported retrospectively and potentially with recall error) and consider only single outcomes (such as a disease or the mortality rate) independently.

2.2 Importance of Observed Individual Heterogeneity

Another approach to minimize bias has been a more thorough inclusion of observed individual-level heterogeneity. For example, some analyses have included objective measures of health (such as BMI or cholesterol) or behaviors correlated with health (such as exercise or drinking). Others have excluded individuals with pre-existing health conditions (National Cancer Institute, 1997) in order to avoid attributing other-cause deaths to smoking. Treating this additional heterogeneity as uncorrelated with the unexplained health outcome error introduces an endogeneity bias if there are common unobservables that influence both morbidity/mortality and the included variables. Also, it is likely that these additional variables are correlated with smoking behavior or, more importantly, that individual-level UH is correlated across the lifetime smoking categories, the included observed heterogeneity (if not exogenous), and morbidity/mortality.

More recently, economists have begun exploring the health consequences of smoking. One of the early economic analyses used smoking status of individuals (described by non-smoker, current smoker, or former smoker and years since quitting) to quantify the mortality benefit of quitting relative to continuing smoking over the subsequent 14-year period (Taylor *et al.*, 2002). However, smoking behavior was treated as exogenous and was ascertained at enrollment only (so quits are based on retrospective and potentially noisy responses and continuing behavior is assumed of current smokers). Additionally, the empirical specification does not account for differences in smoking duration among smokers. The primary control for heterogeneity was exclusion of individuals who were sick in the initial period. While this sample selection mitigates

bias due to poor-health-induced quits, it likely overstates the benefit of cessation since continuing smokers typically engage in less healthy behaviors than do former smokers (United States Surgeon General, 1990).

2.3 Importance of Individual Unobserved Heterogeneity

In order to address the important role of individual UH that influences both smoking behaviors and health outcomes, economists have used different econometric techniques to identify a causal impact. Evans and Ringel (1999) consider the impact of cigarette taxes on the smoking behavior of pregnant women, and the subsequent effect of smoking (and quitting smoking) on birth outcomes. Although the authors are not modeling own health outcomes, their work stresses the importance of accounting for smoking endogeneity as quitting is likely impacted by observables such as taxes as well as unobservables that may be correlated with infant outcomes. Clark and Etil (2002) use data from the first seven waves of the British Household Panel Survey (BHPS) to determine how responsive adult smoking behavior is to health changes. Because they have multiple observations on the same individuals they estimate cigarette demand in first differences using GMM estimation and in levels using twice (or more) lagged variables to instrument lagged demand. Both approaches address endogeneity bias that results from permanent individual unobservables (in the former case) and unobservables that are not autocorrelated of order greater than one (in the latter case). Their work contributes to the evidence of a dynamic relationship between smoking and health: namely that health changes precede smoking reductions. They suggest that the sample of remaining smokers is, therefore, not random.

Bedard and Deschenes (2006) identify the effect of smoking on premature mortality by studying mortality rate differences between cohorts who were more likely to have served in the United States military and, thus, to have received subsidized cigarettes. They find that cohorts with higher veteran rates experienced excess premature mortality, and a large proportion of the excess mortality was due to smoking-attributable diseases such as lung cancer.[7] Their analysis, however, does not include individual-level heterogeneity, does not include smoking history such as duration or quits, and relies on linear cohort controls that make it difficult to separately identify cohort, age, and year effects. Balia and Jones (2011) emphasize the important role of UH in reducing bias in the measured effect of smoking on mortality. Their latent factor model uses smoking

[7]More recently, Carter *et al.* (2015) suggest that much of the excess mortality among current smokers may be due to associations with diseases that have not been formally established as caused by cigarette smoking.

behavior of parents and other household members to exogenously shift an individual's smoking initiation and cessation while having no direct impact on his own mortality.

2.4 Importance of Smoking and Health Dynamics

By solving and estimating a dynamic model of cigarette consumption and mortality where forward-looking individuals explicitly take into account the health consequences of their smoking behaviors, Adda and Lechene (2001) find evidence that individuals with (observed and unobserved) characteristics that result in a higher risk of mortality (net of tobacco-related mortality) are less likely to quit or reduce smoking. The importance of modeling smoking and quitting decisions over the life course, especially when interested in the impact of smoking on morbidity and mortality, is further emphasized by the work of Khwaja (2010) who estimates a life-cycle model of endogenous health input decisions (including health insurance, medical care utilization, alcohol consumption, exercise, and smoking) using panel data on near elderly individuals as a way to correct for dynamic selection associated with survivorship. In addition to the evidence that smoking behaviors are impacted by health, which necessitates the modeling of both observed and unobserved individual characteristics that may be correlated with quitting and continuation, Adda and Lechene (2013) demonstrate that individuals in poorer non-smoking-related health are more likely to initiate smoking.[8]

We build on this literature, which emphasizes the importance of smoking endogeneity. Our first addition stems from our data source. The dataset consists of a very long panel that allows us to accurately observe each individual's smoking dynamics, including quits and relapses. Because the time between observations is short there is little issue of recall error. Also, a common medical examination given at each observation provides uniformity in the health variables. The second addition is in our modeling of smoking behavior. We use explanatory factors that vary over time, cohort, and/or individuals to model the smoking choices in each period. The modeling of the observed choices each period allows us to estimate unbiased marginal effects of very detailed smoking histories. The estimated dynamic model, in turn, allows us to simulate the morbidity and mortality outcomes resulting from a variety of policy- and behavior-relevant smoking decisions. For example, we quantify how the life expectancy benefit of quitting smoking varies by smoking history (i.e., the return for a long time smoker vs. a smoker with a history of relapse).

[8]Their findings, both on the relationship between health and smoking initiation and between smoking and mortality, are based on cross-sectional and some very short panel data, an attempt to capture UH with a time-invariant index of health conditions that are not directly caused by smoking according to the epidemiology literature, and no information on quitting or relapse.

2.5 Theoretical Considerations of Smoking and Health

The theoretical approach to smoking framed by economists is one that emphasizes several important aspects of the smoking/health relationship. Models of the demand for health, initially proposed by Grossman (1972) and extended by several others including most recently Kohn (2008) and Galama (2011), emphasize the role of individual health production and the resulting demand for health inputs. Health is valuable because it determines available time for activities that provide either monetary compensation and/or non-pecuniary reward. The rational addiction model and its variants (Becker and Murphy, 1988; Gruber and Koszegi, 2001; Bernheim and Rangel, 2004) suggest reasons why forward-looking individuals continue to smoke despite its impact on health. Namely, one's history of smoking behavior affects the enjoyment one receives from smoking today as well as withdrawal costs. Thus, individual characteristics such as time preference, risk aversion, and health expectations are important for explaining observed behaviors. Economists also explore the role of information in helping individuals form expectations of uncertainties such as future health and the impact of smoking on own health. While there is evidence that expectations about future longevity are relatively accurate (Viscusi, 1990; Smith *et al.*, 2001; Viscusi and Hakes, 2008), economists have also discovered that individuals respond, with changes in their own smoking behavior, to information that is more personal (e.g., own health decline and shocks, own health markers, and parental health shocks) than general (e.g., Surgeon General's reports, spousal health shocks).[9]

2.6 Overview of Our Approach

In light of these theoretical contributions and empirical explorations, our approach is to estimate, using a panel of individuals followed for much of their adult lives, a dynamic model that approximates the forward-looking decisionmaking that results in observed life-cycle smoking behaviors and health outcomes. Importantly, our empirical model explains smoking behaviors over time as a function of one's health and smoking histories. In turn, that history of smoking behavior impacts morbidity and mortality. In addition to rich observed variation in individual health characteristics, we model both permanent and time-varying UH using a non-linear latent factor approach, or correlated random effects in a FIML framework. Causal impacts of smoking history

[9]See Smith *et al.* (2001); Sloan *et al.* (2002); Darden (forthcoming); Darden and Gilleskie (2016). An important difference between our work, which focuses on accurately estimating the impact of smoking on mortality, and these recent papers is that 1) our data follow individuals frequently (every two years) throughout much of their adult lifetime (up to 46 years vs. 28 years (7 waves) or less) and 2) nearly all individuals are observed until death (88 percent vs. up to 30 percent).

on health outcomes are identified by time-varying exogenous covariates, theoretically-relevant exclusion restrictions, the dynamic relationships in the set of estimated equations, covariance restrictions on the correlated UH, and non-linear estimators. Using FHS data to construct our research sample, we follow 1,464 men for up to 46 years with censoring determined by either death or the end of the available sample period.[10] The data include a variety of health marker measures such as cholesterol and blood pressure, cardiovascular health measures, cancer diagnoses, and smoking information obtained at each health exam approximately two years apart. Eighty-eight percent of our sample die during the observed time frame. After showing that our preferred random effects specification fits the observed smoking and health data well, we simulate the model under different lifetime smoking scenarios to assess the effects of smoking, and smoking cessation, on morbidity and mortality outcomes.

3 Description of the Research Sample

3.1 The Framingham Heart Study

The original cohort of the FHS is well-suited for our analysis because it follows a group of men and women aged 30 to 62 in 1948 who receive a medical examination every two years (on average) to the present. Longitudinal datasets of this length are rare in health and economics. The U.S. Public Health Service began collecting this information in an effort to identify biological and environmental factors contributing to the rapidly rising rates of cardiovascular death and disability.

The FHS original cohort sample consists of two-thirds of the adult population of Framingham, Massachusetts in 1948. The main drawback of these data, for our purposes, is that the sample is drawn from a single, small town with very little racial and ethnic variation. As a result, geographic and demographic characteristics are limited. Additionally, because the focus of data collection is on health, there is no information on socioeconomic characteristics such as income and only limited information on employment or occupation. Finally, it is challenging to measure the sensitivity of smoking to prices, taxes, or regulations, which in our sample vary over time but not across individuals within a year.[11]

[10] We have permission to use the requested set of variables through 1996 only. We were able to obtain death dates of sample participants through 2009.

[11] As we explain later, we gathered detailed data on cigarette prices and advertising expenditures from the late 19th century throughout the 20th century in order to explain differences in contemporaneous smoking behaviors as well as initially-observed smoking histories. We use the variation in age each year to examine possible age-related responses to cigarette price variation over time. Additionally, we

The breadth of health data, however, is overwhelming. Theoretically, measures of height and weight, blood pressure, cholesterol, arthritis, CVD events, cancer, diabetes, death, and cause of death are available each medical exam (every two years). Realistically, values of these health variables are missing sporadically. We impute values based on previous knowledge or interpolation where it makes sense to do so. Other health behaviors are available at points in time, but may not be ascertained at each exam. We do not use these because we cannot adequately construct values over the life cycle.

Questions regarding smoking behaviors are often asked in great detail. For example, for particular exams individuals were asked to report whether they smoked or not, what types of products they smoked, the number of cigarettes smoked per day, whether they used filtered cigarettes or not, whether they inhaled or not, and whether they smoked all or part of the cigarette. However, the questions were not asked every exam and did not always offer the same alternative responses when repeated. For example, for 12 percent of the individuals whom we know to have ever smoked, we know nothing about their smoking intensity (i.e., the number of cigarettes smoked per day). For 42 percent of the person-year observations in which a person smoked, we do not observe intensity. These omissions make it impossible to construct measures of smoking intensity beyond the extensive margin. The sporadic intensity data that we do observe suggests that our sample contains relatively heavy smokers with about 85 percent of smokers reporting smoking a pack or more of cigarettes a day. Even if we made assumptions (from the available data) about the likely level of intensity of someone ever observed to smoke, the associated increases in the dimension of the smoking history vector (i.e., variables representing years of experience and tenure with light smoking and heavy smoking) greatly complicates estimation of the effects of this history on contemporaneous smoking behavior and morbidity and mortality outcomes.

3.2 Structure of Research Sample

To estimate our dynamic model of lifetime smoking behaviors and health outcomes we would like to have data at each exam (approximately every two years) for each participant. In cases where there are gaps in the smoking history, we impute observations when the gaps are minor and drop individuals from the sample when they are larger. (Appendix A contains full details of our procedure.) We begin with a sample of 1754 men with sufficient data on smoking behavior over the course of data collection. If an individual died before the second exam period or we were unable to construct the necessary health history information or initial smoking behavior, we can not include

use the price and expenditure variation at the different reported ages of smoking initiation.

them in our research sample. Our final research sample for this analysis consists of up to 22 biennial observations on 1464 (i.e., 1754 - 290) men beginning in 1952, providing 21,198 person-year observations.[12] Table 1 characterizes our research sample over time. Because individuals enter the study between the ages of 30 and 62, many already have a "positive" history of smoking participation and some have a health history that indicates onset of chronic conditions. We use responses from contemporaneous and retrospective questions administered at the first few exams to construct these initial condition variables. For this reason the initial conditions are represented in our research sample as exam 2 variables and we begin modeling dynamic smoking transitions and health outcomes beginning in exam 3 (around the year 1952).[13]

Because we have data on the same individuals through 1996 (when ages of those still alive range from 74 to 101), we are able to observe the age and cause of death for 87.8 percent of our sample of men (1285 out of 1464). Cause of death is categorized into cardiovascular related, cancer related, and other. Cardiovascular disease (CVD) includes myocardial infarction, angina pectoris, coronary heart disease, stroke, and heart failure. We know the type of cancer (within classes) that an individual acquired, but we aggregate all cancers in our analysis due to sample size limitations.[14] Note that CVD and cancer account for over two-thirds of all observed deaths in our sample, which is comparable to U.S. death rates for this age group during this time frame. Also, deaths related to CVD appear to occur at a higher rate at younger ages, perhaps linked to the national trend where death from cardiovascular disease has fallen substantially over the last 50 years.[15]

Table 1 confirms that the percentage of current cigarette smokers declines over the sample period.[16] This decline reflects quits as well as selective mortality, and we emphasize both in our empirical approach. It is important to measure well the different smoking histories of individuals that accumulate as they age. Rather than rely on retrospective data, the FHS allows us to observe smoking behavior at frequent intervals

[12]Because of likely differences by gender in determinants of smoking and impacts of smoking on health, we examine male behavior and outcomes in this study.

[13]Note that calendar years may overlap across exam numbers. For example, at exam 2, which was administered to a participant sometime between the years 1950 and 1955, ages of individuals range from 32 to 65.

[14]United States Surgeon General (2004) reports that smoking raises the risk of many types of cancer, not only lung cancer.

[15]This positive trend has been attributed to both reductions in smoking (and other risk factors, despite increases in obesity and diabetes) and advances in medical care (Prince *et al.*, 2014).

[16]In general, U.S. smoking rates among this cohort have declined over time from 55.8 and 31.8 percent, for men and women respectively, to 6.6 and 7.8 percent among those still alive 45 years later. As discussed above, there were regulatory changes in the cigarette market as well as dissemination of information about cigarettes during this period.

Table 1: Characterization of Research Sample over Time

FHS exam number	Calendar years	Empirical model period (t)	Sample size at t	Average age at t	Death* by end of t	Cause of death** if died in t		Smoke in t
						CVD	Cancer	
2	1950-55	1	1464	46.6	0.0	0.0	0.0	55.8
3	1952-56	2	1464	48.6	1.7	64.0	20.0	58.8
4	1954-58	3	1439	50.5	2.2	58.1	22.6	59.4
5	1956-60	4	1408	52.3	2.6	62.2	13.5	53.6
6	1958-63	5	1371	54.2	2.7	67.6	24.3	52.4
7	1960-64	6	1334	55.9	3.2	47.6	31.0	52.6
8	1962-66	7	1292	57.7	3.3	65.1	16.3	47.3
9	1964-68	8	1249	59.6	4.4	47.3	27.3	42.1
10	1966-70	9	1194	61.3	4.4	39.6	43.4	37.4
11	1968-71	10	1141	63.1	5.6	46.9	28.1	30.2
12	1971-74	11	1077	64.7	7.1	47.4	18.4	29.2
13	1972-76	12	1001	66.2	7.5	54.7	29.3	27.8
14	1975-78	13	926	67.8	7.1	37.9	34.9	24.4
15	1977-79	14	860	69.3	9.8	40.5	27.4	21.7
16	1979-82	15	776	70.8	10.3	38.6	33.8	19.3
17	1981-84	16	696	72.4	14.8	35.9	23.3	16.0
18	1983-85	17	593	73.7	13.2	32.1	21.8	15.0
19	1985-88	18	515	75.3	16.5	35.3	25.9	13.2
20	1986-90	19	430	76.6	13.0	37.5	33.9	10.5
21	1988-92	20	374	78.2	13.1	38.8	26.5	9.4
22	1990-94	21	325	79.7	17.2	30.4	21.4	7.1
23	1992-96	22	269	81.3	33.5	20.0	22.2	6.7
Total person-observations			21,198	60.6	6.5	42.1	26.3	38.6

Note: $*$ conditional on survival up to t (i.e., the death hazard).
$**$ omitted category is Other.

over much of adult life. Hence, the specific age one quits smoking, as well as ages of relapse, are observed. In our sample of men, 24 percent never smoke. Smoking initiation typically begins at young ages. In fact, only four percent of our sample had both never smoked before we initially observe them and smoked at some point in our data. (Among those with no observed smoking history in 1952 (28 percent of the research sample), only 13 percent initiate smoking before death or 1996, whichever comes first.) Twenty-seven percent of the sample smoke continuously (i.e., every period they are observed once they begin smoking). Among those men we observe ever smoking, 49 percent quit smoking at least once with a 10.1 percent person-period quit rate. Of the men that quit, 74 percent of them do not restart.[17] Among those who relapse within our period of observation, the mean time of smoking cessation between spells of smoking is 3.3 years. The mean age of relapse is 53.9 years. These figures emphasize the non-random patterns of smoking histories and the importance of a long panel consisting of frequent interviews and limited dependence on recall.[18]

Table 2 details the distribution of age of death and cause of death among those who die during our sample period. By exam 23, 87.8 percent of the individuals have died; the average age of those still alive by the end of this exam period is 81.3. The overall mean age of death (conditional on being observed to die) was 72.6 years of age.[19] We distinguish deaths by lifetime smoking pattern (i.e., never smoked, smoker in exam before death, and quit smoking before age 50) and by cause (i.e., cardiovascular disease, cancer, and other). Men who die of cardiovascular disease and cancer die, on average, at ages 70.8 and 71.9 respectively, while those who die of other causes live to age 75.5 on average (not shown in table). For men who never smoke, the average age of death is 75.6 years; and for men who report smoking immediately prior to death, the mean age of death is 66.2 years of age. The difference of 9.3 years is similar to the result

[17]This figure includes quits that may have occurred prior to the beginning of the study. The relapse rate is likely understated because the length of time between reported smoking measures is two years. Among all person-periods where an individual did not smoke last period but had ever smoked, the rate of relapse is 37.5 percent.

[18]Individuals in our sample range from ages 30 to 62 when we first observe them. In estimation we account for potential cohort differences by including an indicator of whether or not an individual is 50 years of age or older when first observed (31 percent of the sample). Those in the older cohort have survived to age 50 and may be healthier in unobserved ways. In fact, they die on average almost five years older than the younger cohort (age 76.4 vs. 71.6). Individuals from the observed older cohort are slightly more likely to smoke in any year after age 50 (57.5 vs. 54.5 percent). In addition to our inclusion of the cohort indicator, our joint estimation of initially-observed smoking and health histories that depend on UH addresses potential selection concerns.

[19]Life expectancy of men born in Massachusetts between 1885 and 1918 (i.e., birth years of men in the original FHS cohort) conditional on reaching age forty was 67 to 69 years (Bureau of the Census, 1949).

reported in Doll *et al.* (2004). Ever smoking is also associated with a higher proportion of cancer-related deaths. Interestingly, the raw data suggest that smoking cessation by age 50 results in a comparable lifespan to those who never smoked, yet they still experience cancer-related deaths in a higher proportion. The Surgeon General Reports use similar life year gains to advocate for smoking cessation programs. We demonstrate with our lifetime model of smoking and health that these figures are biased upward.

Table 2: Distribution of Age and Cause of Death by Lifetime Smoking Pattern

	Age of death distribution (percentile)						Cause of death		
Smoking pattern	Mean	10th	25th	50th	75th	90th	CVD	Cancer	Other
Unconditional on smoking	72.6	58	66	74	80	85	42.1	26.3	31.6
Never smoked	75.6	64	70	76	82	87	44.4	19.9	35.8
Smoked in exam prior to death	66.2	52	59	67	74	79	44.3	28.8	26.9
Quit smoking before age 50	74.1	63	70	75	80	84	39.4	29.2	31.4

Note: Statistics conditional on death by exam 23 (87.8 percent of sample).

4 Empirical Framework

Our goal in this section is to describe an empirically-implementable model that captures the dynamic considerations of forward-looking individuals making optimal smoking decisions in light of uncertain health evolution. The key features of the theory from which we derive our empirical model are: (i) individuals care about discounted lifetime utility; (ii) individuals derive utility from smoking (and other consumption), (iii) the marginal utility of smoking may depend on past levels of smoking, (iv) utility also depends on current health, and (iv) smoking histories, health histories, and current smoking behavior impact the distribution of future health outcomes.

