Abstract

Many studies have linked “candidate” genetic variants (SNPs) with smoking behaviors and with smoking-related health problems. While the biological mechanism is under active investigation, biological research alone cannot answer policy relevant questions on smoking behavior. We develop a life-cycle model of smoking initiation, continuation, cessation, health beliefs, and health outcomes that incorporates genetic heterogeneity. We estimate our model using newly genotyped data from the Health and Retirement Study (HRS). We find the two SNPs that impact smoking to operate through different behavioral channels. Strikingly, both SNPs have effects on smoking-related illness and mortality that are an order of magnitude larger than their measured connection with smoking would suggest. This shows that standard measures of smoking (e.g. maximum cigarette consumption) are only weakly related to the health risks associated with life-cycle smoking.

1 Introduction

Many studies have linked “candidate” genetic variants (SNPs) with smoking behaviors and with smoking-related health problems. While the biological mechanism is under active investigation, biological research alone cannot answer policy relevant questions on smoking behavior. We develop a dynamic life-cycle model of smoking initiation, continuation, cessation, health beliefs, and health outcomes that incorporates genetic heterogeneity. We estimate our model using newly genotyped...
data from the Health and Retirement Study (HRS). We use our model to explore the impact on smoking and health of proactive provision of information on genetic risks.

Our first finding is that two of the candidate SNPs impact smoking through different behavioral channels. The second and more dramatic finding concerns health effects. A common complaint about single SNP studies has been that, while effects may be statistically significant, they explain relatively little of the variation in the phenotype. The effect on measured smoking is indeed of this nature: our SNPs explain less than 0.5% of variation in measured smoking. Yet the effects on later-life health and mortality are an order of magnitude larger than this. For example, amongst smokers, variation in a single SNP accounts for more than a 6.5 percentage point difference in the probability of getting chronic pulmonary obstructive disease, whose primary risk factor is smoking. This is particularly striking given that the base rate among smokers is 18%! This shows that standard measures of smoking in the epidemiological literature are only weakly correlated with the health damage induced by life-cycle patterns of smoking behavior.

The relationship between genes and smoking is relatively well-studied. Moreover, while many behavioral phenotypes are known to be heritable, smoking is unusual in that the biological basis for heritability is particularly well understood. Specifically, biologists have long theorized that smoking may be impacted by genetic variations that alter the functioning of nicotinic acetylcholine receptors (NACHRs) that play a critical role in the processing of nicotine. Epidemiological geneticists been indeed found this to be so (see section 2.1). There is now clear-cut evidence that three SNPs, all of which impact the functioning of NACHRs, impact various forms of smoking behavior (see section 2.2). There are related findings suggesting that these variations impact health through their impact on smoking, although the health findings remain less clear-cut (see section 2.3).

Understanding that nicotinic receptors link underlying genetic variation with smoking and related health consequences represents a profound advance in knowledge. Yet further biological research alone cannot answer policy relevant questions on smoking behavior. What is needed for these purposes is a dynamic model of behavior, health, and learning. Parameters in such a model identify how genes impact smoking decisions over the life cycle. These parameters must be consistent with observations on patterns of smoking initiation, continuation, and cessation. They must also capture how later-life health evolves as a result of the life cycle pattern of smoking behavior. Finally, they must capture the impact of policies such as early provision of information on genetic risks designed
to encourage early cessation. In principle, provision of such information may reduce the number of those who get their bad news through a tragic personal or familial experience. Capturing this rich range of behaviors, beliefs, and outcomes in a reasonable manner is essentially impossible absent a coherent dynamic model.

We develop and estimate a dynamic life-cycle model of beliefs, smoking, and health that incorporates genetic heterogeneity. The model allows for genetic variation to separately affect preferences for nicotine and the utility-costs of reductions in smoking. Hence the parameter estimates allow us to identify which of these channels is affected by each genetic variant linked to smoking. By explicitly specifying health beliefs, the model enables us to capture the impact of policies that involve proactive revelation of genetic information.

We estimate our model parameters using newly genotyped data from the Health and Retirement Study (HRS). As discussed in Section 2, we focus our attention on three SNPs which have been implicated in various dimensions of smoking behavior. Descriptive analyses in the HRS suggest that two of these three candidate SNPs have significant effects on smoking behavior, and it is these two that we explore further in the structural model. The data reveal that these SNPs operate through different behavioral channels. Specifically, one is found to be strongly associated with peak smoking intensity (maximum cigarettes per day), while the other is more strongly associated with cessation later in life. The fact that distinct SNPs operate through distinct behavioral channels highlights the need for an explicit model of life cycle smoking behavior. Such differences increase the salience of genetically-targeted medical interventions that are starting now to gain traction.