4.1 Theoretical Foundation and Derived Empirical Model

To be more specific and to define notation, we write down the individual's optimization problem using a Bellman formulation.[20] From this problem we derive an estimable equation for smoking demand. The lifetime value of each smoking alternative depends on information known by the individual when entering each decision-making period. The information set, denoted Ω_t, includes the vector H_t^S representing an individual's history of smoking decisions up to period t; the vector H_t^D representing his history of diseases up to period t; the vector X_t of exogenous demographic variables in t; and the vector P_t of exogenous period t prices and supply-side characteristics related to the consumption/health input goods. The individual also has private information about his preferences for smoking and his expectations about disease and mortality transitions, denoted by the vector $u_t = [u_t^S, u_t^D, u_t^M]$. Conditional on being alive to make a smoking decision, the lifetime value of smoking alternative $s_t = s$ is

$$\begin{aligned} V_s(\Omega_t, u_t) \quad = \quad & \sum_{d=0}^{D} p(d_t = d | H_t^D, H_t^S, X_t)[U(c_t, s_t = s; H_t^S, d_t = d) + u_t^s \qquad (1) \\ & + \beta(1 - p(m_{t+1} = 1 | H_{t+1}^D, H_{t+1}^S, X_t) V(\Omega_{t+1})] \quad \forall t, s = 0, 1 \end{aligned}$$

where utility is constrained by the per-period budget, $c_t = y_t - P_t s_t$; the price of general consumption c_t is normalized to one; y_t measures income in period t; and P_t includes the price of cigarettes.[21] We capture uncertain health transitions by $p(d_t = d|\cdot)$ where d_t represents a vector of disease states taking on the value d. Current utility depends on the disease state. Contemporaneous utility also depends on one's smoking history, H_t^S, to capture tolerance, reinforcement, and withdrawal effects (i.e., addiction) that vary with an individual's past smoking behaviors. To characterize future utility (line 2 of equation 1), we define β as a measure of how forward-looking an individual may be (i.e., the discount factor); we allow for an absorbing mortality state stochastically with $p(m_{t+1} = 1|\cdot)$ where the value of death is normalized to zero; and we describe the maximal expected value of future utility (unconditional on the future smoking alternative) by $V(\Omega_{t+1}) = E_t[\max_s V_s(\Omega_{t+1}, u_{t+1}^s)]$. The dynamic optimization problem allows smoking and disease histories to impact expected current utility and allows those histories and current behavior and health to affect expectations about future utility.

Optimal smoking decisionmaking requires backward solution from a final period

[20] Individual subscripts n are dropped to simplify notation.

[21] For simplicity, we take income as given and do not model employment, marital, or savings decisions nor the effect of health on these behaviors. This decision is predicated by the fact that the FHS data do not contain this information.

characterized by certain death. Analytic solution also requires functional form assumptions for several components of the problem including the utility function, the disease production function, the mortality function, and the utility error term distribution. Theoretically, the optimization problem can be solved to obtain a decision rule for smoking of the form

$$p(s_t = s) = f(s_t^*) \text{ where } s_t^* = s(H_t^S, H_t^D, X_t, P_t, u_t^S),\ s = 0, 1 \quad \forall t. \tag{2}$$

Notice that the demand (for smoking) equation is a function of all information available to the individual at the beginning of the decision-making period. Specifically, the vector H_t^S (capturing smoking history up to period t) includes previous period smoking status, s_{t-1}; the length of smoking cessation up to t, C_t,; the length of smoking duration up to t, D_t; and the length of smoking experience up to t, E_t.[22] The smoking history is updated at the end of period t (i.e., H_{t+1}^S) to reflect smoking choices made at t. Importantly, the specification also includes exogenous supply-side characteristics of the cigarette market, P_t (e.g., prices, advertising), that vary over time.[23]

An individual faces uncertain health outcomes each period. We model the health production function as

$$p(d_t = d) = f(d_t^*) \text{ where } d^* = d(H_t^D, H_t^S, X_t, u_t^D),\ d = 0, \ldots, D \quad \forall t \tag{3}$$

where the "disease" variable may take on several discrete values. In practice, we estimate the number of cardiovascular disease (CVD) events in period t; the probability of cancer diagnosis (by the end of period t conditional on no cancer diagnosis up to t); the probability of diabetes diagnosis (conditional on no diabetes diagnosis up to t); and body mass in period t.[24] The vector H_t^D (capturing disease history up to period t) includes variables constructed from the health outcomes that are modeled (i.e., any CVD events entering period t, the number of CVD events entering period t, cancer diagnosis ever, diabetes diagnosis ever, and BMI in period $t-1$) as well as variables that

[22] We summarize the history of smoking behavior using (polynomials of) these four variables rather than including indicators of smoking behavior at each age, which would be computationally intractable. Are we missing important aspects of past behavior? We do not capture variation in smoking intensity, for the reasons described in section 3.2, and we remind the reader that most smokers during this time smoked a pack or more a day. We also do not explicitly account for the number of attempts to quit smoking, which could reveal information about an individual's unobserved type. However, the variables and specification we include is rich. Additional non-linearities did not alter the results.

[23] We return to discussion of these theoretically-important variables when we discuss initial conditions and identification below.

[24] Body mass is modeled as a continuous distribution of the body mass index (BMI), which is a normalized function of height and weight.

we treat as exogenous that also describe one's health (i.e., per-period systolic blood pressure, diastolic blood pressure, cholesterol levels, and an indicator of arthritis).[25]

Mortality, or the probability of death at the end of period t (i.e., an individual dies before making it to the next exam) is

$$p(m_{t+1} = m) = f(m^*_{t+1}) \text{ where } m^* = m(H^D_{t+1}, H^S_{t+1}, X_t, u^M_t),\ m = 0, 1 \quad \forall t. \qquad (4)$$

Because death is an absorbing state, decision rules and health production are conditioned on being alive in period t. Non-random mortality, therefore, creates important selection into the sample of (remaining) individuals whose characteristics explain the modeled smoking behaviors and health outcomes. Note that the probability of survival to the next period $(t+1)$ depends on the updated (i.e., accounting for current period observed outcomes) disease and smoking histories.

As should be evident, the period t demand for smoking is identified by variation in period t supply-side conditions (P_t) that, according to an economic theory of individual optimizing behavior, impact the smoking decision but do not independently impact health production or mortality conditional on period t smoking behavior. In other words, equations 2, 3, and 4 form a set of structural (demand and production) equations that can be empirically identified and estimated. These dynamic equations explain smoking choices, s_t, and health outcomes, d_t and m_{t+1}, from periods $t = 2$ to $t = 22$, where t denotes the two-year period between exams in our data.

4.2 Initial Conditions

Because we first observe individuals between the ages of 32 and 65 (in period $t = 2$), we must account for the endogeneity of initially-observed smoking history [denoted by E_2 (smoking experience entering $t = 2$ where $E_2 = 0$ implies never smoked and $E_2 > 0$ implies ever smoked), s_1 (smoking status in $t = 1$ conditional on ever smoking prior to $t = 2$), and D_2 (years of smoking duration entering $t = 2$ conditional on smoking in $t = 1$)] as well as disease history [denoted by CVD_2 (any CVD events entering $t = 2$) and BMI_1 (body mass index in $t = 1$)].[26] For simplicity we denote the three initial

[25] The medical literature tells us quite plainly that blood pressure and cholesterol are impacted by smoking. Arthritis, however, appears to have little association with smoking (to our knowledge), but does impact health transitions. For the sake of parsimony, we have chosen to present results from a model where these outcomes are not jointly modeled with the set of equations defined below. Conclusions about the effects of smoking on morbidity and mortality were not appreciably different using a larger model that treated these health variables as endogenous.

[26] We do not include equations for initially-observed cancer or diabetes because very few individuals in our sample enter the FHS with these diseases. We also exclude an equation for initially-observed years of smoking cessation conditional on having quit prior to the first health exam because the small

smoking variables entering period $t = 2$ by the vector I_2^S and the two initial disease variables entering period $t = 2$ by the vector I_2^D. We specify the initial conditions by the following non-dynamic equations[27]

$$\begin{aligned} I_2^S &\equiv [E_2, S_1, D_2] = s'(X_1, P_1, Z_1, u_1^{I^S}) \\ I_2^D &\equiv [\text{CVD}_2, \text{BMI}_1] = d'(X_1, P_1, Z_1, u_1^{I^D}) \end{aligned} \tag{5}$$

The initial condition equations are included in the set of jointly-estimated structural equations. Note that these initial condition equations are static and expressed in their reduced form; they do not contain any lagged endogenous variables. In addition to the exogenous cigarette market characteristics, we also include exogenous shifters, denoted Z_1, to aid in identification (discussed below).

4.3 Individual-level Unobserved Heterogeneity

Unobserved individual characteristics (i.e., latent heterogeneity) also impact smoking demand, morbidity, and mortality (represented by equations 2-5). It is important to model this correlated UH for several reasons. First, it is reasonable to believe that unobserved individual differences impact smoking behavior and health. In fact, the ability of observed variables to explain health outcomes is notoriously low. These differences include permanent unobserved characteristics such as health-related genetic endowments or cohort effects and non-health related personality or preference characteristics. They also include differences such as unobserved health events or stress events that vary over time. Second, the outcomes we model are functions of endogenous explanatory variables and, as such, the error term in the equation of interest is correlated with the explanatory variable, creating endogeneity bias in the estimated coefficients. Third, measurement error cannot be ruled out, so allowing for a source of this error reduces measurement error bias in marginal effects of interest. Accounting for these unobserved differences is necessary for obtaining unbiased causal impacts of the variables of interest.

To model these potential sources of correlated UH, the composite error term in each equation j, u_t^j, is decomposed into three parts: a permanent individual heterogeneity component (μ), a time-varying, serially-uncorrelated individual heterogeneity component (ν_t), and an idiosyncratic component (ϵ_t). More specifically, $u_t^j = \mu^j + \nu_t^j + \epsilon_t^j$. The latent heterogeneity captured by μ allows for correlation across smoking behavior and

number of quitters does not provide enough variation for estimation.

[27]The prime superscript on the functions is meant to distinguish them from the previously defined dynamic smoking and disease functions.

health outcomes within a period and across time. The heterogeneity captured by ν_t allows for correlation in behaviors and outcomes within a period. We assume that ϵ_t is a vector of independent and identically-distributed errors (Extreme Value or Normally distributed depending on the functional form of the contributions to the likelihood function).

Although we do not allow persistence in the time-varying shocks (ν_t), we note an important feature of our dynamic model specification: the inclusion of rich histories of endogenous smoking behavior and health outcomes. As an example, consider a health shock that occurs in period t. As modeled, the random, serially-uncorrelated shocks define period t health events (i.e., the number of cardiovascular disease events, cancer diagnosis, diabetes diagnosis, and body mass index variation) conditional on observed variables, and have no independent effects on subsequent smoking behavior and health outcomes conditional on the histories of these health events. That is, the observed health events in period t absorb all impacts of the health shock on future health. As a counter example, one could imagine a hip fracture in t that may or may not alter an individual's observed health events in t but that leads to a sedentary lifestyle that explains future health events. Our model would not capture this type of serially-correlated period t health shock. However, given that CVD (which includes stroke in our model), cancer, and diabetes explain over two-thirds of deaths in the U.S. (in 2010), we have endogenously modeled the main contributors. Non-infectious airway diseases (or chronic lower respiratory diseases) surpassed stroke as the number three killer in 2008 and currently account for 7% of deaths; we do not model these health events due to lack of consistent information in the FHS data. Similarly, period t smoking shocks (e.g., a stressful year at work or home) do not have independent effects on subsequent smoking behavior or health outcomes, conditional on the rich history of endogenous smoking behavior. We believe that allowing for such dependence is a second order (and computationally challenging) concern.

The permanent heterogeneity, which is correlated across outcomes and over time, is captured by the joint distribution of $\mu = [\mu^{I^S}, \mu^{I^D}, \mu^S, \mu^D, \mu^M]$. The time-varying heterogeneity is defined by the joint distribution of $\nu_t = [\nu_t^S, \nu_t^D, \nu_t^M]$.[28] We could assume that these multivariate distributions are normal, for example, and estimate the cross-equation correlation coefficients along with the coefficients on the observable covariates. However, we do not wish to impose a specific distribution. Rather, we model the UH as random effects and approximate their unknown distributions discretely, esti-

[28]Time-varying heterogeneity does not enter the equations for the variables describing one's initial smoking and health histories entering period two because those variables summarize behavior and outcomes from all periods prior to inclusion in the study.

mating both the discrete mass points along the support of the unobserved components as well as the associated probability weights (termed a Discrete Factor Random Effects (DFRE) method or latent factor method). This flexible estimation technique (Heckman and Singer, 1984; Mroz and Guilkey, 1992; Cunha and Heckman, 2008) does not impose a specific distribution on the error terms as is standard with many maximum likelihood techniques.[29] Additionally, the discrete distributions of the random effects add only a fraction of the additional parameters (and associated loss in degrees of freedom) required by the fixed-effects method (which would be inconsistent in nonlinear models).[30]

The latent factor approach allows individual characteristics that are unobserved by the researcher to impact all jointly estimated equations (in a non linear way) and integrates over their distributions when constructing the likelihood function. Twelve probabilities or densities, presented generally in equations 2-5, form the likelihood function. The contribution of individual n to the likelihood function, unconditional on the

[29]Using Monte Carlo simulation, Mroz (1999) shows that when the true distribution of the error terms is jointly normal the DFRE method performs as well as maximum likelihood estimation assuming normality. When the simulated distribution is not normal, the DFRE method performs better in terms of precision and bias. Mroz (1999) and Guilkey and Lance (2014) describe the econometric properties of the DFRE estimator using Monte Carlo studies.

[30]While the method we use is called a "random effects estimator," it is important to recognize that the estimated "random effect" is not assumed to be independent of endogenous explanatory variables, provided that we model the dependence of such endogenous explanatory variables and the outcome of interest on the random factor. Any explanatory variable that we do not explicitly model as a function of the random factor is assumed to be independent of the random factor.

correlated UH (μ and ν_t), is

$$
\begin{aligned}
L_n(\Theta, \mu, \nu_t, \rho, \psi) \;\; = \;\; & \\
& \sum_{k=1}^{K} \rho_k \Bigg\{ p(E_2 = 0 | \mu_k^{I_1^S})^{\mathbf{1}[E_{n2}=0]} \\
& \times \Big[[1 - p(E_2 = 0 | \mu_k^{I^{S_1}})] p(s_1 = 1 | \mu_k^{I^{S_2}})^{s_{n1}} [1 - p(s_1 = 1 | \mu_k^{I^{S_2}})]^{(1-s_{n1})} \phi^{S_3}(D_{n2} | \mu_k^{I^{S_3}}) \Big]^{\mathbf{1}[E_{n2}>0]} \\
& \times p(\mathrm{CVD}_2 = 0 | \mu_k^{I^{D_1}})^{(1-\mathrm{CVD}_{n2})} [1 - p(\mathrm{CVD}_2 = 0 | \mu_k^{I^{D_1}})]^{\mathrm{CVD}_{n2}} \phi^{D_2}(\mathrm{BMI}_{n1} | \mu_k^{I^{D_2}}) \\
& \times \prod_{t=2}^{T} \sum_{\ell=1}^{L} \psi_\ell \Bigg[p(s_t = 1 | \mu_k^S, \nu_{t\ell}^S)^{s_{nt}} \big[1 - p(s_t = 1 | \mu_k^S, \nu_{t\ell}^S) \big]^{(1-s_{nt})} \\
& \quad \times \prod_{d^1=0}^{2} p(d_t^1 = d^1 | \mu_k^{D^1}, \nu_{t\ell}^{D^1})^{\mathbf{1}[d_{nt}^1 = d^1]} \\
& \quad \times p(d_t^2 = 1 | \mu_k^{D^2}, \nu_{t\ell}^{D^2})^{d_{nt}^2} \big[1 - p(d_t^2 = 1 | \mu_k^{D^2}, \nu_{t\ell}^{D^2}) \big]^{(1-d_{nt}^2)} \\
& \quad \times p(d_t^3 = 1 | \mu_k^{D^3}, \nu_{t\ell}^{D^3})^{d_{nt}^3} \big[1 - p(d_t^3 = 1 | \mu_k^{D^3}, \nu_{t\ell}^{D^3}) \big]^{(1-d_{nt}^3)} \times \phi(d_{nt}^4 | \mu_k^{D^4}, \nu_{t\ell}^{D^4}) \\
& \quad \times \big[1 - p(m_{t+1} = 1 | \mu_k^M, \nu_{t\ell}^M) \big]^{(1-m_{nt+1})} \\
& \quad \times \Big[p(m_{t+1} = 1 | \mu_k^M, \nu_{t\ell}^M) \times \prod_{c=0}^{2} p(m_{t+1}^C = c | \mu_k^{M^C}, \nu_{t\ell}^{M^C})^{m_{nt+1}^C} \Big]^{m_{nt+1}} \Bigg] \Bigg\}
\end{aligned}
$$

where Θ defines the vector of parameters of the model and $p(\cdot)$ represents the logit or multinomial logit probabilities (or densities) of the observed behaviors and outcomes. The vectors ρ and ψ denote mass-point specific estimates of the joint probabilities of the permanent and time-varying heterogeneity, respectively. ρ_k is the estimated joint probability of the k^{th} permanent mass point, which is given by

$$\rho_k = \mathrm{P}(\mu^{I^S} = \mu_k^{I^S}, \mu^{I^D} = \mu_k^{I^D}, \mu^S = \mu_k^S, \mu^{D^1} = \mu_k^{D^1}, \ldots, \mu^{D^4} = \mu_k^{D^4}, \mu^M = \mu_k^M, \mu^{M^C} = \mu_k^{M^C} \;\;).$$

ψ_ℓ is the estimated joint probability of the ℓ^{th} time-varying mass point and is given by

$$\psi_\ell = \mathrm{P}(\nu_t^S = \nu_{t\ell}^S, \nu_{t\ell}^{D^1} = \nu_{t\ell}^{D^1}, \ldots, \nu_t^{D^4} = \nu_{t\ell}^{D^4}, \nu_t^M = \nu_{t\ell}^M, \nu_t^{M^C} = \nu_{t\ell}^{M^C} \;\;).$$

4.4 Variables used in the Empirical Specification

Table 3 summarizes the dependent variables describing the dynamic smoking behavior and health outcomes that we seek to explain. The morbidity measures of disease that we model over time include the number of cardiovascular disease events, cancer diagnosis, diabetes diagnosis, and body mass index, while the mortality measures are

death and cause of death. Appendix Table B1 summarizes the jointly estimated set of twelve behaviors and outcomes, and their determinants, that form the likelihood function. The determinants are divided into pre-determined endogenous variables, exogenous variables, and unobserved heterogeneity (columns 2, 3, and 4). The table serves two purposes: it summarizes sources of identification in relevant equations based on theoretical restrictions and it depicts the correlation across equations and over time coming from both observed and unobserved heterogeneity. The specification of each equation (i.e., how these explanatory variables enter the equations) includes higher moments and interactions of some variables if relevant since the equations represent nth order approximations of the non-linear and dynamic demand and production functions.

The equation system of smoking behavior and health outcomes captures the inherent dynamics over an individual's lifetime. Namely, observed smoking and morbidity outcomes depend on information known at the beginning of the period. Specifically, they depend on one's smoking history and disease history up to the current period. Table 4 provides summary statistics for explanatory variables entering period t (i.e., those that enter the smoking and disease equations). Mortality occurs during the period (or prior to period $t+1$) and depends on the *updated* histories of these variables. That is, the probability of death before period $t+1$ (conditional on not dying prior to period t) depends on one's smoking and disease histories up to period t as well as behavior and disease outcomes in period t. All equation specifications also include polynomials and interaction terms and a flexible time trend as explanatory variables.

4.5 Identification

Having defined all of the equations in our jointly-estimated system, we can now thoroughly discuss identification. We have four sources of identification for estimation of causal marginal impacts: theoretically-justified variables in our behavioral equations that are excluded in outcome equations, variation in the histories of all exogenous explanatory variables captured by our dynamic equation specification, additional exogenous variables that explain the jointly-estimated initial condition equations, and functional forms of the non-linear estimators as well as covariance restrictions on the error structure across equations and over time.

Our main behavioral equation explains smoking in the current period and our main health equations capture morbidity in the period and mortality at the end of the period. Importantly, the smoking equation includes theoretically-justified, supply-side, exogenous variables that influence cigarette demand, namely the mean real price of cigarettes at time t (for 5 cartons or 1000 cigarettes in year 2000 dollars) and real per

Table 3: Dependent Variables in the Jointly-Estimated Set of Equations

Variable	Mean	SD
Smoking behavior		
Smoke at t	0.386	0.487
Morbidity outcomes		
CVD events at t		
0 CVD events (omitted category)	0.946	0.225
1 CVD event	0.042	0.201
2+ CVD events	0.012	0.107
* Ever had CVD event up to t (person-exams)	0.198	0.399
* Ever had CVD event (persons)	0.484	0.500
* Number of CVD events up to t; [0, 14]	0.339	0.843
Cancer diagnosis at t | no cancer up to t	0.013	0.112
* Ever diagnosed with cancer up to t (person-exams)	0.048	0.214
* Ever diagnosed with cancer (persons)	0.179	0.383
Diabetes diagnosis at t | no diabetes up to t	0.011	0.102
* Ever diagnosed with diabetes up to t (person-exams)	0.061	0.240
* Ever diagnosed with diabetes (persons)	0.145	0.352
Body Mass Index at t /10; [1.4, 5.4]	2.644	0.350
Mortality outcomes		
Death hazard by end of t	0.065	0.247
Cause of death in t — death in t		
CVD	0.421	0.494
Cancer	0.263	0.440
Other (omitted category)	0.316	0.465
Initial conditions		
Never smoked	0.272	0.445
Current smoker | ever smoked	0.863	0.344
Years of smoking/10 | current smoker; [0.2, 5.5]	2.512	0.938
Any CVD	0.034	0.182
Body Mass Index/10; [1.7, 4.0]	2.588	0.338

Note: Starred rows are additional statistics, not dependent variables.
Ranges of continuous variables are in brackets.

Table 4: Explanatory Variables Entering Period t

Variable	Abbreviation	Mean	SD	Min	Max
Time-varying variables					
Endogenous					
Smoker in $t-1$	$\texttt{s}_{t-1}$	0.409	0.492	0	1
Years of cessation entering t	$\texttt{C}_t$	5.035	9.799	0	68
Years of duration entering t	$\texttt{D}_t$	13.238	18.317	0	74
Years of experience entering t	$\texttt{E}_t$	21.208	18.585	0	74
1 CVD event in $t-1$	$\texttt{1[CVD}_{t-1}=1\texttt{]}$	0.037	0.189	0	1
2+ CVD events in $t-1$	$\texttt{1[CVD}_{t-1}>1\texttt{]}$	0.011	0.106	0	1
Ever had CVD event up to $t-1$	$\texttt{E\_CVD}_{t-1}$	0.167	0.373	0	1
Number of CVD events up to $t-1$	$\texttt{N\_CVD}_{t-1}$	0.274	0.742	0	12
Cancer diagnosed in $t-1$	$\texttt{CAN}_{t-1}$	0.010	0.098	0	1
Ever diagnosed with cancer up to $t-1$	$\texttt{E\_CAN}_{t-1}$	0.036	0.187	0	1
Diabetes diagnosed in $t-1$	$\texttt{DIA}_{t-1}$	0.010	0.099	0	1
Ever diagnosed with diabetes up to $t-1$	$\texttt{E\_DIA}_{t-1}$	0.052	0.222	0	1
Body mass index in $t-1$	$\texttt{BMI}_{t-1}$	26.436	3.458	13	54
Exogenous					
Systolic blood pressure in $t-1$	$\texttt{SBP}_{t-1}$	136.422	20.538	80	260
Diastolic blood pressure in $t-1$	$\texttt{DBP}_{t-1}$	81.934	11.595	38	140
Cholesterol level in $t-1$	$\texttt{CHO}_{t-1}$	222.752	39.461	81	551
Arthritis in $t-1$	$\texttt{ART}_{t-1}$	0.268	0.443	0	1
`BMI, SBP, DBP` missing		0.063	0.243	0	1
`CHO` missing		0.131	0.338	0	1
Age (years)		60.597	12.170	32	101
Time-invariant variables (Exogenous)					
Education: grade school		0.274	0.446	0	1
Education: some high school		0.161	0.367	0	1
Education: high school degree		0.282	0.450	0	1
Education: some college		0.087	0.281	0	1
Education: college degree		0.098	0.298	0	1
Education: post college		0.098	0.298	0	1
Born outside U.S.		0.170	0.376	0	1
Italian ancestry		0.233	0.423	0	1
Older cohort: Age 50+ at $t=1$		0.309	0.462	0	1

Note: Table summarizes variables entering period t that explain smoking and disease in t. Mortality at end of t depends on updated endogenous variables that include period t smoking and disease. Equation specifications also contain interactions, polynomials, and time trends.

capita expenditures on cigarette advertising (in year 2000 dollars) at time t, denoted by P_t. The cigarette market characteristics do not impact period t morbidity and end of period t mortality outcomes conditional on observed smoking histories. Because individuals live in the same community, cigarette advertising and price vary over time but not across individuals. Thus, we interact these time-varying variables with age and previous smoking status. We omit the levels of these variables due to collinearity with year.[31] Below we discuss the cigarette market data that aid in identification (and represent the traditionally-used theoretically-justified restrictions in single equation or IV reduced form analyses).

Additionally, our main equations for smoking and health are dynamic (i.e., depend on endogenous past outcomes) and we model (estimate) the endogenous smoking behavior and health outcomes for all observed periods spanning up to 46 years of a person's lifetime. This dynamic specification allows the entire history of previous exogenous covariates to serve as implicit instrumental variables for the lagged endogenous variables (Arellano and Bond, 1991; Bhargava and Sargan, 1983). That is, they directly influence past (but not current) behavior. Recall that these dynamic equations depend on health markers (i.e., blood pressure and cholesterol levels) and arthritis, which we treat as exogenous. The smoking equation also accounts for the exogenous history of cigarette prices and advertising.

Because we cannot use dynamic equations to explain our initially-observed endogenous variables (i.e., smoking and health histories up to the point we first observe someone in our sample), we specify unique static equations that are jointly estimated with the main dynamic equations. Theory tells us that smoking demand in period t is a function of one's smoking history, H_t^S. Dynamic substitution confirms that period t smoking demand is a function of initial smoking history, H_2^S, entering period $t = 2$ (the first period we can model smoking behavior dynamically). The equations for each variable defining smoking history entering period $t = 2$, I_2^S, are functions of the initial cigarette

[31] We have run regressions where we include the levels of cigarette prices and advertising expenditure (as well as higher moments) and can confirm that the signs of the effects are in the expected directions: higher prices reduce smoking probabilities and higher advertising expenditures increase them. However, we note the possible collinearity with aggregate year effects and choose to remove the level variables and include a cubic time trend. We retain interactions of the price/advertising variables with age. We capture different effects of cigarette prices and advertising at different ages. At the extreme, adolescents are differently affected by price variation than adults. Likewise, adolescents are differently affected by advertising of cigarettes than are adults. In fact, the literature shows us this (for example, Grossman *et al.* (1993) and Pollay *et al.* (1996)). We also expect price elasticities to differ by income, which we do not have in our data. Yet we know income is correlated with age (and education, for that matter). Lastly, there is ample evidence that smokers and non-smokers are differently sensitive to price. To capture these effects, we interact the supply-side variables with age and lagged smoking status.

market characteristics, P_1, which we observe.[32] Because one's initially-observed disease history also depends on previous smoking behavior, we include cigarette market characteristics in the equations that explain initial cardiovascular disease and body mass, I_2^D, and do not include individual smoking history.

We now return to discussion of the cigarette market data. Appendix C provides intuition on why these these supply shifters would causally influence demand. The argument is relatively straightforward for prices, but one could be concerned that advertising simply shifts demand between brands. There is quasi-experimental evidence that advertising has a causal impact on smoking rates. In the late 1960s and early 1970s a series of government regulations restricted and then banned cigarette ads on television and radio.[33] Appendix Figure C1 shows that real advertising spending rose continuously in the decades before and after these regulations, but fell by forty percent in the five-year period after the regulations were introduced. This reduction is almost surely driven by regulations and not changes in demand or new information about the consequences of smoking (i.e., it is several years after the 1964 Surgeon General's Report). Cigarette consumption for adults fell over five percent during the beginning of this period, while it rose for the decades before and just after the ban. (See Table 2 in American Lung Association (2011); Figure 2 in Harris (1979).) Appendix C also documents that both supply shifters vary substantially over time, and that these changes are often linked to government or judicial policies.

The cigarette market data we use come from a variety of sources discussed in Appendix C. We assemble a time series of cigarette prices and industry-wide advertising by cigarette companies from 1893 to 2009 to help identify the period-by-period smoking equations (using contemporaneous changes in these variables over our observation period) and the initial condition (using values during each persons childhood). In the main equations, we use the data from 1950-1994 to represent time-varying market char-

[32]Specifically, the initial conditions are functions of the advertising expenditures. Recall that individuals in our sample are between 32 and 65 when we first observe them. Most current cigarette smokers began when they were young. The individuals in our data set were "young" (i.e., age 10 to 14 or 15 to 18) in different calendar years. We use the advertising expenditure levels in the years individuals were "young" to shift initial smoking behavior. These pre-1940 advertising levels have no independent effect on smoking behavior in 1948 and beyond (to 1996) when we estimate the biennial smoking patterns of individuals in our sample. We use different age groupings because ever smoked may be correlated with smoking experimentation at younger ages while duration smoking conditional on being a current (period $t = 1$) smoker likely reflects continued smoking into adolescence rather than simply experimentation. We wanted to capture these two likely age scenarios. We do not use cigarette prices in these initial condition equations because we did not have enough variation in the early years to identify effects. See Figure C.2 in the Appendix.

[33]In 1968 the FCC required TV and radio stations to air anti-cigarette commercials if they also broadcast cigarette ads. In 1971 federal law banned all cigarette ads on TV and radio.

acteristics affecting contemporaneous smoking behavior. While these supply variables have statistically-important effects on smoking behavior, we also find empirical support for omitting them from the health equations.[34] We use the data through 2009 in simulations, based on our estimated model, until death. We use the data from 1895-1939 to account for variation in the cigarette market that may have influenced smoking behavior early in one's life and, in particular, smoking initiation. Recall that individuals are different ages (i.e., 32 to 65) when we originally observe them. Hence, they were "young" at different points in history. We argue that variation in cigarette prices and advertising *when individuals in the FHS were in their teenage years* may explain propensities to begin smoking. Indeed, we show that real per-capita cigarette advertising expenditures *averaged over the years in which an individual was between 10 and 14 (or 15 and 18)* positively predict the initially-observed smoking history when an individual enters the FHS. Our "advertising expenditure during the ages of 10 and 14 (or 15 and 18)" variable is not perfectly correlated with age or calendar year because ages at the first exam vary and the first exam of each individual in the FHS was administered sometime between 1948 and 1953.[35] So, for example, when the Supreme Court dissolved the Tobacco Trust in 1911 cigarette prices fell and advertising rose. These changes are likely to have a different effect on older men of our sample (who were adults at the time of the breakup) relative to younger men. Support for this sort of differential effect is presented in the literature review in Appendix C.

In addition to the time-varying cigarette market characteristics that serve as a source of exogenous variation that impacts the initial conditions, we also include information about sibling structure in the initial condition equations. We argue that smoking initiation, which generally occurs at young ages, may be influenced by siblings or one's ordering among siblings (Gilleskie and Strumpf, 2005; Kelly *et al.*, 2011; Black *et al.*, 2015). These variables include the number of siblings, an indicator of being an only child, a linear birth order variable, and an indicator of being a first-born child. The coefficients

[34]These variables are jointly significant in the smoking equation and not significant in the health equations conditional on smoking history. To demonstrate that the supply-side variables have an economically meaningful effect on individual choices, we used the estimates discussed in the next section to simulate smoking behavior when we forced the advertising and price variables to be one standard deviation above and below its observed value. We find, for example, that smoking propensity is 13 percentage points lower when we increase cigarette prices by one standard deviation (results available upon request). These figures imply an elasticity close to -1 at age 50. Note that this elasticity represents the dynamic influence of a price increase each year. It reflects the per period change in smoking behavior as well as the cumulative impact of past changes in smoking behavior induced by the price increase.

[35]A historical cigarette price variable was similarly constructed to reflect average prices during an individual's teenage years. This variable did not satisfy the identification criteria and is, therefore, not used in the initial conditions equations.

on these variables are jointly significant in the initial condition equations, and are not significant in the main equations once we control for smoking and health histories. We provide summary statistics for variables that capture the cigarette market and sibling structure in Table 5. To summarize, our model parameters are (over-)identified using theoretically-relevant exclusion restrictions where appropriate, the entire history of exogenous time-varying variables given the dynamic equation specification, covariance restrictions associated with estimation of the correlated UH, and non-linear estimators.

Table 5: Variables that Serve in Identification

Variable	Mean	SD	Min	Max
Cigarette market (values in year 2000 $)				
Mean cigarette price for 5 cartons in year t - using years 1996-2009*	212.87	49.65	129.52	283.79
Mean cigarette price for 5 cartons in year t - using years 1950-1994	87.45	13.84	70.07	125.97
Advertising expenditure per capita in year t - using years 1996-2009*	3.07	1.35	1.29	5.10
Advertising expenditure per capita in year t - using years 1950-1994	6.57	1.93	2.83	10.60
Advertising expenditure per capita at ages 15-18 - using years 1899-1939	2.21	1.85	0.01	5.92
Advertising expenditure per capita at ages 10-14 - using years 1895-1935	1.82	1.80	0.07	5.92
Sibling structure				
Number of siblings	4.45	2.89	0	20
Only child	0.04	0.20	0	1
Birth order (up to 5th)	2.75	1.49	1	5
First born child	0.27	0.45	0	1
Sibling information missing	0.18	0.38	0	1

Note: * 1996-2009 values are used when simulating behavior beyond our sample observation period. The price and expenditure time series are depicted in Appendix C.

5 Results

Our empirical analysis begins with FIML estimation of the 12 equation system (see Appendix Table B1) — representing smoking behavior (1 equation), morbidity outcomes (4 equations), mortality outcomes (2 equations), and initial conditions (5 equations) — that allows for individual-level correlation across equations and over time. We discuss parameter estimates from two versions of the model: one that allows for correlated individual-level UH (our preferred FIML model) and one that does not. The model without correlated UH amounts to estimation of a single equation separately with endogenous regressors treated as exogenous. Having estimated the parameters, we demonstrate the ability of our preferred model to fit the observed data. We also show that the model is able to predict out-of-sample (post-1996) ages of death for the 12 percent of individuals in the estimation sample who had not died by 1996. We then use the model to simulate morbidity and mortality outcomes for a wide array of smoking patterns that could be exhibited by individuals over the life course. These results, compared to those from models typically used in this literature, demonstrate the differences we find regarding the impact of smoking cessation on morbidity and mortality.

5.1 Parameter Estimates

Our preferred model (labeled 'FIML with correlated UH') is the one that explicitly introduces and estimates the correlated UH that might impact both smoking behavior and morbidity/mortality in order to capture the selection inherent in smoking history variation and the confounding associated with health outcomes variation. Table 6 presents estimates of and standard errors on coefficients of selected determinants of smoking. It is difficult to infer marginal effects from this table, since the variables are part of a dynamic (e.g., smoking history is defined by lagged smoking as well as years of smoking duration, experience, and cessation) and larger system (e.g., smoking history impacts health history which also influences current smoking behavior).[36]

As expected, the point estimate on lagged smoking is large and significant, suggesting state dependence in smoking. Furthermore, the longer one has smoked the more

[36]We calculate (and discuss in the next sections) marginal effects using simulation techniques to account for the dynamic feedback and large number of polynomials and interactions in the specification. We focus our discussion of the effects of smoking and health histories. Other exogenous determinants included in each equation are age, education, ancestry, origin, cohort, and year trends (and results are available from the authors). Estimates of the correlated UH contributions and their distributions are presented in Appendix Table B2. Parameter estimates for other equations are also in Appendix B: cause of death (B3); morbidity outcomes (B4, B5, and B6), and initial conditions (B7).

likely he is to continue smoking (at a diminishing rate). With regard to the endogenous health variables, blood pressure has a significant effect (negative for SBP and positive for DBP) and a CVD event in the previous period reduces the probability of smoking. Body mass and cholesterol levels in the previous period do not impact the probability of smoking.

For completeness, we also provide coefficient estimates and standard errors from a model with no correlated UH (i.e., the equations are estimated separately and coefficients on endogenous variables reflect bias associated with selection and confounding). This model (labeled 'single equation without correlated UH') extends the models often used by practitioners and policymakers to measure the impact of smoking on health outcomes and to calculate the benefits of smoking cessation by including a richer description of observed smoking and disease histories. Yet, it does not attempt to capture potential (permanent and time-varying) correlated individual UH and, therefore, may still produce biased impacts of endogenous smoking and health histories despite providing an improved fit. Indeed, changes in the significance and signs of endogenous variables is evident in the table. A comparison of the estimates from the model without correlated UH (columns 1 and 2) and with correlated UH (columns 3 and 4) reveals differences in the significance and functional relationship between one's smoking history and his propensity to smoke currently. In particular, notice that variation in years of smoking cessation no longer significantly impacts contemporaneous smoking when correlated UH is introduced, while additional years of smoking duration and experience increase the probability of smoking at a decreasing rate.[37] (The sizes of the coefficients also exhibit differences, but changes in marginal effects are difficult to assess at this point. We further examine marginal effects, and address the role of UH, after we introduce simulations from the models.)

Before proceeding, we note the significance of the cigarette market variables in the smoking equations. These variables, serving as a source of identification, are not significant when included in the morbidity and mortality equations conditional on one's smoking history. The cigarette price and advertising variables impact smoking but appear to do so in an unexpected direction. We estimated the model using prices and advertising expenditures in level terms and found the expected signs: prices have a negative effect on the smoking probability while advertising expenditure has a positive effect. We do not include these level variables in our main specification because it is hard to disentangle these terms, which vary only over time, from temporal effects.

[37]Determination of significance involves a joint test when variables enter as polynomials or interactions.

Table 6: Selected Parameter Estimates: Smoking in period t

	Single equation without correlated UH			FIML with correlated UH		
Variable	Estimate	Std. error		Estimate	Std. error	
Smoker in $t-1$, $\mathtt{s}_{t-1}$	0.784	0.337	**	1.631	0.871	*
Years of cessation, $\mathtt{C}_t$	0.260	0.061	***	-0.109	0.086	
$\mathtt{C}_t^2/100$	-3.146	0.553	***	-0.576	0.653	
Years of duration, $\mathtt{D}_t$	0.110	0.010	***	0.151	0.025	***
$\mathtt{D}_t^2/100$	-0.134	0.021	***	-0.156	0.041	***
Years of experience, $\mathtt{E}_t$	0.105	0.011	***	0.197	0.022	***
$\mathtt{E}_t^2/100$	-0.115	0.021	***	-0.207	0.042	***
$\mathtt{1[CVD}_{t-1}=1\mathtt{]}$	-0.447	0.213	**	-0.538	0.296	*
$\mathtt{1[CVD}_{t-1}>1\mathtt{]}$	-0.684	0.324	**	-1.216	0.557	**
$\mathtt{CAN}_{t-1}$	-0.620	0.452		-0.389	0.674	
$\mathtt{DIA}_{t-1}$	0.129	0.495		0.386	0.574	
$\mathtt{E\_CVD}_{t-1}$	-0.929	0.285	***	-1.394	0.355	***
$\mathtt{N\_CVD}_{t-1}$	0.242	0.112	**	0.404	0.113	***
$\mathtt{E\_CAN}_{t-1}$	-0.118	0.487		-0.135	0.540	
$\mathtt{E\_DIA}_{t-1}$	-0.228	0.410		-0.578	0.508	
$\mathtt{E\_CVD}_{t-1}$ * $\mathtt{S}_{t-1}$	0.714	0.249	***	0.893	0.330	***
$\mathtt{E\_CAN}_{t-1}$ * $\mathtt{S}_{t-1}$	0.232	0.544		0.110	0.662	
$\mathtt{E\_DIA}_{t-1}$ * $\mathtt{S}_{t-1}$	0.592	0.467		0.376	0.695	
$\mathtt{BMI}_{t-1}$	-0.069	0.104		-0.193	0.238	
$\mathtt{BMI}_{t-1}^2/100$	0.078	0.193		0.220	0.441	
$\mathtt{SBP}_{t-1}$	0.039	0.015	***	0.044	0.020	**
$\mathtt{SBP}_{t-1}^2/100$	-0.013	0.005	**	-0.015	0.007	**
$\mathtt{DBP}_{t-1}$	-0.068	0.027	**	-0.067	0.033	**
$\mathtt{DBP}_{t-1}^2/100$	0.036	0.016	**	0.035	0.020	*
$\mathtt{CHO}_{t-1}$	0.006	0.005		0.009	0.007	
$\mathtt{CHO}_{t-1}^2/100$	-0.001	0.001		-0.002	0.002	
BMI, SBP, DBP missing	-1.000	1.559		-2.799	3.342	
CHO missing	0.918	0.673		1.327	0.908	
Cigarette price at t * $\text{Age}_t/10$	0.327	0.100	***	0.467	0.151	***
Cigarette price at t squared/100 * $\text{Age}_t/10$	-0.231	0.058	***	-0.314	0.088	***
Cigarette price at t * $\text{Age}_t/10$ * $\mathtt{S}_{t-1}$	0.029	0.009	***	0.019	0.012	
Ad expenditure at t * $\text{Age}_t/10$	-0.300	0.083	***	-0.273	0.105	***
Ad expenditure at t * $\text{Age}_t/10$ * $\mathtt{S}_{t-1}$	0.098	0.072		0.009	0.093	
Constant	2.281	1.827		4.916	4.232	

Note: Specifications also include controls for age, education, ancestry, origin, cohort, and year trends. Standard errors are in parentheses.
*** indicates joint significance at the 1% level; ** 5% level; * 10% level.