In a counterfactual policy experiment, we are exploring the effects of revealing genetic information (making individuals aware of their genotype at SNPs rs169 and rs4950). Absent such revelation individuals in our model are not perfectly aware of their propensity to experience difficulty in reducing or quitting cigarette use. It is for this reason that revealing such information can be of potential value to individuals. We are also conducting a number of robustness checks, in particular running a specification that ignores any correction for sample selection bias (individuals must survive to be genotyped). The sensitivity of the structural parameter estimates identifies the importance of dealing with selection bias when interpreting reduced-form associations between genes and behavior.

Looking forward, our paper suggests interdisciplinary gains from trade as genetic data becomes
increasingly available. Economists will not make any substantial progress understanding genetic effects without successful biological discovery. The pooling of data sets and analysis of underlying biology is essential if one is to avoid following false leads. Yet geneticists and other biologists have a great deal to learn from economic modeling methods whenever behavior intervenes between gene and phenotype, as it does for smoking. Other addictive behaviors and dietary choices are clear candidates for further interdisciplinary and collaborative work.

Section 2 provides a brief review of recent findings on genes, smoking, and health, and details our procedure for selecting candidate genes for our analysis. In section 3 we introduce the genomic HRS and present important correlations, many of which align with those in the epidemiological literature. In section 4 we introduce the structural model. In section 5 we present our structural results. In section 6 we conduct robustness exercises and analyze issues of sample election. Section 7 presents our policy analysis. Section 8 concludes.

2 Background to Study

In section 2.1 we provide relevant background on the biology of smoking. In section 2.2 we summarize the literature on candidate SNPs and smoking. In section 2.3 we summarize the somewhat more controversial findings concerning possible health impacts. In section 2.4 we rationalize our selection of three candidate SNPs and summarize findings concerning their possible impacts on other phenotypes.

2.1 Genetics and Smoking

Human DNA is composed of a sequence of about 3 billion pairs of nucleotide molecules, each of which can be indexed by its location in the sequence. At the overwhelming majority of locations, there is virtually no variation in the nucleotides across individuals. The segments of DNA where individuals do differ are called genetic polymorphisms. The most common kind of genetic polymorphisms are called a single-nucleotide polymorphism (SNP). SNPs are locations in the DNA sequence where individuals differ from each other in terms of a single nucleotide. At the vast majority of SNP locations ("loci"), there are only two possible nucleotides that occur. The nucleotide of a SNP that is more common in the population is called the major allele, and the nucleotide that
is less common is called the minor allele.

There is a large body of evidence indicating that most behavioral traits—including risk and time preferences, cognitive ability, and personality—are heritable, with up to half of the variation across individuals statistically accounted for by genes taken as a whole (Benjamin et al. 2012). For most of these traits, however, efforts to identify specific genes associated with the trait have so far largely failed. Smoking (like a few other behaviors such as alcohol consumption and food preferences) is an exceptional outcome variable on two counts. First and most importantly, smoking is more “biologically proximal” than other behavioral traits in the sense that variation across individuals in their genes matters fairly directly and powerfully, with relatively few other intervening factors, for how the body processes nicotine, as we discuss below. Second, unlike other behavioral outcomes and because of its relevance for health, smoking has been measured in many datasets with genotyped individuals, and many large-scale genetic association studies have been published.

The most comprehensive study of the genetics of smoking conducted to date was a collaborative genome-wide association (GWAS) that looked for SNPs associated with various dimensions of smoking behavior (initiation, cessation and cigarettes per day, CPD) (Consortium 2010). In a GWAS, which has become the standard tool for gene discovery in modern epidemiology, researchers test hundreds of thousands of SNPs for association with the outcome of interest, one SNP at a time. The consortium GWAS identified three genomic regions (“loci”) containing SNPs that whose association reached the conventional significance threshold $5 \cdot 10^{-8}$ (lower than standard threshold because of the large number of hypotheses tested). Interestingly, many of the identified variants were found in a cluster of nicotinic receptor genes located on chromosome 15. This cluster comprises three genes - \textit{CHRNA3}, \textit{CHRNA5} and \textit{CHRNB4} - which encode subunits of the nicotinic acetylcholine receptors (NACHR’s) and the genes had therefore featured prominently in the earlier literature on “candidate genes”. The specific SNPs we study, however, had not been studied prior to the recently published GWA studies and many of the associations identified in the candidate gene literature had inconsistent replication records (Munafo et al. 2004).