Table 7 presents estimates of and standard errors on the coefficients of the observed determinants of death by the end of period t conditional on being alive in period t. We also examine cause of death conditional on dying in Appendix Table B3. Note that the specification reflects updated values of the endogenous variables (i.e., includes the period t behavior and health events). As with the previous table where we cannot draw firm conclusions simply by examining the coefficients, we nonetheless can discuss some interesting findings. First, while it may appear that current smoking reduces the probability of death, one should note that each year of smoking duration significantly increases the probability of death and current period smokers are likely to have a long history of smoking. Second, disease is an important predictor of death, as expected. Cardiovascular disease events and cancer diagnosis in the current period increase the probability of death by the end of the period. While such events in the previous period specifically do not have a statistically significant effect, having ever had these diseases increases the death hazard. Higher levels of current health markers, such as body mass, diastolic blood pressure, and cholesterol, predict eminent death. Third, it appears that years of smoking cessation has no statistically significant effect on the probability of dying, but this interpretation ignores the indirect channels embedded in the entire system of equations. Continued smoking (i.e., a positive number of years of smoking duration) significantly increases the probability of both morbidity and mortality. Thus, smoking cessation (which sets duration to zero) eliminates an important detrimental impact. Of course, experience is still positive after a quit. These dynamic effects will be clearer when we simulate behavior of the individuals under different lifetime smoking patterns.

Differences in the coefficient signs and significance across the models with and without correlated UH are more apparent in the cause of death equation. Current smoking significantly predicts death due to cancer, conditional on dying, and has a large but imprecise impact on death due to cardiovascular disease. Diagnosis of cancer in the current period or the previous period does explain cancer deaths before the next period and cardiovascular events in the current period explain CVD deaths.

Appendix Tables B4, B5, and B6 present estimates of and standard errors on the coefficients of the observed determinants of endogenous disease events in period t: the number of cardiovascular events (B4); cancer diagnosis conditional on no diagnosis prior to the current period (B5); diabetes diagnosis conditional on no diagnosis prior to the current period (B5); and the continuous health marker body mass index (B6). These per-period disease events define health history variables that explain the dynamic smoking patterns of individuals over a lifetime. Because we allow for correlation between these endogenous events and smoking behavior and mortality outcomes, we reduce bias

Table 7: Selected Parameter Estimates: Mortality by end of period t

	Single equation without correlated UH			FIML with correlated UH		
Variable	Estimate	Std. error		Estimate	Std. error	
Smoker in t, $\mathtt{s}_t$	-0.624	0.153	***	-1.428	0.506	***
Years of cessation, $\mathtt{C}_t$	0.009	0.010		0.004	0.012	
$\mathtt{C}_t^2/100$	-0.018	0.024		-0.012	0.026	
Years of duration, $\mathtt{D}_t$	0.074	0.013	***	0.101	0.023	***
$\mathtt{D}_t^2/100$	-0.102	0.023	***	-0.132	0.033	***
Years of experience, $\mathtt{E}_t$	-0.014	0.009		-0.008	0.011	
$\mathtt{E}_t^2/100$	0.031	0.018	*	0.026	0.021	
$\mathtt{1[CVD}_t = 1\mathtt{]}$	0.476	0.112	***	0.700	0.273	**
$\mathtt{1[CVD}_t > 1\mathtt{]}$	0.692	0.193	***	0.745	0.278	***
$\mathtt{CAN}_t$	0.874	0.175	***	1.224	0.349	***
$\mathtt{DIA}_t$	0.413	0.275		0.688	0.435	
$\mathtt{1[CVD}_{t-1} = 1\mathtt{]}$	0.021	0.130		0.019	0.139	
$\mathtt{1[CVD}_{t-1} > 1\mathtt{]}$	0.460	0.217	**	0.437	0.242	*
$\mathtt{CAN}_{t-1}$	0.288	0.219		0.342	0.254	
$\mathtt{DIA}_{t-1}$	0.095	0.268		0.099	0.277	
$\mathtt{E\_CVD}_{t-1}$	0.396	0.107	***	0.377	0.116	***
$\mathtt{N\_CVD}_{t-1}$	0.151	0.046	***	0.162	0.047	***
$\mathtt{E\_CAN}_{t-1}$	0.604	0.122	***	0.667	0.144	***
$\mathtt{E\_DIA}_{t-1}$	0.555	0.111	***	0.445	0.123	***
$\mathtt{BMI}_t$	-0.276	0.079	***	-0.350	0.097	***
$\mathtt{BMI}_t^2/100$	0.432	0.143	***	0.497	0.172	***
$\mathtt{SBP}_t$	0.015	0.012		0.015	0.013	
$\mathtt{SBP}_t^2/100$	-0.005	0.004		-0.005	0.005	
$\mathtt{DBP}_t$	-0.053	0.022	**	-0.052	0.023	**
$\mathtt{DBP}_t^2/100$	0.040	0.013	***	0.040	0.014	***
$\mathtt{CHO}_t$	-0.021	0.005	***	-0.019	0.007	***
$\mathtt{CHO}_t^2/100$	0.004	0.001	***	0.004	0.001	***
BMI, SBP, DBP missing	-4.770	1.136	***	-6.164	1.442	***
CHO missing	-0.921	0.579		-0.708	0.740	
$\mathtt{ART}_t$	-0.035	0.070		-0.049	0.074	
Constant	-1.557	1.740		1.285	2.288	

Note: Specifications also include controls for age, education, ancestry, origin, cohort, and year trends. Standard errors are in parentheses.
*** indicates joint significance at the 1% level; ** 5% level; * 10% level.

in the estimated marginal impacts of interest in our study. The results presented in the tables indicate that the parameters differ in sign as well as economic and statistical significance between the specifications. Marginal impacts are described below using simulations from the estimated model.

5.2 Ability of Empirical Model to Fit the Observed Data

Given the many features of our dynamic model, and the associated inability to fully comprehend a variable's impact by looking only at coefficients, we simulate smoking behavior and health outcomes using the estimated model. To conduct our simulations, we replicate the exogenous variables of each individual in the estimation sample R=50 times. For each replication we simulate the initial conditions using the estimated reduced-form equations. We then simulate smoking behavior and health outcomes (morbidity and mortality) for one period using the dynamic equations. That is, we use the estimated model and draws from the estimated correlated UH error distribution and the i.i.d. error distributions to simulate the endogenous outcomes. We then update the smoking and disease histories and simulate behavior and outcomes in the subsequent period for those who are not simulated to die. Analogously, we simulate outcomes until everyone in the simulation sample has died.[38] Note that our simulations use the parts of the model that are explained by observed heterogeneity (i.e., the estimated coefficients and observed exogenous and endogenous variables) as well as unobserved heterogeneity captured by the estimated permanent UH, the estimated time-varying UH, and the random error draw for each outcome each period. Predictions only capture what is explained by observed variation. All three sources of UH play an important role in our simulations because they determine current behavior and outcomes that impact subsequent behavior and outcomes through our system of dynamic, correlated equations. Thus, the fit we capture is one that demonstrates the comprehensive ability of our model to explain life-cycle observations.

The top panel of Table 8 displays the distribution of age and cause of death for the observed sample (used in estimation of the model) and the simulation sample (generated from the estimated model). Our purpose here is to show how well our estimated dynamic data generating process fits the observed data. Our preferred model captures the age and cause of death distributions quite well, and correctly simulates the percent who died by the end of 1996 (or 22 periods of the model).

[38]In cases where we simulate someone to survive who, in the data, is observed to die, we must impute values of his exogenous variables. Age is increased by two years every period that the replicated individual is alive. The health markers (i.e., systolic and diastolic blood pressure and cholesterol levels) are imputed based on an individual's last observed values and averages among individuals his age.

Table 8: Age and Cause of Death: Observed and Simulated Data

Sample	Percent died	Age of death distribution (percentile) Mean	10th	25th	50th	75th	90th	Cause of death CVD	Cancer	Other
Deaths observed through 1996 (right-censored; estimation sample)										
Observed	87.8	72.6	58	66	74	80	85	42.6	26.3	31.6
Simulated	89.0	72.3	58	66	73	79	85	43.9	26.2	30.0
Deaths observed through 2009										
Observed	100.0	74.4	59	68	76	82	88			
Simulated	100.0	74.0	59	67	75	82	88			

Note: Observations are right-censored if the individual (observed or simulated) has not died through 1996 (two years after exam 23).

In the lower panel of Table 8, we evaluate our model's ability to predict age of death outside of the sample used in estimation of the model. When we began this project, we were granted access to the FHS data from NHLBI through 1996. However, the original cohort of the FHS continued to be followed. We recently acquired age of death (but not cause of death) for the original sample through 2009. Using our estimated model to simulate smoking behavior and health outcomes until death for the replicated sample, we can determine how well our model captures the true observed age of death of individuals used in estimation who had not died through 1996. By 2009, everyone in our estimation sample had died. The average age of death was 74.4 years. When we use our model of lifetime smoking, morbidity, and mortality to simulate the sample until death, we find a simulated average age of death of 74.0. This ability of our model to match the observed out-of-sample death ages gives us additional confidence that the model explains lifetime smoking and health very well.

Comparisons of the simulated data to the observed data provide measures of how well our model captures the dynamic behaviors and outcomes of interest. Up to this point, we have described the FHS data over time, with calendar year or period (two years) being the unit of observation per person. However, the main purpose of the empirical model is to explain smoking behavior and health outcomes over an individual's lifetime (while controlling for aggregate variation that affects all individuals over time). From this point forward we discuss the model in terms of its ability to fit lifetime profiles of smoking, morbidity, and mortality by age.

First, we graphically compare the age-specific outcomes of the simulation sample, for those years when the replicated individual is observed to be alive, with the outcomes of the estimation sample. Figure 2 presents the model fit for each of the dependent

variables (excluding cause of death) by age.[39] We slightly overstate smoking behavior at younger ages and understate it at older ages, but generally capture the overall decline in smoking by age quite well. Our mortality model fits very well even into ages above 70 when observed death rates are less precise due to small sample sizes. The model also accurately predicts, by age, small probability events such as one or more CVD events, cancer diagnosis, and diabetes diagnosis. We also fit body mass as measured by BMI very well. Additionally, chi-squared tests indicate that we cannot reject that the averages of the simulated data from our data generating process are equivalent to those of the observed data.

Second, we use the simulated data to assess the ability of the model to capture important smoking transitions such as quitting and relapse, rather than simply levels of smoking. Figure 3 displays the probability of quitting smoking at each age conditional on smoking in the previous period (two years earlier on average). Again, our model does an exceptional job of capturing the quitting trend by age.[40] The probability of relapse is more difficult to graph, yet average simulated relapse rates of 44.1 percent among men who quit smoking is not statistically different from the observed relapse rate of 37.5 percent.[41] Among those who relapse, the observed and simulated mean length of smoking cessation between spells of smoking is 3.3 and 4.2 years, respectively; the mean age of relapse is 53.9 and 54.6 years.

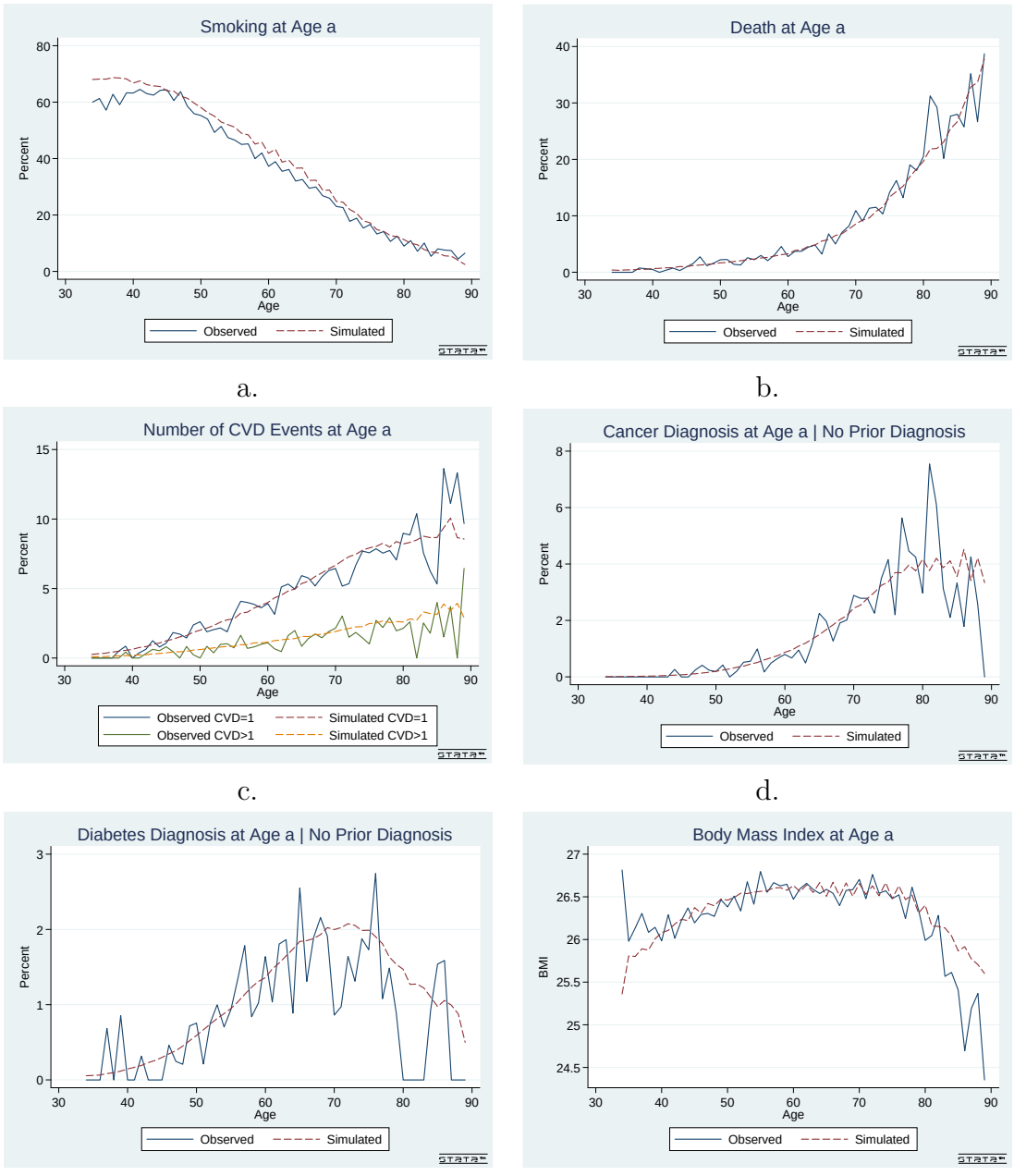
Appendix D shows the empirical importance of modeling correlated individual UH. In our model UH captures unobservables, both those that are common across a lifespan (like genetics) and ones that vary over time (such as unobserved stress), which would bias the estimated impacts of smoking and health histories. We show there are important differences in smoking behavior, mortality distribution, and the link between smoking and death across the various permanent UH types. In addition to providing a better fit through reduced selection and endogeneity, these results highlight the importance of distributional issues that can factor into the choice of policy variables such as cigarette taxes or regulations.

[39]We do not describe the fit of the five reduced-form (not dynamic) initial condition equations; this information is available from the authors. The initial conditions, which are correlated with the permanent individual UH, are estimated jointly with the per-period equations in order to aid in identification of the UH distributions. Appendix Table B7 provides coefficient estimates for the initial condition equations.

[40]Note that behavior at the youngest and oldest ages reflects small sample sizes for these groups in our data.

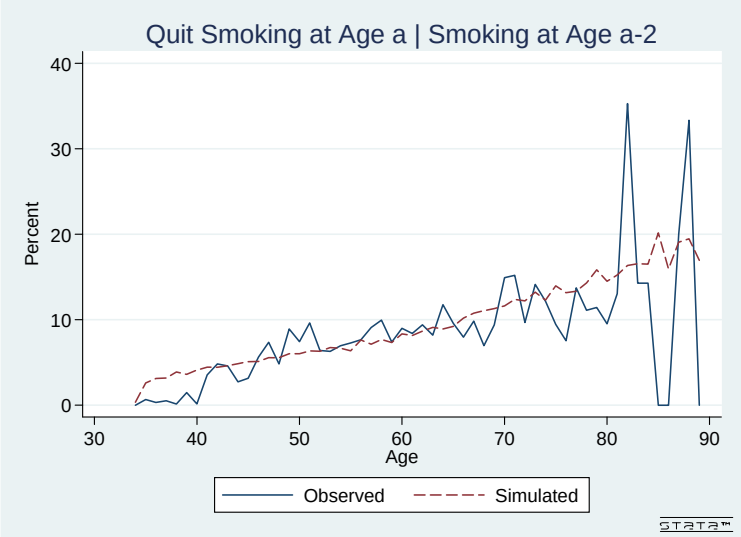
[41]In this calculation, we condition on quits observed after age 30.

Figure 2: Model Fit of Smoking, Morbidity, and Mortality Outcomes



Note: Solid lines indicate averages of observed sample behavior and outcomes. Dashed lines indicate predicted probabilities from simulation of our preferred FIML model with correlated UH. Figures depict rates or levels of: a) smoking, b) mortality, c) cardiovascular disease, d) cancer, e) diabetes, and f) body mass.

Figure 3: Model Fit of Quit Behavior conditional on Smoking in Previous Period



Note: Solid line indicates average of observed sample behavior. Dashed line indicates average of simulated behavior from our preferred model with UH.

5.3 Simulated Lifetime Smoking Scenarios

Having demonstrated the ability of our estimated model to capture observed behavior and outcomes accurately and to predict mortality well out-of-sample, we now turn to assessment of the impact of smoking on morbidity and mortality. To do so, we conduct several simulations in which we impose different lifetime smoking patterns. For this analysis, we simulate smoking behavior and health outcomes until death. That is, there is no right censoring of the data; every simulated individual is observed until death.[42]

Row 1 of Table 9 presents the age and cause of death distributions when all individuals in the simulation sample (i.e., a sample composed of R (=50) replications of exogenous variables of the N (=1464) sample observations) are simulated to never smoke (where death is determined by the model based on sequential updates of morbidity throughout the lifecycle). Average age of death is 75.5 years. Row 2 describes the mortality outcomes of the simulation sample when everyone is assumed to smoke from age 18 through death.[43] Age of death is, on average, 71.2, or 4.3 years earlier than

[42]Simulations are conducted until the year 2024 (or hypothetical exam 42). Recall that individuals are age 30-62 in 1952. Every replicated individual is simulated to die. In the baseline simulation (where we impose no smoking pattern but use the model to simulate smoking histories), only 0.8% of individuals are simulated to die after age 100. In all simulations in which someone is simulated to die after age 100, we assume death occurs at age 100 for summary calculations.

[43]For these simulations, initial smoking duration is set to initial age - 18. We also simulate the

a non-smoker. For comparison, the (biased) unconditional difference-in-means in age of death between lifelong smokers and nonsmokers is 9.3 years using the right-censored FHS estimation sample of men.[44] We contend that the observed and unobserved heterogeneity that we model reduces the bias in estimates of the impact of life cycle smoking behavior and morbidity on mortality. Additionally, CVD accounts for 16.6 percent (or (46.4 - 39.8 =) 6.6 percentage points) more deaths for smokers versus non-smokers, while death from cancer rises 40 percent (or (28.0-20.0=) 8.0 percentage points).

Table 9: Age and Cause of Death: Simulated Data by Smoking Scenario

		Age of death distribution (percentile)					Cause of death		
Smoking scenario	Mean	10th	25th	50th	75th	90th	CVD	Cancer	Other
Never smoked	75.5	61	69	77	83	89	39.8	20.0	40.2
Smoked continuously from age 18	71.2	57	64	71	79	85	46.4	28.0	25.5
Started smoking at 18 and quit at:									
Age 40	75.9	61	69	77	84	90	38.7	28.0	33.3
Age 50	74.2	55	67	76	83	89	39.6	25.4	35.0
Age 60	71.6	57	61	72	80	87	42.1	23.7	34.1
25 Years of smoking:									
Ages 13-48	74.3	56	67	76	83	89	39.4	25.8	34.9
Ages 23-58	72.8	57	63	74	82	88	41.8	23.6	34.6
Continuous smoking with a gap between:									
Ages 30-35	71.0	60	66	71	77	82	52.5	27.7	19.8
Ages 40-45	72.2	62	68	73	78	82	53.1	27.8	19.2
Ages 50-55	73.0	57	69	75	79	83	52.2	27.3	20.5

More importantly for policy purposes, we calculate the expected gain (in life years) of quitting smoking at particular ages. Relative to smoking continuously, quitting smoking at ages 60, 50, and 40 implies an increase in longevity of 0.4, 3.0, and 4.7 years, respectively. The heavily cited Doll, *et al.* work finds increases of 3, 6, and 9 years. Our findings suggest that these commonly-used figures are inflated by nearly 50 percent. Note that while quitting by age 40 produces a lifespan distribution that is almost identical to that of never smokers, the likelihood of death by cancer is still

smoking scenarios in Table 9 with smoking initiation at age 13 rather than 18. The age of death distribution is shifted to the left slightly (i.e., younger). Results are available from the authors.

[44]Restricting our calculations to death before 1998 (i.e., the period when we observe both smoking and health in our data), the difference in the simulated death ages of never smokers and continual smokers is 3.8 years.

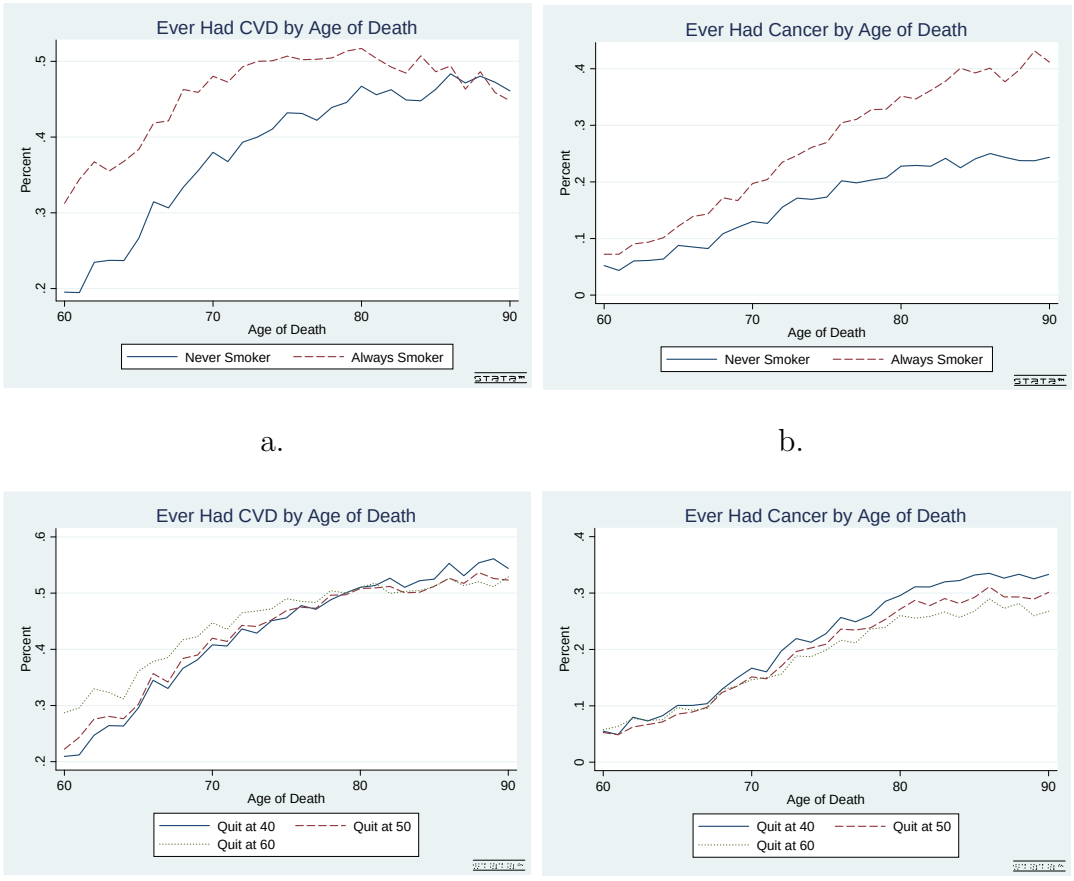
proportionately higher. Interestingly, quitting smoking lowers the probability of death by CVD relative to cancer or other causes, yet the history of smoking among former smokers still manifests itself in a higher probability of a cancer-related death. Related, Taylor *et al.* (2002) find that quitting smoking at age 35 extends life expectancy of men by 6.9 to 8.5 years relative to those who continue to smoke. They also find that quitting earlier is more beneficial than quitting later. While their empirical results are based on a larger sample of individuals than ours, it is not a nationally representative sample and only 20 percent of the sample had died during the study period. Our findings using their same specifications are similar to theirs. However, when we evaluate results using our preferred model with additional sources of heterogeneity and more observed deaths, our estimates of the benefits of quitting decrease significantly.

Also detailed in Table 9, we examine differences in age and cause of death for individuals with the same smoking experience but different ages of initiation (and hence also different ages of quitting, conditional on survival). We find that starting smoking later in life (i.e., age 23 versus 13) leads to a lower life expectancy by 1.5 (=74.3 - 72.8) years. Yet, death by cancer is more likely when smoking is initiated earlier and death by CVD is more likely when smoking occurs at older ages.

Lastly, we examine the impact of smoking cessation followed by relapse. We simulate individuals to have a 5-year reprieve from smoking at the ages of 30-35, 40-45, or 50-55. In all simulations the individuals began smoking at age 18 and smoked until death (following the single 5-year cessation period). A small spell of cessation has no statistically significant difference on life expectancy (from that of continuous smokers) if it occurs at younger ages. If the 5-year cessation occurs later, there is a slight increase in average ages of death. Death attributable to cancer or CVD receives similar weights, relative to other causes, regardless of the age of cessation. Interestingly, CVD deaths are 13 percent more prevalent (about 6 percentage points) for individuals with a 5-year gap in smoking than those who have smoked continuously since age 18.

While both age of death and cause of death (i.e., mortality) vary by smoking behavior over the lifecycle, we also evaluate differences in *quality* of life (i.e., morbidity). We calculate the proportion of our simulation sample who are ever diagnosed with cancer or ever experience a cardiovascular disease incident and report these proportions by lifetime smoking pattern. However, we do not present disease incidence at a given age due to the problem of dynamic selection: given our mortality results, those simulated to never smoke have more life years in which to potentially become ill. Rather, Figure 4 presents simulated lifetime disease incidence by simulated age of death across different smoking patterns. Our results suggest large differences in lifetime disease incidence by smoking pattern. For example, among those who die at age 70, our simulations show

that smoking continuously from age 18 produces an 8 percentage points higher likelihood of ever being diagnosed with cancer than never smoking (panel b). While the cancer gap between always and never smokers grows as age of death increases, we find that the fairly constant gap associated with cardiovascular disease disappears at death ages above 80 (panel a). Differences in age of quitting smoking do not explain CVD and cancer incidence of smokers (panels c and d), except for those with long longevity. Smokers who live longer (older than age 75 or 80) exhibit higher incidence of CVD and cancer in their lifetimes if they quit earlier. These findings suggest that policy aimed at smoking prevention more so than quits is beneficial if the goal is to improve quality-adjusted lifespans.



a. b.

c. d.

Note: Each figure represents the simulated incidence of ever being diagnosed with or experiencing the respective event by the simulated age of death and by different counterfactual smoking histories.

Figure 4: Morbidity Outcomes by Age of Death

6 Discussion

Based on our findings, we concur with the universal evidence in the medical and economics literatures that smoking is detrimental to health measured by both morbidity and mortality. Our results suggest, however, that the mortality consequences of smoking typically cited and used by policymakers are overstated by as much as 50 percent (i.e., a difference in age of death of 4.3 years on average versus the existing evidence of 9.3 years). As an example of why accurate estimates are important, consider the U.S. Food and Drug Administration's recent evaluation of the costs and benefits of smoking cessation in order to determine the appropriateness of a particular regulatory action that will impact smoking behaviors (Chaloupka *et al.*, 2014). The economic analysis has received much attention due to the suggestion that the benefits be discounted to reflect a smoker's lost happiness that would accompany smoking reductions; less attention has been paid to the calculations of the morbidity and mortality consequences of reduced smoking. Irrespective of the discounting issue, the "inflated" figures being used to evaluate policy and regulatory decisions could lead to costly implementation with significantly reduced impacts.

Our results reiterate the importance of quitting at younger ages, with improvements in life expectancy of, for example, 4.7 years versus 3.0 years if the cessation occurs at age 40 versus 50. Additional new findings show the importance of relapse avoidance: short spells of non-smoking followed by relapse has very little benefit. In addition to policies that encourage quitting, emphasis should be placed on quit maintenance. Cessation programs without follow-up support for former smokers will not be effective in extending life if relapse occurs. Our model also demonstrates that rates of death attributed to CVD and cancer as well as lifetime incidence of disease (i.e., morbidity) differ by lifetime smoking patterns. By applying our findings and accurate pecuniary costs and measures of pain and suffering by disease, more comprehensive cost-effectiveness analyses can be used to evaluate smoking policy recommendations expected to impact smoking behaviors differently.

Admittedly, the nature of smoking has changed for recent generations compared to that of the FHS original cohort. Changes include the age of initiation, the use of filters, the levels of tar, the modes of smoking, etc. Some of these changes are technological and might change underlying estimated parameters; others are behavioral and should support our findings. The offspring and the third generation of the FHS original cohort make it possible for us to follow up our analysis using the same techniques with a more recent cohort. We emphasize the importance of continued and new data acquisitions, like the FHS, that follow individuals at frequent intervals over a long period of time.

References

Adda, J. and Lechene, V. (2001). Smoking and Endogenous Mortality: Does Heterogeneity in Life Expectancy Explain Differences in Smoking Behavior, Discussion Paper Series, Department of Economics, University of Oxford, ISSN 1471-0498.

— and — (2013). Health Selection and the Effect of Smoking on Mortality. *Scandinavian Journal of Economics*, **115**, 902–931.

American Lung Association (2011). Trends in Tobacco Use. http://www.lung.org/finding-cures/our-research/trend-reports/Tobacco-Trend-Report.pdf.

Arellano, M. and Bond, S. (1991). Some Tests of Specification for Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *Review of Economic Studies*, **58**, 277–297.

Balia, S. and Jones, A. (2011). Catching the Habit: A Study of Inequality of Opportunity in Smoking-related Mortality. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **174**, 175–194.

Becker, G. S. and Murphy, K. (1988). A Theory of Rational Addiction. *Journal of Political Economy*, **96** (4), 675–700.

Bedard, K. and Deschenes, O. (2006). The Long-Term Impact of Military Service on Health: Evidence from World War II and Korean War Veterans. *American Economic Review*, **96**, 176–194.

Bernheim, B. D. and Rangel, A. (2004). Addiction and Cue-Triggered Decision Processes. *American Economic Review*, **94** (5), 1558–1590.

Bhargava, A. and Sargan, J. (1983). Estimating Dynamic Random Effects Models from Panel Data Covering Short Time Periods. *Econometrica*, **51** (6), 1635–1659.

Black, S. E., Devereux, P. J. and Salvanes, K. G. (2015). Healthy(?), Wealthy and Wise: Birth Order and Adult Health, NBER Working Paper 21337.

Bureau of the Census (1949). *Historical Statistics of the United States, 1789-1945: a supplement to the Statistical Abstract of the United States*. Tech. rep., Washington DC: US Department of Commerce.

Carter, B. D., Abnet, C. C., Feskanich, D., Freedman, N. D., Hartge, P., Lewis, C. E., Ockene, J. K., Prentice, R. L., Speizer, F. E., Thun, M. J. and Jacobs, E. J. (2015). Smoking and Mortality: Beyond Established Causes. *New England Journal of Medicine*, **372** (7), 631–640.

CBO (2012). *Raising the Excise Tax on Cigarettes: Effects on Health and the Federal Budget*. Tech. rep., Congressional Budget Office of the United States of America.

Chaloupka, F., Warner, K., Acemoglou, D., Gruber, J., Laux, F., Max, W., Newhouse, J., Schelling, T. and Sindelar, J. (2014). *An Evaluation of FDAs Analysis of the Costs and Benefits of the Graphic Warning Label Regulation*. Tech. rep., National Bureau of Economic Research.

Clark, A. and Etil, F. (2002). Do Health Changes Affect Smoking? Evidence from British Panel Data. *Journal of Health Economics*, **21** (4), 533–562.

Cunha, F. and Heckman, J. (2008). Formulating, identifying and estimating the technology of cognitive and noncognitive skill formation. *Journal of Human Resources*, **43** (4), 738–782.

Darden, M. (forthcoming). Smoking, Expectations, and Health: A Dynamic Stochastic Model of Lifetime Smoking Behavior. *Journal of Political Economy*.

— and Gilleskie, D. B. (2016). The Effects of Parental Health Shocks on Adult Offspring Smoking Behavior: Evidence from a Long Panel. *Health Economics*, **25**, 939–954.

Doll, R., Peto, R., Boreham, J., Gray, R. and Sutherland, I. (2004). Mortality in Relation to Smoking: 50 Years' Observations on Male British Doctors. *British Medical Journal*, **328**, 1519–1528.

—, —, Wheatley, K., Gray, R. and Sutherland, I. (1994). Mortality in Relation to Smoking: 40 Years' Observations on Male British Doctors. *British Medical Journal*, **309**, 901–911.

Evans, W. N. and Ringel, J. S. (1999). Can Higher Cigarette Taxes Improve Birth Outcomes? *Journal of Public Economics*, **72** (1), 135–154.

Galama, T. (2011). A Contribution to Health Capital Theory. *RAND Working Paper 831*.

Gilleskie, D. B. and Strumpf, K. S. (2005). The Behavioral Dynamics of Youth Smoking. *Journal of Human Resources*, **40** (4), 822–866.

Grossman, M. (1972). On the Concept of Health Capital and the Demand for Health. *Journal of Political Economy*, **80** (2), 223–255.

—, Sindelar, J. L., Mullahy, J. and Anderson, R. (1993). Policy watch: Alcohol and cigarette taxes. *Journal of Economic Perspectives*, **7** (4), 211–222.

Gruber, J. and Koszegi, B. (2001). Is Addiction "Rational"? Theory and Evidence. *The Quarterly Journal of Economics*, **116** (4), 1261–1303.

Guilkey, D. and Lance, P. (2014). Program Impact Estimation with Binary Outcome Variables: Monte Carlo Results for Alternative Estimators and Empirical Examples. In R. C. Sickles and W. C. Horrace (eds.), *Festschrift in Honor of Peter Schmidt*, Springer New York, pp. 5–46.

Harris, J. (1979). *Smoking and Health: A Report of the Surgeon General: Appendix: Cigarette Smoking in the United States, 1950-1978*. Tech. rep., US Department of Health, Education, and Welfare, http://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/NNBCMD/p-segmented/true.

Heckman, J. J. and Singer, B. (1984). A Method for Minimizing the Impact of Distributional Assumptions in Econometric Models for Duration Data. *Econometrica*, **52** (2), 271–320.

Honore, B. and Lleras-Muney, A. (2006). Bounds in Competing Risks Models and the War on Cancer. *Econometrica*, **74**, 1675–1698.

Jha, P., Ramasundarahettige, C., Landsman, V., Rostron, B., Thun, M., Anderson, R., McAfee, T. and Peto, R. (2013). 21st-Century Hazards of Smoking and Benefits of Cessation in the United States. *New England Journal of Medicine*, **368**, 341–350.

Kelly, A. B., O'Flaherty, M., Connor, J. P., Homel, R., Toubourou, J. W., Patton, G. C. and Williams, J. (2011). The Influence of Parents, Siblings and Peers on Pre- and Early-teen Smoking: A Multilevel Model. *Drug and Alcohol Review*, **30** (4), 381–387.

Khwaja, A. (2010). Estimating Willingness to Pay for Medicare using a Dynamic Life-cycle Model of Demand for Health Insurance. *Journal of Econometrics*, **156** (1), 130–147.

Kohn, J. (2008). The Change in Health, Consumption and the Demand for Medical Care. *Working Paper*.

Mroz, T. (1999). Discrete Factor Approximations in Simultaneous Equation Models: Estimating the Impact of a Dummy Endogenous Variable on a Continuous Outcome. *Journal of Econometrics*, **92** (2), 233–274.

— and Guilkey, D. (1992). Discrete Factor Approximations for Use in Simultaneous Equation Models with Both Continuous and Discrete Endogenous Variables. *mimeo, Dept. of Economics, University of North Carolina, Chapel Hill*.

Naqvi, N., Rudrauf, D., Damasio, H. and Behara, A. (2007). Damage to the Insula Disrupts Addiction to Cigarette Smoking. *Science*, **315**, 531–534.

National Cancer Institute (1997). *Changes in Cigarette-Related Disease Risks and Their Implication for Prevention and Control. Smoking and Tobacco Control Monograph Number 8*. Tech. rep., National Cancer Institute. Department of Health and Human Services, Public Health Service, National Institutes of Health. Bethesda: U.S.

Peto, R., Darby, S., Deo, H., Silcocks, P., Whitley, E. and Doll, R. (2000). Smoking, Smoking Cessation, and Lung Cancer in the UK since 1950: Combination of National Statistics with Two Case-control Studies. *British Medical Journal*, **321**, 323–329.

Pirie, K., Peto, R., Reeves, G., Green, J. and for the Million Women Study Collaborators, V. B. (2013). The 21st Century Hazards of Smoking and Benefits of Stopping: A Prospective Study of One Million Women in the UK. *Lancet*, **381**, 133–141.

Pollay, R. W., Siddarth, S., Siegel, M., Haddix, A., Merritt, R. K., Giovino, G. A. and Eriksen, M. P. (1996). The last straw? cigarette advertising and realized market shares among youths and adults, 1979-1993. *Journal of Marketing*, **60** (2), 1–16.

Prince, M. J., Wu, F., Guo, Y., Robledo, L. M. G., O'Donnell, M., Sullivan, R. and Yusuf, S. (2014). The Burden of Disease in Older People and Implications for Health Policy and Practice. *The Lancet*, **385** (9967), 549–562.

Sloan, F. A., Smith, V. K. and Taylor, D. H. (2002). Information, Addiction, and 'Bad Choices': Lessons from a Century of Cigarettes. *Economic Letters*, **77**, 147–155.

Smith, V. K., Taylor, D. H., Sloan, F. A., Johnson, F. R. and Desvousges, W. H. (2001). Do Smokers Respond to Health Shocks? *The Review of Economics and Statistics*, **83** (4), 675–687.

Taylor, D. H., Hasselblad, V., Henley, S. J., Thun, M. J. and Sloan, F. A. (2002). Benefits of Smoking Cessation for Longevity. *American Journal of Public Health*, **92** (6), 990–996.

United States Surgeon General (1990). *The Health Benefits of Smoking Cessation: A Report of the Surgeon General's Report*. Tech. rep., United States Department of Health and Human Services, Public Health Service, Office on Smoking and Health.

United States Surgeon General (2004). *The Health Consequences of Smoking: A Report of the Surgeon General*. Tech. rep., United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Promotion, Office on Smoking and Health, Atlanta, GA.

United States Surgeon General (2014). *The Health Consequences of Smoking–50 Years of Progress: A Report of the Surgeon General*. Tech. rep., United States Department of Health and Human Services, Centers for Disease Control and Prevention, National Center for Chronic Disease Prevention and Promotion, Office on Smoking and Health, Atlanta, GA.

Viscusi, W. K. (1990). Do Smokers Underestimate Risks? *Journal of Political Economy*, **98** (6), 1253–1269.

— and Hakes, J. K. (2008). Risk Beliefs and Smoking Behavior. *Economic Inquiry*, **46** (1), 45–59.

Smoking and Mortality: New Evidence from a Long Panel Online Appendix*

Michael Darden[1], Donna B. Gilleskie[2], and Koleman Strumpf[3]

[1]*Department of Economics, Tulane University*
[2]*Department of Economics, University of North Carolina at Chapel Hill*
[3]*School of Business, University of Kansas*

March 2017

*The Framingham Heart Study (FHS) is conducted and supported by the NHLBI in collaboration with the FHS Study Investigators. This manuscript was prepared using a limited access dataset obtained from the NHLBI and does not necessarily reflect the opinions or views of the FHS or the NHLBI. We appreciate the comments of Robert Kaestner, Don Kenkel, Edward Norton, Tom Mroz, and seminar participants and discussants at Georgia State, Indiana, Lund, Notre Dame, SUNY-Stony Brook, Tulane, York, the Health Econometrics Symposium at Leeds University, the 5th (Spain) and 10th (Dublin) iHEA World Congress, the Econometric Society Summer Meetings, the Southeastern Health Economics Study Group, and the Triangle Health Economics Workshop. Financial support from the National Institutes for Health (grant #1RO1HD42256-01) is gratefully acknowledged. Correspondence: mdarden1@tulane.edu; donna_gilleskie@unc.edu; cigar@ku.edu.

A Sample Construction

If smoking behavior is missing for any exam(s) it becomes difficult to construct variables that accurately reflect the history of one's behavior (e.g., years of cessation, duration, and experience). Missing smoking observations require that we either impute behavior or drop individuals from our research sample. Table A1 details our progress with assigning smoking behavior during missing exams by using observed individual responses from other exams. As our sequentially-applied assumptions introduce more potential noise, the resulting sample size increases. Ultimately, we select a sample of men (sample E) that minimizes this type of imputation while maximizing the number of individuals we can follow. Additional individuals are dropped from the sample if they are missing other important variables.