2.2 Candidate SNPs

Our analyses in this paper are all based on two SNPs in this cluster and a third SNP located on chromosome 8. All have been robustly associated with at least one dimension of smoking behavior.
The first SNP, rs16969968, has been shown to confer a functional effect on receptor function through an amino acid change. In smokers, there is consistent evidence that the SNP is robustly associated with the number of cigarettes smoked (Bierut 2009, Berrettini et al. 2008, Spitz et al. 2008, Stevens et al. 2008, Consortium 2010) and nicotine dependence (Saccone et al. 2007, Saccone et al. 2009, Spitz et al. 2008, Stevens et al. 2008). There is some evidence that this SNP is associated with cessation, especially in the face of health shocks, though this evidence is not always consistent (Baker et al. 2009, Breitling et al. 2009, Conti et al. 2008, Freathy et al. 2009, Thorgeirsson et al. 2008, Consortium 2010). Studies attempting to relate the SNP to other smoking phenotypes, such as age at initiation or age at which an individual starts smoking regularly, have not yielded any consistent findings (Consortium 2010, Weiss et al. 2006). The SNP is colloquially referred to as “Mr Big”, because compared to other behavior traits, the effect of the SNP is quite large.

Our second SNP, rs680244, is correlated with rs16969968, but it nevertheless has been shown to be additionally related to number of cigarettes smoked conditional on rs16969968 (Saccone et al. 2010). Our final SNP rs4950 (in gene \textit{CHRNB3}), has also been linked to nicotine dependence (Saccone et al. 2009, Zeiger et al. 2008) though its link to number of cigarettes smoked is mixed (Thorgeirsson et al. 2010, Zeiger et al. 2008). Unsurprisingly, since these SNPs have been found to be linked to smoking behavior, they also have been linked to a variety of respiratory and circulatory diseases (Amos et al. 2010, Hung et al. 2008, Pillai et al. 2009, Spitz et al. 2008, Thorgeirsson et al. 2008).

We note that both of the genes, CHRNA5 and CHRNB3, code for subunits of the NAChr’s. These receptors are an important part of the body’s dopamine system. In natural “reward” processes, when acetylcholine binds with the NAChr’s, it sets off a chain reaction on the cell that ultimately releases dopamine in the brain, indicative of a positive reward prediction error that incentivizes repetition. Due to structural similarities between acetylcholine and nicotine, nicotine ingested through smoking can also bind to the NAChr’s and trigger this reaction, leading to addiction. The precise mechanism through which mutations in CHRNA5 and CHRNB3 operate in the NAChr’s is not well understood, though lab experiments suggest that the risk-conferring alleles in rs16969968 may influence the strength of the reaction of NAChr’s to nicotine (Bierut 2009) and also the expression of the CHRNA5 gene in brain and lung tissue (Wang et al. 2009, Falvella et al. 2009).
3 Data and Reduced Form Findings

3.1 HRS and Genomic Data

The data for our analysis come from the Health and Retirement Study (HRS), which is a nationally representative longitudinal survey of Americans over 50 years of age and their spouses. The initial HRS sample was collected in 1992 and included individuals born between 1931 and 1941. The survey is administered every two years with only minor adjustments from wave to wave. More cohorts have been added over time, making the current HRS sample representative of individuals born between 1890 and 1954 who survived until the sample period.

From 2006 to 2008, 12,507 HRS respondents were genotyped from saliva samples. To avoid detecting spurious genetic associations due to genotyping errors, it is important to analyze data that have undergone quality control filtering (see Beauchamp et al. 2011 for discussion). We work with the public-release version of the genotypic data which has been quality controlled by researchers at the University of Washington (2012). We further restrict our sample to caucasians, since the genetic associations that motivate our study are largely found in all caucasian samples. Our final genotyped sample consists of 68,288 person-year observations on 8,122 unique individuals. The Appendix offers a complete discussion of the criteria used to select this sample. Table 1 presents some basic cross-sectional characteristics of the individuals in our sample, with the variables measured as of an individual’s most recent appearance in the panel.

The HRS survey contains several questions on past and current cigarette consumption that we use in our descriptive analysis. As indicated in Table 1, about 57% of our sample report having ever smoked, and conditional on smoking, the average maximum number of cigarettes consumed per day is just over 25.