Here we address two potential concerns with our sample selection: (i) omitting individuals with particular patterns of missing smoking data may lead to an unrepresentative sample; (ii) filling in missing observations should be done in a limited fashion so as to avoid adding data artifacts and to not contribute to item (i). Our detailed consideration of these concerns fall into three general categories:

- individuals with filled-in data are not observationally different
- most individuals have only a few missing observations
- some omissions that one might be likely to assume possible are not common here

In addition it is worth recalling that these issues are endemic to the empirical smoking literature where high quality panel data is quite rare. Consider, for example, the NELS:88 data set (https://nces.ed.gov/surveys/nels88/), which has been used to study teen smoking behavior. In the first three waves of NELS:88 (1988, 1990, 1992) even if we limit the sample to kids who were on grade or permanent drop outs, 17% (2237 of 12954 total kids) have missing smoking data. Our study involves a far longer period (eight times as many periods) and a much older period, so the rather complete smoking questions is actually a very significant strength of our analysis.

The first step is to consider the role of sample selection. Table A2 reports Pearson χ^2-tests of whether individuals in our final sample (Sample E, those with no missing data or some filled-in data) differ from those from the FHS that are omitted from our analysis. We consider a variety of covariates used in the empirical analysis. For none of the variables can we reject the null that the individuals in the two groups are identical.

The next step is to examine the patterns of missing smoking data. In both the full population and the sample we examine (Sample E) there are few cases with large

Table A1: Research Sample Selection based on Assignment Method for Missing Smoking Values

Assumptions about smoking behaviors used to determine research sample	Lifetime smoking observed (% of individuals)	Smoking values missing (% of person-years)	Resulting sample size	Research sample name
Begin with original sample of males who have at least two consecutive exams (N=2274). Use reported per-exam smoking info	27.4	22.5	622	A
+ fill in missing per-exam smoking values with data from subsequent retrospective questions	32.5	19.9	739	B
+ fill in up to 4 consecutive missing values when same smoking choice observed before & after missing spell	55.1	13.2	1253	C
+ fill in all missing values when at least 50% of smoking choices are observed and that smoking behavior is the same	55.6	12.9	1265	D
+ fill in missing values before death when at least 3 consecutive exams of the same smoking behavior is observed before death	77.1	8.1	1754	E
+ fill in missing values before death with last observed smoking choice before death	85.1	4.2	1935	F

Table A2: Pearson χ^2-test: Comparisons of Sample E (N=1754) vs Individuals omitted from sample (N=520)

Covariate	χ^2	df	p
Education	3.24	5	0.663
Country of birth	1.13	1	0.288
Italian Ancestry	2.35	1	0.126
Age (by decade)	6.90	4	0.142

Age is based on age at first exam and is grouped into decade intervals (20s, 30s, 40s, 50s, 60s, 70s) to ensure adequate sample size in each cell.

numbers of missing data, and it is uncommon to have missing data interspersed between periods where data are present. This means the patterns of missing data that one might think of are actually quite rare. It is not possible to literally consider every such pattern, but the most challenging cases involve either large number of missing data or particular streaks of missing data, and we address both of these below.

Table A3 presents some descriptive statistics on missing data for the full population and our Sample E. In each case we present mean, medians and (to show what extreme values look like) the ninetieth percentiles of the distribution. The first two rows list the number of missing and imputed observations. There are typically only a few missing observations per person. And among those in the population but not in Sample E (those not in our empirical analysis, N=520), the median is four missing observations. The average person has about two missing observations imputed (using the rules for Sample E).

The remaining rows examine streaks of missing or available data (i.e., adjacent periods with no/available smoking information in the original raw data). We find most streaks of available data are relatively long, of missing data are relatively short, and most people only have a few such streaks. Together these indicate that anomalous patterns one might think occur cannot be common, since many involve a deviation from at least one of these results. The third row shows that for most people the longest streak of missing data is about two. Even among the population not in Sample E, the median maximum missing streak is two periods. The next two rows show that most people only have a few streaks of missing and available data (the median person has one streak of missing data and two streaks of available data in both the population and sample (and, even at the extreme ninetieth percent, a person would have seven total streaks).

The last two rows report the length of streaks of available data. In the population the average person's shortest streak is about eight periods and the longest is ten. The first point is of particular interest since most of the patterns that would require unusual omissions of an individual or excessive imputations would imply that there would be very short streaks of available data. Note that while the median values for the minimum streak are low, this is largely due to individuals who die early. (Calculated as a median of the number of observed periods, the median minimum streak is 0.18 for the population and 0.25 for Sample E.)

Table A3: Patterns of Missing Smoking Data (24 periods)

Statistic	Mean	Median	90% Percentile	Std Dev.
Full Population (N=2274)				
Number of Missing Observations	3.246	2	9	3.95
Number of Imputed Observations	2.040	1	5	2.69
Maximum Streak of Missing Observations	2.284	1	6	3.28
Number of Streaks of Missing Observations	1.501	1	3	1.28
Number of Streaks of Available Data	2.068	2	4	2.07
Maximum Streak of Available Data	10.440	9	22	7.97
Minimum Streak of Available Data	7.661	2	22	8.68
Sample E (N=1754)				
Number of Missing Observations	2.303	2	5	2.85
Number of Imputed Observations	2.251	2	5	2.83
Maximum Streak of Missing Observations	1.509	1	3	2.04
Number of Streaks of Missing Observations	1.327	1	3	1.26
Number of Streaks of Available Data	1.997	2	3	1.08
Maximum Streak of Available Data	12.068	13	24	7.87
Minimum Streak of Available Data	9.239	2	24	9.01

One last issue is: why is there missing data?. This could be due to either missed exams or individuals not answering the smoking question conditional on being at the exam. The table below shows that in almost all of the cases it is due to a missed exam, so no data at all is available for that person-exam. The main exceptions are exams 1 and 4 (where there are relatively few cases of missing data) and exam 11.

Table A4: Proportion of Missing Smoking Observations Due to Missed Exam

Exam Number	Proportion
1	0.240
2	1.000
3	1.000
4	0.703
5	0.995
6	1.000
7	0.950
8	1.000
9	0.991
10	0.984
11	0.000
12	0.981
13	0.995
14	0.991
15	0.994
16	1.000

First Sixteen Exams only (Kannel *et al.*, 1988). The other explanation for the missing data is not answering the smoking question at the exam. Note that this table is based on the raw exam number from the FHS rather than the timing convention used in the empirical analysis (defined in Section 3.2).

B Additional Estimation Results

Table B1: Summary of Specifications for Jointly Estimated Behaviors and Outcomes via FIML

Behavior/ Outcome	Explanatory Variables: Pre-determined/ Endogenous	Exogenous	Unobserved Heterogeneity	Likelihood Contribution
Smoking behavior				
Smoke at t	H_t^S, H_t^D	X_t, P_t	$\mu^S, \nu_t^S, \epsilon_t^S$	$p(s_t = s), s = 0, 1$
Morbidity outcomes				
CVD events at t	H_t^S, H_t^D	X_t	$\mu^{D^1}, \nu_t^{D^1}, \epsilon_t^{D^1}$	$p(d_t^1 = d^1), d^1 = 0, 1, 2$
Cancer diagnosis at t | no cancer up to t	H_t^S, H_t^D	X_t	$\mu^{D^2}, \nu_t^{D^2}, \epsilon_t^{D^2}$	$p(d_t^2 = d^2), d^2 = 0, 1$
Diabetes diagnosis at t | no diabetes up to t	H_t^S, H_t^D	X_t	$\mu^{D^3}, \nu_t^{D^3}, \epsilon_t^{D^3}$	$p(d_t^3 = d^3), d^3 = 0, 1$
Body Mass Index at t	H_t^S, H_t^D	X_t	$\mu^{D^4}, \nu_t^{D^4}, \epsilon_t^{D^4}$	$\phi(d_t^4)$
Mortality outcomes				
Death hazard by end of t	H_{t+1}^S, H_{t+1}^D	X_t	$\mu^M, \nu_t^M, \epsilon_t^M$	$p(m_{t+1} = m), m = 0, 1$
Cause of death in t | death in t	H_{t+1}^S, H_{t+1}^D	X_t	$\mu^{M^C}, \nu_t^{M^C}, \epsilon_t^{M^C}$	$p(m_{t+1}^C = c), c = 0, 1, 2$
Initial conditions				
Never smoked up to $t = 2$		X_1, P_1, Z_1	$\mu^{I^{S^1}}, \epsilon_1^{I^{S^1}}$	$p(E_2 = e), e = 0, 1)$
Current smoker in $t = 1$ | ever smoked		X_1, P_1, Z_1	$\mu^{I^{S^2}}, \epsilon_1^{I^{S^2}}$	$p(s_1 = s), s = 0, 1)$
Years of smoking up to $t = 2$ | smoker in $t = 1$		X_1, P_1, Z_1	$\mu^{I^{S^3}}, \epsilon_1^{I^{S^3}}$	$\phi^{S^3}(D_2)$
Any CVD up to $t = 2$		X_1, P_1, Z_1	$\mu^{I^{D^1}}, \epsilon_1^{I^{D^1}}$	$p(\text{CVD}_2 = e), e = 0, 1)$
Body Mass Index in $t = 1$		X_1, P_1, Z_1	$\mu^{I^{D^2}}, \epsilon_1^{I^{D^2}}$	$\phi^{D^2}(\text{BMI}_1)$

$p(\cdot)$ is a logit or multinomial logit probability of a dichotomous or polychotomous variable and $\phi(\cdot)$ is the normal density of a continuous variable.

Table B2: Estimation Results: Unobserved Heterogeneity Distributions

Mass point	Smoking		Mortality		Cause of death: CVD		Cause of death: Cancer		CVD=1		CVD>1		Cancer diagnosis	Diabetes diagnosis		BMI		Mass point probability
Time-invariant unobserved heterogeneity																		
μ_2	-2.799	***	-0.588		1.031	*	0.612		0.045		0.562		0.505	-2.024	**	-0.040	***	0.297
	(0.357)		(0.375)		(0.543)		(0.623)		(0.255)		(0.475)		(0.475)	(0.861)		(0.008)		
μ_3	-0.521		-0.179		1.037		0.132		-0.144		0.511		0.126	0.020		0.010		0.032
	(0.454)		(0.476)		(0.837)		(0.998)		(0.411)		(0.637)		(0.638)	(0.626)		(0.018)		
μ_4	-2.205	***	0.089		1.312	**	-0.304		0.252		0.846		0.332	0.201		0.072	***	0.147
	(0.641)		(0.354)		(0.588)		(0.703)		(0.24)		(0.534)		(0.542)	(0.511)		(0.009)		
μ_5	-2.095	***	0.010		-0.020		-1.180		0.861	**	0.386		0.127	0.707		0.171	***	0.034
	(0.691)		(0.435)		(0.732)		(0.846)		(0.354)		(0.961)		(0.767)	(0.67)		(0.018)		
μ_6	-3.061	***	-0.198		1.384	**	0.508		0.168		0.513		0.521	-0.755	*	0.007		0.353
	(0.353)		(0.476)		(0.596)		(0.713)		(0.213)		(0.526)		(0.495)	(0.387)		(0.007)		
Time-varying unobserved heterogeneity																		
ν_{t2}	-0.566		-0.531		-0.444		-0.126		-0.896		-3.761	***	-0.505	-0.760		0.976	***	0.023
	(1.674)		(0.677)		(1.257)		(1.431)		(0.654)		(1.195)		(1.288)	(1.134)		(0.084)		
ν_{t3}	4.508	**	-0.099		1.356		-0.109		-1.606	**	-3.353	***	-0.953	-2.800		0.689	***	0.548
	(2.05)		(0.738)		(0.845)		(1.126)		(0.644)		(0.626)		(1.19)	(2.863)		(0.06)		
ν_{t4}	0.291		-1.572	*	1.822		1.512		-0.704		-2.707	***	0.037	-0.143		0.703	***	0.391
	(1.692)		(0.868)		(1.201)		(1.293)		(0.58)		(0.667)		(1.13)	(1.011)		(0.062)		
ν_{t5}	3.578		-0.310		1.338	*	0.429		0.380		-0.403		0.706	1.040		0.414	***	0.034
	(2.196)		(0.699)		(0.811)		(1.171)		(0.606)		(0.555)		(1.169)	(0.998)		(0.044)		
ν_{t6}	0.044		0.280		0.329		-1.862		-0.945		-2.339	***	-1.461	-2.616	**	1.586	***	0.001
	(2.292)		(1.051)		(1.611)		(1.926)		(1.4)		(0.622)		(1.166)	(1.128)		(0.171)		

Note: First mass point vector values normalized to zero for both distrubutions, and occur with probabilities 0.137 and 0.003.
Standard errors are in parentheses. *** indicates joint significance at the 1% level; ** 5% level; * 10% level.

Table B3: Selected Parameter Estimates: Cause of Death conditional on Death by end of period t

Variable	Cardiovascular disease						Cancer					
	Single Equation Without Correlated UH			FIML With Correlated UH			Single Equation Without Correlated UH			FIML With Correlated UH		
	Estimate	Std. error		Estimate	Std. error		Estimate	Std. error		Estimate	Std. error	
Smoker in t, $\mathtt{s}_t$	0.537	0.317	*	0.974	0.747		0.381	0.337		1.306	0.588	**
Years of cessation, $\mathtt{C}_t$	0.002	0.007		0.008	0.008		0.012	0.008		0.018	0.009	*
Years of duration, $\mathtt{D}_t$	-0.006	0.007		-0.005	0.014		0.000	0.007		-0.011	0.011	
Years of experience, $\mathtt{E}_t$	-0.003	0.005		-0.005	0.005		0.001	0.005		-0.002	0.006	
$\mathtt{1[CVD}_t = 1\mathtt{]}$	0.967	0.254	***	1.067	0.371	***	-0.188	0.333		-0.340	0.408	
$\mathtt{1[CVD}_t > 1\mathtt{]}$	1.428	0.463	***	1.373	0.646	**	0.094	0.639		-0.083	0.787	
$\mathtt{CAN}_t$	1.664	0.652	**	1.643	0.647	**	3.156	0.626	***	3.023	0.676	***
$\mathtt{DIA}_t$	-0.797	0.659		-0.851	0.985		0.355	0.687		0.148	0.970	
$\mathtt{1[CVD}_{t-1} = 1\mathtt{]}$	0.014	0.284		-0.016	0.313		-0.489	0.388		-0.557	0.400	
$\mathtt{1[CVD}_{t-1} > 1\mathtt{]}$	0.024	0.434		0.047	0.475		-1.014	0.735		-1.017	0.778	
$\mathtt{CAN}_{t-1}$	0.498	0.653		0.460	0.727		1.269	0.563	**	1.299	0.667	*
$\mathtt{DIA}_{t-1}$	-0.332	0.631		-0.217	0.570		0.679	0.684		0.874	0.669	
$\mathtt{E\_CVD}_{t-1}$	0.460	0.225	**	0.607	0.236	***	-0.137	0.299		-0.107	0.358	
$\mathtt{N\_CVD}_{t-1}$	0.122	0.095		0.100	0.100		-0.099	0.146		-0.098	0.181	
$\mathtt{E\_CAN}_{t-1}$	-0.086	0.286		-0.149	0.310		1.639	0.257	***	1.661	0.294	***
$\mathtt{E\_DIA}_{t-1}$	0.256	0.230		0.397	0.267		-0.180	0.300		0.004	0.353	
$\mathtt{BMI}_t$	0.047	0.024	**	0.060	0.041		0.006	0.026		0.075	0.049	
$\mathtt{SBP}_t$	0.005	0.004		0.005	0.005		0.002	0.005		0.002	0.006	
$\mathtt{DBP}_t$	-0.003	0.008		-0.002	0.009		-0.009	0.010		-0.009	0.010	
$\mathtt{CHO}_t$	0.010	0.003	***	0.011	0.003	***	0.002	0.003		0.002	0.003	
BMI, SBP, DBP missing	1.399	0.766	*	1.704	1.147		-0.578	0.877		1.072	1.380	
CHO missing	1.677	0.581	***	1.849	0.594	***	0.218	0.650		0.174	0.668	
$\mathtt{ART}_t$	-0.334	0.153	**	-0.348	0.161	**	-0.393	0.174	**	-0.404	0.190	**
Constant	-8.689	2.947	***	-12.363	2.463	***	-9.692	3.530	***	-14.383	1.919	***

Note: Specifications also include controls for age, education, ancestry, origin, cohort, and year trends.
Standard errors are in parentheses. *** indicates joint significance at the 1% level; ** 5% level; * 10% level.

Table B4: Selected Parameter Estimates: Cardiovascular Disease Event in period t

Variable	CVD=1 (relative to none)						CVD>1 (relative to none)					
	Single Equation Without Correlated UH			FIML With Correlated UH			Single Equation Without Correlated UH			FIML With Correlated UH		
	Estimate	Std. error		Estimate	Std. error		Estimate	Std. error		Estimate	Std. error	
Smoker in $t-1$, $\mathtt{s}_{t-1}$	0.184	0.251		0.279	0.275		0.007	0.535		0.227	0.603	
Years of cessation, $\mathtt{C}_t$	-0.004	0.011		-0.002	0.013		0.015	0.024		0.027	0.026	
$\mathtt{C}_t^2/100$	0.018	0.028		0.013	0.030		-0.059	0.065		-0.088	0.066	
Years of duration, $\mathtt{D}_t$	0.035	0.017	**	0.032	0.017	*	0.081	0.035	**	0.081	0.036	**
$\mathtt{D}_t^2/100$	-0.067	0.030	**	-0.064	0.030	**	-0.139	0.058	**	-0.145	0.057	**
Years of experience, $\mathtt{E}_t$	0.002	0.010		0.005	0.012		-0.026	0.018		-0.036	0.019	*
$\mathtt{E}_t^2/100$	-0.001	0.020		-0.005	0.024		0.048	0.037		0.065	0.037	*
$\mathtt{1[CVD}_{t-1}=1\mathtt{]}$	0.441	0.128	***	0.474	0.149	***	0.540	0.231	**	0.664	0.251	***
$\mathtt{1[CVD}_{t-1}>1\mathtt{]}$	0.694	0.205	***	0.744	0.224	***	0.857	0.308	***	0.952	0.319	***
$\mathtt{CAN}_{t-1}$	-0.173	0.322		-0.164	0.358		-0.144	0.664		-0.044	0.662	
$\mathtt{DIA}_{t-1}$	0.018	0.274		-0.076	0.284		-1.189	0.748		-1.432	0.727	**
$\mathtt{E\_CVD}_{t-1}$	0.529	0.130	***	0.499	0.143	***	0.164	0.236		0.114	0.250	
$\mathtt{N\_CVD}_{t-1}$	0.174	0.049	***	0.177	0.059	***	0.345	0.068	***	0.383	0.065	***
$\mathtt{E\_CAN}_{t-1}$	0.349	0.179	*	0.323	0.188	*	0.161	0.360		0.076	0.376	
$\mathtt{E\_DIA}_{t-1}$	0.376	0.139	***	0.324	0.166	*	0.869	0.223	***	0.881	0.247	***
$\mathtt{E\_CVD}_{t-1}$ * $\mathtt{S}_{t-1}$	-0.269	0.164		-0.050	0.348		-0.265	0.303		-0.082	0.664	
$\mathtt{E\_CAN}_{t-1}$ * $\mathtt{S}_{t-1}$	-0.025	0.347		0.817	0.281	***	-0.004	0.694		-0.629	0.792	
$\mathtt{E\_DIA}_{t-1}$ * $\mathtt{S}_{t-1}$	0.755	0.258	***	-0.269	0.185		-0.734	0.651		-0.321	0.325	
$\mathtt{BMI}_{t-1}$	-0.064	0.078		0.026	0.108		0.247	0.201		0.379	0.112	***
$\mathtt{BMI}_{t-1}^2/100$	0.170	0.135		-0.077	0.181		-0.361	0.358		-0.711	0.247	***
$\mathtt{SBP}_{t-1}$	0.009	0.014		0.007	0.015		0.030	0.024		0.027	0.024	
$\mathtt{SBP}_{t-1}^2/100$	0.000	0.004		0.001	0.005		-0.005	0.008		-0.004	0.008	
$\mathtt{DBP}_{t-1}$	-0.036	0.024		-0.034	0.025		-0.042	0.042		-0.040	0.043	
$\mathtt{DBP}_{t-1}^2/100$	0.022	0.014		0.021	0.015		0.031	0.024		0.031	0.025	
$\mathtt{CHO}_{t-1}$	0.002	0.006		0.005	0.005		-0.005	0.007		0.001	0.006	
$\mathtt{CHO}_{t-1}^2/100$	0.000	0.001		0.000	0.001		0.002	0.001		0.001	0.001	
BMI, SBP, DBP missing	-0.997	1.260		-0.342	1.699		4.386	2.913		5.352	1.471	***
CHO missing	0.721	0.670		1.105	0.623	*	-0.208	0.940		0.472	0.838	
$\mathtt{ART}_{t-1}$	0.124	0.078		0.108	0.081		-0.226	0.151		-0.246	0.166	
Constant	-11.303	1.876	***	-12.022	2.261	***	-17.530	3.843	***	-17.436	1.311	***

Note: Specifications also include controls for age, education, ancestry, origin, cohort, and year trends.
Standard errors are in parentheses. *** indicates joint significance at the 1% level; ** 5% level; * 10% level.

Table B5: Selected Parameter Estimates: Cancer Diagnosis and Diabetes Diagnosis in period t

Variable	Cancer diagnosis | No previous diagnosis						Diabetes diagnosis | No previous diagnosis					
	Single Equation Without Correlated UH			FIML With Correlated UH			Single Equation Without Correlated UH			FIML With Correlated UH		
	Estimate	Std. error		Estimate	Std. error		Estimate	Std. error		Estimate	Std. error	
Smoker in $t-1$, $\mathtt{s}_{t-1}$	-0.001	0.004		0.387	0.433		0.543	0.333		0.197	0.362	
Years of cessation, $\mathtt{C}_t$	0.000	0.000		0.009	0.006		-0.007	0.008		-0.012	0.008	
Years of duration, $\mathtt{D}_t$	0.000	0.000		0.004	0.009		-0.021	0.009	**	-0.024	0.010	**
Years of experience, $\mathtt{E}_t$	0.000	0.000		0.001	0.004		-0.005	0.005		0.003	0.006	
$\mathtt{E\_CVD}_{t-1}$	0.003	0.004		0.197	0.247		0.063	0.259		-0.110	0.277	
$\mathtt{N\_CVD}_{t-1}$	-0.002	0.002		-0.122	0.125		0.341	0.111	***	0.355	0.115	***
$\mathtt{E\_CAN}_{t-1}$							-0.991	0.519	*	-1.132	0.522	**
$\mathtt{E\_DIA}_{t-1}$	-0.005	0.004		-0.218	0.246							
$\mathtt{BMI}_{t-1}$	0.000	0.000		-0.004	0.035		0.143	0.018	***	-0.018	0.060	
$\mathtt{SBP}_{t-1}$	0.000	0.000		0.001	0.004		0.009	0.004	**	0.009	0.004	**
$\mathtt{DBP}_{t-1}$	0.000	0.000		-0.006	0.008		-0.012	0.008		-0.015	0.008	*
$\mathtt{CHO}_{t-1}$	0.000	0.000	***	-0.006	0.002	***	-0.002	0.002		-0.002	0.002	
BMI, SBP, DBP missing	-0.015	0.009		-0.787	1.075		2.179	0.973	**	-2.255	1.826	
CHO missing	-0.005	0.006		-0.992	0.472	**	-0.028	0.490		-0.078	0.532	
$\mathtt{ART}_{t-1}$	0.001	0.002		0.083	0.134		0.043	0.155		0.012	0.163	
Constant	1.014	0.021	***	-20.791	3.585	***	-22.242	2.882	***	-17.731	3.854	***

Note: Specifications also include controls for age, education, ancestry, origin, cohort, and year trends.
Standard errors are in parentheses. ** indicates joint significance at the 5% level; * 10% level.

Table B6: Selected Parameter Estimates: Body Mass Index in period t

Variable	Single Equation Without Correlated UH			FIML With Correlated UH		
Variable	Estimate	Std. error		Estimate	Std. error	
Smoker in $t-1$, $\mathtt{s}_{t-1}$	0.005	0.005		0.001	0.005	
Years of cessation, $\mathtt{C}_t$	0.000	0.000		-0.001	0.000	**
$\mathtt{C}_t^2/100$	0.001	0.000	*	0.001	0.001	**
Years of duration, $\mathtt{D}_t$	-0.001	0.000	***	-0.001	0.000	***
$\mathtt{D}_t^2/100$	0.002	0.001	**	0.002	0.001	***
Years of experience, $\mathtt{E}_t$	0.000	0.000	*	0.001	0.000	***
$\mathtt{E}_t^2/100$	-0.001	0.001	**	-0.002	0.001	***
$\mathtt{1[CVD}_{t-1}=1\mathtt{]}$	0.004	0.005		-0.001	0.005	
$\mathtt{1[CVD}_{t-1}>1\mathtt{]}$	0.017	0.009	*	0.014	0.009	
$\mathtt{CAN}_{t-1}$	0.020	0.011	*	0.007	0.011	
$\mathtt{DIA}_{t-1}$	-0.018	0.010	*	0.004	0.012	
$\mathtt{E\_CVD}_{t-1}$	0.003	0.005		0.005	0.006	
$\mathtt{N\_CVD}_{t-1}$	-0.003	0.002		-0.006	0.004	
$\mathtt{E\_CAN}_{t-1}$	-0.009	0.007		-0.001	0.007	
$\mathtt{E\_DIA}_{t-1}$	0.000	0.005		-0.021	0.007	***
$\mathtt{E\_CVD}_{t-1}$ * $\mathtt{S}_{t-1}$	-0.003	0.006		-0.019	0.016	
$\mathtt{E\_CAN}_{t-1}$ * $\mathtt{S}_{t-1}$	-0.029	0.013	**	-0.006	0.013	
$\mathtt{E\_DIA}_{t-1}$ * $\mathtt{S}_{t-1}$	-0.010	0.011		0.002	0.006	
$\mathtt{BMI}_{t-1}$	0.102	0.002	***	0.110	0.011	***
$\mathtt{BMI}_{t-1}^2/100$	-0.016	0.004	***	-0.044	0.022	**
$\mathtt{SBP}_{t-1}$	-0.001	0.000		-0.001	0.000	
$\mathtt{SBP}_{t-1}^2/100$	0.000	0.000		0.000	0.000	
$\mathtt{DBP}_{t-1}$	0.002	0.001	**	0.002	0.001	**
$\mathtt{DBP}_{t-1}^2/100$	-0.001	0.000	**	-0.001	0.000	**
$\mathtt{CHO}_{t-1}$	0.000	0.000		0.000	0.000	
$\mathtt{CHO}_{t-1}^2/100$	0.000	0.000		0.000	0.000	
BMI, SBP, DBP missing	2.682	0.043	***	2.637	0.190	***
CHO missing	-0.014	0.019		0.000	0.020	
$\mathtt{ART}_{t-1}$	0.000	0.002		-0.001	0.002	
Constant	-0.003	0.048		-0.734	0.150	***

Note: Specifications also include controls for age, education, ancestry, origin, cohort, and year trends. Standard errors are in parentheses.
*** indicates joint significance at the 1% level; ** 5% level; * 10% level.

Table B7: Parameter Estimates: Initial Conditions (FIML with correlated UH)

	Never Smoked			Current Smoker			Smoking Duration			CVD			BMI		
Age (years)	0.067	0.023	***	-0.117	0.035	***	0.110	0.010	***	0.132	0.061	**	0.003	0.003	
Educ: grade school	-0.123	0.199		-0.534	0.299	*	0.058	0.054		-0.647	0.479		-0.019	0.031	
Educ: some high school	-0.524	0.222	**	0.412	0.356		0.027	0.046		-0.693	0.589		-0.028	0.030	
Educ: some college	-0.291	0.272		-1.143	0.395	***	-0.128	0.061	**	-0.656	0.683		-0.025	0.036	
Educ: college degree	0.371	0.248		-0.628	0.385		-0.179	0.057	***	-0.578	0.623		-0.012	0.029	
Educ: post college	0.635	0.249	**	-0.216	0.440		-0.280	0.078	***	0.288	0.574		0.048	0.032	
Born outside U.S.	0.040	0.197		-0.231	0.293		-0.072	0.062		-0.200	0.459		-0.007	0.026	
Italian ancestry	0.351	0.176	**	-0.219	0.259		-0.016	0.042		-0.510	0.560		0.142	0.025	***
Older cohort	-0.288	0.269		-0.144	0.392		-0.207	0.099	**	-0.016	0.627		-0.033	0.037	
Ad expenditure at age 10-14	0.177	0.111		-0.428	0.172	**				0.221	0.335		-0.009	0.011	
Ad expenditure at age 15-18							0.065	0.033	**						
Number of siblings	0.019	0.032		-0.001	0.046		0.006	0.009		-0.006	0.088		-0.007	0.003	**
Only child	-0.869	0.441	**	1.238	0.806		0.106	0.065		-0.464	1.092		-0.013	0.037	
Sibling information missing	-0.568	0.272	**	1.849	0.441	***	0.149	0.065	**	-0.323	0.769		0.008	0.036	
Birth order (up to 5th)	-0.008	0.058		0.024	0.091		-0.012	0.016		-0.383	0.212	*	0.014	0.009	*
First born child										-0.739	0.655		0.005	0.026	
Constant	-5.556	1.354	***	6.958	1.721	***	-2.288	0.490	***	-7.307	3.187	**	2.469	0.141	***
Time-invariant unobserved heterogeneity															
μ_2	0.719	0.812		2.042	0.458	***	-0.049	0.072		-1.495	0.607	**	-0.309	0.046	***
μ_3	1.823	0.844	**	0.553	0.619		0.023	0.109		0.795	0.743		0.722	0.062	***
μ_4	2.086	0.854	**	1.355	0.640	**	-0.066	0.075		-2.477	1.168	**	0.040	0.063	
μ_5	-1.060	5.472		2.757	0.674	***	-1.991	0.224	***	-2.666	1.069	**	0.025	0.110	
μ_6	1.307	0.892		0.611	0.440		-0.118	0.112		-2.275	1.720		0.322	0.056	***

Note: Standard errors are in parentheses. ** indicates joint significance at the 5% level; * 10% level.
We use age 10-14 exposure to explain initiation of smoking, and we use age 15-18 exposure to explain observed duration of smoking when we first observe an individual in our data. Individuals who smoke into adulthood, generally, "developed the habit" in adolescence rather than "just experimented". First born child is omitted for all initial smoking equations because there was not enough variation in the variable in the smaller sample sizes of the conditioned equations: current smoker|ever smoked and duration|current smoker.

C Historical Data

In this appendix we discuss the cigarette advertising and price data used to construct important cigarette market variables over the 19th and 20th centuries. For each variable (i.e., average advertising expenditure and average price), we first provide a justification of its use as an instrument for cigarette smoking and then discuss details associated with construction of the advertising expenditure and price time series. Some of the discussion focuses on the state of Massachusetts, since the FHS data are from the town of Framingham.

C.1 Advertising and Cigarette Consumption

We use industry-wide advertising spending to instrument for cigarette initiation during the years 1895-1939 and for smoking behavior over the years 1950-1996. There are two key conditions needed for identification: smoking initiation must be responsive to advertising and trends in advertising spending must be aimed at market expansion rather than brand switching. We deal with each of these issues in turn making reference to the literature.

The first condition is that firm advertising impacts smoking behavior. There are several channels by which this could occur. For example, during the pre-World War II period cigarette advertising increased social acceptability of smoking (particularly for women for whom it had been considered taboo), promoted the image of smokers as independent and glamorous, and listed health benefits such as hunger suppression (Brandt, 2007). There is empirical evidence linking advertising to youth smoking initiation (and almost all smokers in our data begin smoking by the teen years). In their survey of the economics of smoking, Chaloupka and Warner (2000) note that advertising has a positive and significant impact on teen smoking initiation in studies using individual-level data. Borden (1942), Tennant (1950), and Pierce and Gilpin (1977) note that cigarette advertising during our study period was primarily targeted to groups, such as female youths, which had not smoked previously, and that these groups experienced greater increases in smoking initiation rates at those times. Telser (1962) provides estimates which show that firm-level cigarette advertising increased overall smoking levels during 1925-1939. (Participants in the FHS original cohort were born between 1886 and 1918 and were in their teens between 1900 and 1932.)

The second condition deals with the intentions underlying the decision to advertise. Advertising can both increase demand (the focus here) and also lead to brand switching (which might not increase smoking initiation). The main threat to identification would be if the latter effect predominates or if it changes in importance over time. In the

period through 1912 this is not a major concern since cigarettes and all other forms of tobacco were sold by a monopolist, the American Tobacco Company, also referred to as the Tobacco Trust. Since there was limited variation of prices and market segmentation at this time, there would be little advertising related to brand-names. In the post-Trust period, the industry largely moved in lock-step. The main cigarette manufacturers were convicted in 1941 of violating the Sherman Act, both Section 1 (restraint of trade) and Section 2 (monopolization). For example, the wholesale prices of all leading brands were identical from 1928 to 1946 and virtually identical prior to that with manufacturers changing prices within days of one another. In such an environment of likely tacit collusion, an important feature of advertising was to increase smoking overall as much as to promote individual brands. Echoing the goals of smoking advertising in the last paragraph, George Washington Hill, president of American Tobacco, testified at the 1941 anti-trust trial: "The impetus of those great advertising campaigns not only built this for ourselves, but built the cigarette business as well ... You don't benefit yourself most, I mean, altogether ... you help the whole industry if you do a good job" (p. 137, Tennant (1950)). There were two periods of relatively strong competition: the period immediately following the dissolution of the Tobacco Trust and the 1930s with a short-lived rise of economy cigarettes. Counter to what would be expected under brand-switching, advertising moved erratically in the first period and decreased during the latter period (see Figure C1). Also Telser (1962) shows that advertising at the brand-level was market expanding and that brand-stealing effects are small in magnitude during the 1920s and 1930s.

C.2 Construction of Advertising Expenditures Time Series

Annual nominal advertising spending on cigarettes, exclusive of free goods (e.g., give-aways of cigarettes) and other non-traditional advertising, comes from a variety of sources. Spending for the years 1893-1913 are from United States Department of Commerce (1915), which lists advertising spending per cigarette and also total cigarette sales. These totals include the entire cigarette business of the American Tobacco Company (the Tobacco Trust), exclusive of exports and foreign manufacturing business as well as Turkish cigarettes. Spending for the years 1893-1910 and the spending by the Trust's successor companies for 1912-1913 are government assembled totals completed in the wake of the the Supreme Court's break-up of the Trust in 1911. (No data are available for 1911 and spending is interpolated for that year).

Advertising expenditure for the years 1914-1928 are based on Nicholls (1951). Nicholls lists R.J. Reynolds Tobacco Company's cost of advertising, exclusive of gratis goods.

Largely due to its Camel brand, over most of this period Reynolds was the leading cigarette producer and it annually sold between a third and almost half of all cigarettes. The aggregate spending on cigarettes is approximated by dividing this total by Reynolds' share of total cigarettes and multiplying this by the share of cigarettes among all tobacco products.

Expenditures for the years 1929-1949 are drawn directly from Fujii (1980). He uses a variety of primary and secondary sources to create an index of corporate cigarette advertising. Expenditures for the years 1950-1962 come from Schneider *et al.* (1981). They credit their series to a telephone interview with Television Bureau of Advertising, Inc. Both of these sources list real spending.

Advertising expenditure values from 1963 onwards are from the Federal Trade Commission (2013). Starting in this year the FTC began collecting information on cigarette spending across a variety of media including TV, radio, print and others. In all cases we net out totals related to price promotion, promotional allowances, and other specific channels which were added in later years.

We consider a variety of robustness checks to ensure that differences between these sources do not create artificial variation. Several of the series overlap and the patterns discussed below remain when we use values for the other series. These overlaps include United States Department of Commerce (1915) and Nicholls (1951) which both include data for 1913; Nicholls (1951) and Fujii (1980) which both include data for 1929-1949 (Nicholls' data are for Reynolds' total traceable advertising expenditures over 1939-1949); Fujii (1980) and Schneider *et al.* (1981) which both include data for 1950-1973; Schneider *et al.* (1981) and Federal Trade Commission (2013) which both include data for 1963-1978. A second check was to include additional company's advertising spending during 1914-1928. Nicholls (1951) includes data for American Tobacco for 1925-1928, and aggregate spending on cigarettes is not sharply changed when the same approach described earlier is used. Advertising costs for American for 1929-1939 and Ligett & Myers for 1935-1939 is also available and is used to compare Nicholls (1951) and Fujii (1980) in the first robustness check. Finally, as a robustness check we compare these assembled values to other data sources. Borden (1942) includes various measure of total advertising over 1929-1939 for Camels and for all brands that are comparable to the values in Nicholls (1951) and Fujii (1980). Tennant (1950) presents several series that are identical or follow a similar pattern as United States Department of Commerce (1915) and Nicholls (1951).

Additionally, our assembled time series cigarette advertising expenditure data are converted into per capita terms using the United States population figures from United States Census Bureau (2000), United States Census Bureau (2011), and United States Cen-

sus Bureau (2012). In all cases, terms are converted to year 2000 dollars using Bureau of Labor Statistics (2013) for 1913 onwards and Sahr (2013) for earlier years.

Figures C2a and C2b depict the resulting time series in levels and per capita terms, respectively. A few common features are present in both series. There is a run-up in advertising after the break-up of the Tobacco Trust (i.e., annual spending tripled within three years) as well as a reduction in advertising during each of the World Wars and The Depression. There was another steady increase in the post-war period (i.e., annual spending went up almost eight fold from 1945 to 1967), and then fell starting in 1967 with the FCC's ruling in that year that the fairness doctrine required anti-smoking ads on TV and radio and the 1971 ban in ads on those media. Advertising again climbed in the mid-1970s to the mid-1980s, after which it steadily declined.

C.3 Prices and Cigarette Consumption

Standard price theory suggests that own prices should impact cigarette consumption. The important distinction is that, as highlighted in the theoretical foundation section, smoking decisions are inherently dynamic: smoking impacts future health and future utility (via preferences that capture addiction). The rational addiction literature shows forward-looking agents alter their smoking behavior based on both current and expected future cigarette prices (Becker and Murphy (1988); Gruber and Koszegi (2001)). In the analysis below we focus on contemporaneous prices.

There is a large literature documenting economically and statistically significant effects of prices (Chaloupka and Warner, 2000). For youth the price elasticity is -0.5 to -1.5, reflecting the responsiveness of smoking initiation (which determines the initial conditions in our estimation framework). For adults the price elasticity is -0.2 to -0.5, reflecting responsiveness of quits, relapse, and conditional intensity (captured by the contemporaneous smoking equations in our estimation framework). These results come mainly from analyses of recent data. It is noted that our data involve an earlier period with different technologies (e.g., filtered cigarettes were not introduced until the 1950s) and social mores with regards to smoking.

For identification purposes, we argue that our price data are exogenous. There are two main threats to this argument. The first issue is that firms might set prices strategically in response to consumer demand, and such reverse causality will lead to bias. While tobacco companies have some market power, it is important to remember that other factors shape consumer prices. The additional factors are federal and state excise taxes on cigarettes, state sales taxes, and state-imposed price regulations. These factors change for reasons that are largely exogenous to cigarette demand: the introduction

Figure C1: Annual Real Aggregate Cigarette Advertising Expenditure

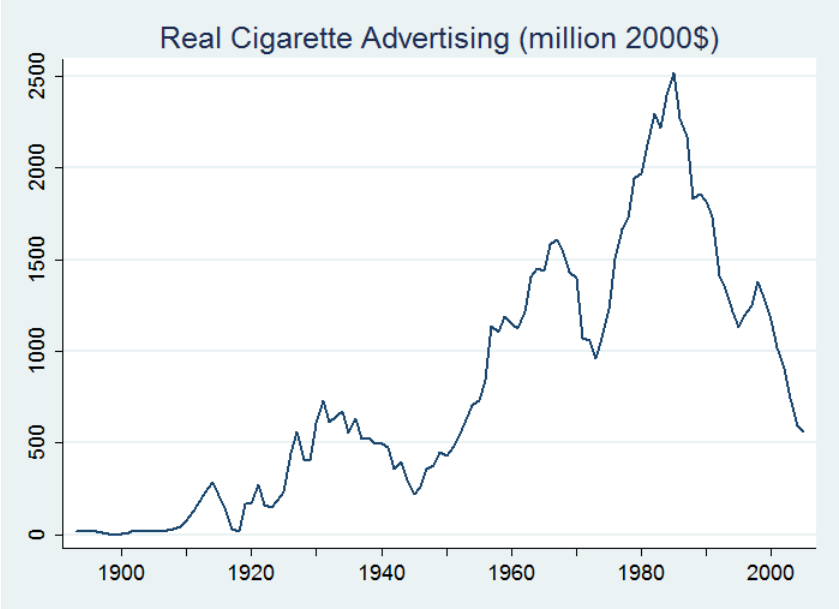


Figure C1a.

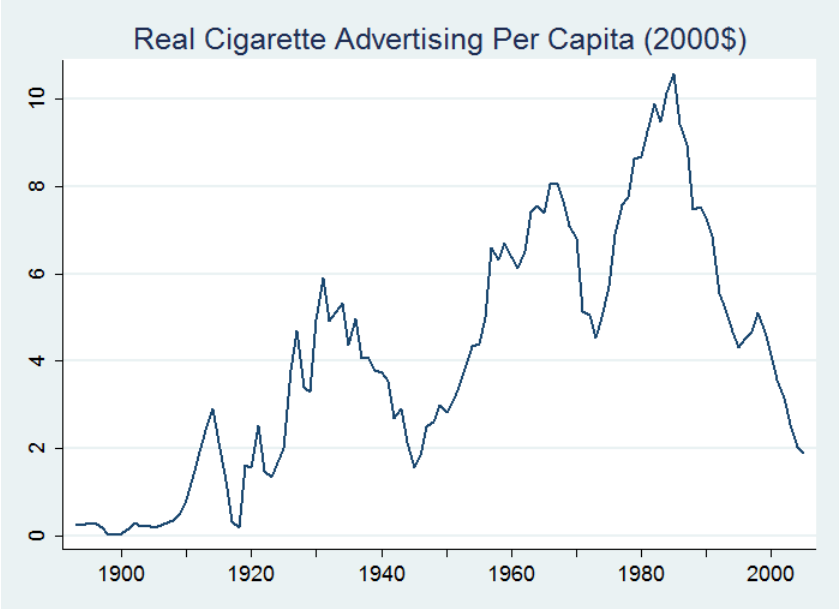


Figure C1b.

and subsequent increases in Massachusetts cigarette excise taxes through at least 1950 were instituted as emergency measures related to budget shortfalls; Massachusetts implemented a minimum cigarette price law in 1945 and over time continued to tinker with its formula (e.g., the mark-up rate, whether the state excise tax is included, differential treatment of less expensive brands, differential prices for non-chain stores). We show in the next subsection that taxes comprise on average half of the consumer price, and this share varies substantially over time. The minimum price rule makes it difficult for cigarette manufacturers to set final consumer prices; while the minimum price is based on wholesale prices, the specific formula continually changes (Annotated Laws of Massachusetts, 2007).

The second concern is that consumers buy cigarettes in other states that have lower prices. (There is far more price variation between- rather than within-states due to the role of state taxes and regulations.) Merriman (2010) shows that large tax differences lead to substantial cross-border shopping particularly over short distances. If this is true then observed prices at the state level would not reflect the true price that consumers face, and the extent of the mismeasurement would vary based on the size of the price differential. In the case of Framingham Massachusetts, the nearby states are Connecticut, New Hampshire, Rhode Island, and Vermont. For the years 1955-2011 Connecticut and Rhode Island have comparable prices as Massachusetts (Orzechowski and Walker, 2011), so cross-border shopping is not an issue. New Hampshire and Vermont both have lower prices over at least a portion of this period. Still, it is unlikely that cross-border shopping was a big issue over much of our sample, due to the relatively high cost of inter-state transportation until at least the 1950s and 1960s. There is also indirect evidence against cross-state traffic: the price differential grows over time, so if there is more inter-state purchases then sales between the states should become more lopsided over time. Per capita sales in New Hampshire and Vermont increased relative to Massachusetts when the price differential first started to become significant (i.e., the 1960s for New Hampshire and 1970s for Vermont). But, in the next decade as the price differential continued to grow, sales stopped shifting to the other states or even shifted back to Massachusetts.

C.4 Construction of Prices Time Series

This subsection discusses construction of the cigarette price time series for the period 1901-2011. The series is for Massachusetts, the smallest area for which we could collect prices. (We argue this is reasonable given the relatively small size of Massachusetts). In all cases prices are per one thousand cigarettes (an industry standard), include all

state and federal taxes, and are converted to year 2000 dollars using Bureau of Labor Statistics (2013) for 1913-2011 and Sahr (2013) for earlier years. We generate various summary statistics including the unweighted-average, minimum across brands, and these values exclusive of generics/economy brands. (Generic/economy brands were prominent for three periods: 1901-1910, 1931-1950 and 1991-2011.)

Prices through 1950 come from a variety of sources that list price at the brand-level and at the national level. (Taxes and other Massachusetts-specific factors are discussed below.) Prices for 1901-1911 are from United States Department of Commerce (1915). Prices are available annually for the principal brands of the American Tobacco Company (the Tobacco Trust), exclusive of Turkish cigarettes. The principal brands comprised a majority of sales, with one brand accounting for three-fourths of domestic sales in the beginning of the period. These are government assembled totals completed in the wake of the the Supreme Court's break-up of the Trust in 1911. (No data are available for 1911 and prices are interpolated for that year). There is also data from this source for 1912-1913 for the Trust's successor companies, which is combined with the sources listed below.

Prices for 1912-1950 are based on Nicholls (1951). This source lists the date and level of all list price changes for the main brands. The market was quite concentrated during this period and just the three leading brands (Lucky Strikes, Camel, Chesterfield) accounted for almost all sales until the 1930s and two-thirds of sales through 1950 (Maxwell, various years; Nicholls, 1951). The data include prices for all brands, including economy/generics, which account for virtually all domestic cigarette sales. The price level and date of change were checked against Tennant (1950) and there are only a few and relatively minor discrepancies. Massachusetts cigarette excise taxes (a per unit tax) were first introduced 11 August 1939 and are added onto these prices. (Federal excise taxes are included in the list price.)

No data are available for 1951-1954 and interpolation is used. The only change in taxes during this time was a one cent per pack increase in the federal excise tax on cigarettes on 01 November 1951.

Data for 1955-2011 are from Orzechowski and Walker (2011), which lists average retail price by state. Prices are the market share-weighted average of prices of all brands based on surveyed consumer prices in Massachusetts for fiscal years ending 30 June. A separate series, which includes generics brands, is included starting in 1991. Prices include state and federal cigarette excise taxes but do not include sales taxes. Prices were adjusted to reflect the sales tax after Massachusetts removed the exemption for cigarettes on 01 July 1988.

Figure C2 graphs the resulting price time series. This figure uses the minimum

price across brands and omits generics. In our main analysis we focus on the average price exclusive of generics. The omission of generics is relatively innocuous since for the years of our main model (1952-1996) the different summary statistics (of prices with and without generics) are virtually identical (i.e., generic/economy brands were prominent for three periods — 1901-1910, 1931-1950, and after 1990 — which only overlap with the very end of our observation period). Figure C2a shows prices over the century. While they appear relatively stable, note the wide-range of the vertical axis. (In fact, the post-2000 period is omitted from the graph since prices continue to rise and this would further obscure the variation.) Prices collapse almost in half after the dissolution of the Tobacco Trust in 1911. Prices then rise and fall repeatedly during the 1920s , 1930s, and 1940s. Prices then steadily rise for the next two decades, dip again, and finally increase sharply in the 1990s. Figure C2b shows that, on average, half of this price is composed of taxes (i.e., state and federal excise taxes on cigarettes as well as state sales tax). This information is helpful for identification since the tax share is one of the main sources of variation in prices, and it oscillates for non-demand reasons (e.g., taxes rise during World War 2 when fiscal demands necessitated the creation of the state excise tax, initially an emergency measure, and increases in the federal tax).

Figure C2: Annual Real Cigarette Prices and Taxes

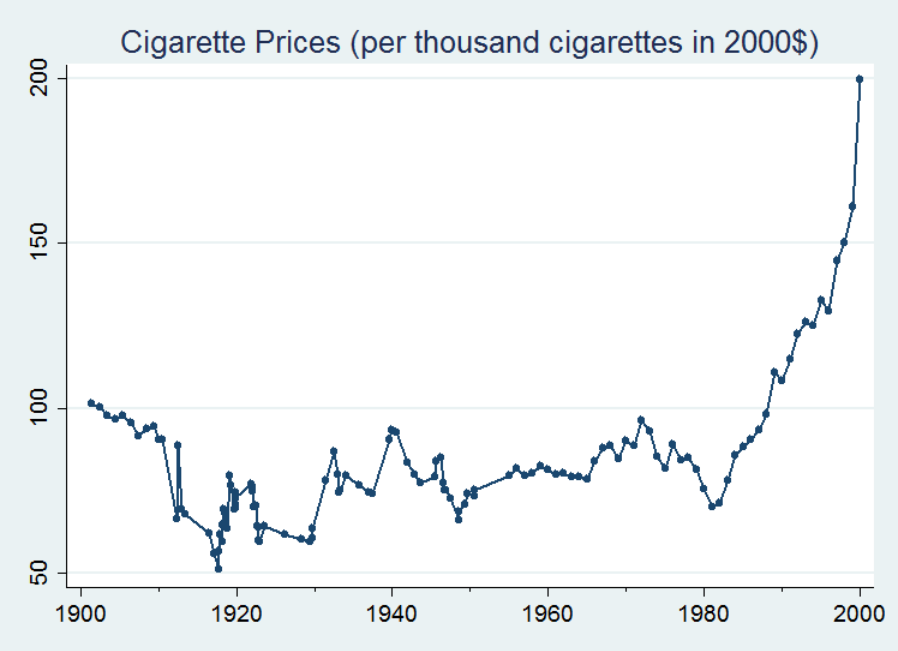


Figure C2a.

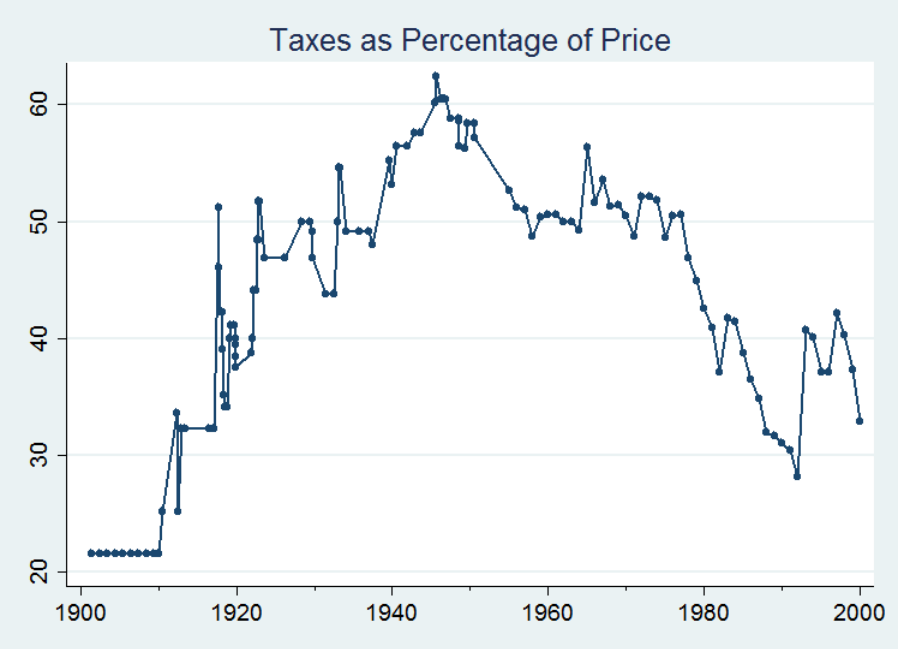


Figure C2b.

D The Role of Unobserved Heterogeneity

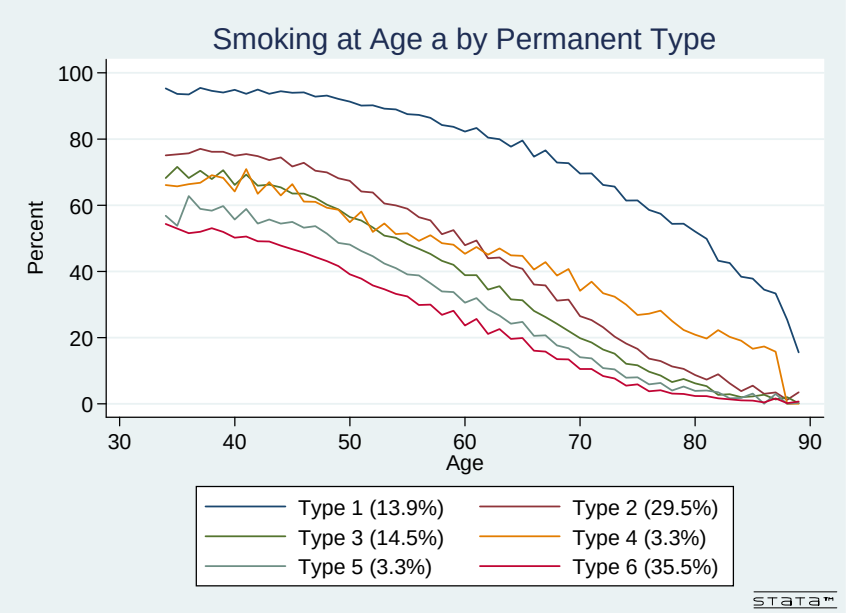
While the rich observed lifetime health and smoking heterogeneity of individuals plays an important role in explaining the mortality rates of individuals with different lifetime smoking patterns by addressing concern over confounding (using observable data), correlated individual UH also plays an important role. Its main function is to capture the correlation, through unobservables, among the modeled behaviors and outcomes that would otherwise bias estimated impacts of the smoking and health histories. Our jointly estimated model allows for UH that is likely to be common across a lifespan (such as genetics, risk-aversion, time preference or self-esteem, for example) as well as differences that may vary over time (such as unobserved stress or health, for example). We model these two types of UH using discretized distributions characterized by mass point vectors that describe the impact of each type of heterogeneity on the outcomes of interest. Appendix Table B2 displays the estimated coefficients and standard errors that capture the distributions. Estimated probability weights of each discrete mass point vector are listed in the last column.

We use simulation in the paper to evaluate our estimated model. Recall that in these simulations we replicate, R times, the exogenous characteristics of all individuals in our estimation sample, N. For each replication, we use the estimated correlated UH distributions to draw a permanent "type" that is common for that replicated individual across all time periods and draw, at every time period, a second "type" that may be different each period. Using the estimated model to simulate lifetime behavior and outcomes (from an individual's observed initial age through age 100), we show that lifetime smoking probabilities differ by these unobserved types. While it is difficult to depict the differences associated with the time-varying UH, we can condition on (simulated) permanent UH type and plot the resulting smoking rates by age (Figure D1). We order the "types" by the simulated smoking probability at age 40 (i.e., highest to lowest).

The distribution of permanent UH suggests that about 14 percent of the sample (type 1) are as much as 20 percentage points more likely to smoke at any given age than the other 86 percent. The figure correctly shows that this time-invariant unobserved determinant of smoking shifts smoking probabilities uniformly, unconditional on smoking and health histories, at each age.[1] Additionally, the model allows for behavior and outcome shifters each two-year period based on a draw from the distribution capturing time-varying unobservables. A likelihood ratio test comparing the goodness of

[1]The permanent UH is linearly added to the operand of the linear (OLS) and non-linear (LOGIT or MLOGIT) operators. For the latter, this does not translate into an intercept shift but also depends on the level of the product of observed variables and their coefficients.

Figure D1: Smoking Probabilities by Permanent Unobserved Heterogeneity



Note: The colored lines indicate each heterogeneous unobserved type, with simulated probabilities (drawn from the estimated probability distribution) in parentheses.

fit of the nested models with and without correlated UH suggests that the model with such UH fits significantly better.

Using the same ordering based on smoking propensity, we also report the mortality outcomes for each permanent UH type in Table D1. We see that those individuals with an UH type that makes them more likely to smoke also experience the shortest lifespan with an average age of death of 71.8. They also have the highest proportion of cancer deaths. Types 4 and 6, who are less likely to smoke, have the highest mean (at 75.1 and 74.3, respectively) and percentile ages of death. Type 3 individuals are much more likely to have a CVD-related death than any of the other types. Finally, note that type 2 captures individuals with high smoking rates yet longer than average expected lifetimes.

These variations in behavior and health outcomes are picked up by our modeling of the correlated UH. While we cannot know specifically what the UH captures, we can hypothesize its role. One example that we have not previously discussed — an aspect of behavior that we cannot include explicitly due to data limitations — is smoking intensity, which may have different impacts on morbidity and mortality outcomes as well as smoking behavior over the lifetime. The sporadic intensity data that we do observe suggests that our sample contains relatively heavy smokers with only about

Table D1: Age and Cause of Death by Permanent Unobserved Heterogeneity

Permanent	Simulated		Age of death distribution (percentile)					Cause of death		
UH Type	Percent	Mean	10th	25th	50th	75th	90th	CVD	Cancer	Other
1	13.9	71.3	57	64	72	79	85	29.3	30.7	40.0
2	29.5	73.6	58	66	75	82	88	35.9	29.6	34.5
3	14.5	72.9	58	66	74	81	87	56.9	16.1	27.0
4	3.3	75.1	62	68	75	82	88	43.5	24.4	32.1
5	3.3	73.8	59	67	75	81	87	43.0	14.9	42.1
6	35.5	74.3	59	67	75	82	88	46.3	24.5	29.1

15 percent of smokers reporting smoking less than a pack of cigarettes a day. The permanent and time-varying UH that we model potentially addresses the variation in smoking intensity that our dichotomous smoking indicator "averages over".

Generally, these conditional (on UH type) death distributions reflect 1.) differences in lifespan due to unobserved permanent factors (like genetics or time preferences) as well as 2.) differences in smoking behavior (as illustrated in Figure D1). While we cannot say exactly what the UH captures, knowing these different smoking and mortality patterns by type gives us insight into both the estimation results and policy recommendation. First, the (unconditional on type) death distribution would be different if UH were ignored. (We see this in the biased coefficients of the model without UH.) It is not simply that inclusion of UH improves precision by reducing important selection, endogeneity, and measurement error biases, but it allows different lifetime smoking patterns which in turn have non-linear feedback effects (on both health and subsequent smoking) via the dynamic system of equations. Second, policy evaluation should be more sensitive to distributional issues knowing there is heterogeneity in the population in terms of smoking initiation rates, quit rates, relapse rates, and mortality rates. We find, for example, that some individuals are more predisposed to smoke, but only some of these have shorter expected life spans.

References

Annotated Laws of Massachusetts (2007). *PART I ADMINISTRATION OF THE GOVERNMENT. TITLE IX TAXATION. Chapter 64C Cigarette Excise.* Tech. rep., General Court of the Commonwealth of Massachusetts, https://malegislature.gov/Laws/GeneralLaws/PartI/TitleIX/Chapter64C.

Becker, G. S. and Murphy, K. (1988). A Theory of Rational Addiction. *Journal of Political Economy*, **96** (4), 675–700.

Borden, N. (1942). *The Economic Effects of Advertising.* Richard D. Irwin, Inc.: Chicago.

Brandt, A. (2007). *The Cigarette Century.* Basic Books: New York.

Bureau of Labor Statistics (2013). *Consumer Price Index, All Urban Consumers - (CPI-U), U.S. city average, All items.* Tech. rep., Bureau of Labor Statistics, http://www.bls.gov/cpi/.

Chaloupka, F. J. and Warner, K. E. (2000). The Economics of Smoking. *Handbook of Health Economics*, **1B**, 1539–1627.

Federal Trade Commission (2013). *Cigarette Report for 2011.* Tech. rep., Federal Trade Commission, http://www.ftc.gov/reports/federaltradecommission-cigarettereport2011.

Fujii, E. (1980). The Demand for Cigarettes: Further Empirical Evidence and Its Implications for Public Policy. *Applied Economics*, **12**, 479–489.

Gruber, J. and Koszegi, B. (2001). Is Addiction "Rational"? Theory and Evidence. *The Quarterly Journal of Economics*, **116** (4), 1261–1303.

Kannel, W. B., Wolf, P. A. and Garrison, R. (1988). *The Framingham Study: An Epidemiological Investigation of Cardiovascular Disease.* Tech. rep., US Department of Health and Human Services, National Institutes of Health.

Maxwell, J. C., Jr. (various years). *Historical Sales Trends in the Cigarette Industry.* Tech. rep., Wheat, First Securities, Inc.

Merriman, D. (2010). The Micro-geography of Tax Avoidance: Evidence from Littered Cigarette Packs in Chicago. *American Economic Journal: Economic Policy*, **2**, 61–84.

Nicholls, W. (1951). *Price Policies in the Cigarette Industry: A Study of 'Concerted Action' and Its Social Control, 1911-50.* Nashville: Vanderbilt University Press.

Orzechowski and Walker (2011). *Tax Burden on Tobacco: Historical Compilation, vol 46.* Tech. rep., Orzechowski and Walker Consulting Firm, http://www.taxadmin.org/fta/tobacco/papers/Tax_Burden_2011.pdf.

Pierce, J. and Gilpin, E. (1977). A Historical Analysis of Tobacco Marketing and the Uptake of Smoking by Youth in the United States: 1890–1977. *Health Psychology*, **14**, 1–9.

Sahr, R. (2013). *Inflation Conversion Factors for Years 1774 to estimated 2018.* Tech. rep., Oregon State University, http://oregonstate.edu/cla/polisci/faculty-research/sahr/infcf17742007.pdf1.

Schneider, L., Klein, B. and Murphy, K. (1981). Governmental Regulation of Cigarette Health Information. *Journal of Law and Economics*, **23**, 575–612.

Telser, L. (1962). Advertising and Cigarettes. *Journal of Political Economy*, **70**, 471–499.

Tennant, R. (1950). *The American Cigarette Industry: A Study in Economic Analysis and Public Policy..* New Haven, Yale University Press.

United States Census Bureau (2000). *Historical National Population Estimates: July 1, 1900 to July 1, 1999 .* Tech. rep., United States Census Bureau, http://www.census.gov/popest/data/national/totals/pre-1980/tables/popclockest.txt.

United States Census Bureau (2011). *Table 1. Intercensal Estimates of the Resident Population by Sex and Age for the United States: April 1, 2000 to July 1, 2010 (US-EST00INT-01).* Tech. rep., United States Census Bureau, http://www.census.gov/popest/data/intercensal/national/tables/US--ST00INT01.csv.

United States Census Bureau (2012). *Historical Statistics of the United States, Colonial Times to 1970.* Tech. rep., United States Census Bureau.

United States Department of Commerce (1915). *Report of the Commissioner of Corporations On the Tobacco Industry, Part III: Prices, Costs and Profits.* Tech. rep., Department of Commerce, Washington: Government Printing Office.