In the descriptive analyses that follow, we estimate regressions of the following form:

\[
y_i = \beta_0 + \sum_j \beta_j SNP_{ji} + X_i'\gamma + \epsilon_i,
\]

where \( SNP_{ji} \in \{0, 1, 2\} \) is the genotype of individual \( i \) at SNP \( j \) and \( X_i \) is a matrix of controls. Controlling for potential confounds that may be correlated with genotype is critical in order to avoid spurious findings.
In practice, the most common concern is confounding due to population stratification: different groups within the sample differ in allele frequencies and also differ in their outcome for non-genetic reasons.\footnote{A famous illustration of stratification is the “chopsticks effect” (Lander & Schork, 1994). Imagine a study that tries to identify genetic markers for chopstick use by comparing a Japanese population (cases) to a Caucasian population (controls). Without controlling for population stratification, any markers which differ appreciably in frequency between the Caucasian and Asian populations will be found to be associated with chopstick use, but those associations are of course spurious. This example might seem to suggest that a simple fix would be to control for race or ethnicity. Indeed, it is standard practice to restrict a genetic association study to subjects of a common ethnic background, as we do here. It has been found, however, that allele frequencies can differ substantially even within ethnically homogeneous populations, such as different regions within Iceland (Price et al., 2009).} For this reason, it is common practice in genetic association studies to include as control variables the first 10 or 20 principal components of all the genotypes measured in the dense SNP chip. These principal components seem to pick up much of the subtle genetic structure within a population (Price et al. 2006). Our analyses therefore control for the first ten principal components, provided by the HRS. We also include a dummy for Male gender, a full set of age dummies, and interactions between the Male and Age dummies.

3.2 Genes, Smoking, and Health Outcomes

Table 2 presents estimates of the relationship between our SNPs of interest and some commonly analyzed measures of smoking behavior. We start with at the maximum number of cigarettes consumed per day. In Columns 1-3, we consider the influence of each SNP separately. We find that each additional copy of the reference allele (G/A) of \textit{rs16969968} is associated with an increase of 1.32 cigarettes, an association that is very close to the point estimate reported in ?. Each extra copy of \textit{rs4950} is associated with a reduction of 0.72 cigarettes, although this estimate is only statistically significant at the 0.10 level. We find no statically significant association between \textit{rs680244} and maximum cigarettes per day. In Column 4 of Table 1, we include all three SNPs in a single regression, and only the estimated coefficient on \textit{rs16969968} remains statistically significant.

Though informative, the maximum number of cigarettes consumed per day is only one facet of life-cycle smoking. Reduction or cessation represents another critical feature of smoking behavior. In Columns 4-5 of Table 2, we investigate the relationship between our SNPs and two measures of change in extensive-margin behavior. In Column 1, our outcome variable is an indicator for
smoking status conditional on ever-smoking, and the sample is restricted to an individual’s most recent year in the HRS. We find no statistically significant association between later-life smoking and either rs16969968 or rs680244. However, each additional copy of the minor allele at rs4950 is associated with a 2.3 percentage point reduction in the probability of continued smoking. Given that less than 17% of ever-smokers remain smokers as of their last observation in the HRS, the size of this relationship is substantial. In Column 6, we examine a measure of in-sample quitting. We restrict ourselves to individuals who were smoking in their penultimate observation in the HRS (two years before the final observation), and define a quit as taking place if an individual does not report smoking in their last observation in the HRS. We again find statistically significant results only for rs4950. One extra copy of the minor allele for this SNP is associated with a 5 percentage point increase in the probability of quitting.

One advantage of the HRS for this analysis is the availability of life-cycle data on health outcomes and mortality. This allows us to directly estimate the relationship between specific SNPs related to cigarette consumption and major illnesses associated with smoking. We are particularly interested in lung disease, heart disease, and cancer, since these are the major conditions directly linked to smoking. HRS respondents are asked about their current health status in each of these three categories, along with a series of follow-up questions. For example, the first question about lung disease asks subjects if they have ever been told by a doctor that they have a lung condition such as “chronic bronchitis or emphysema”. In subsequent surveys, respondents are asked if their medical conditions is improving or deteriorating and information is also collected about any treatment received or medications prescribed.

Tables 2-3 reports estimates of linear probability models explaining health outcomes as a function of the SNPs, the maximum number of cigarettes smoked per day, and the controls listed above. The dependent variables are indicators for the incidence of (non-cancerous) lung illness, heart illness, and cancer. These are cross-sectional regressions with samples restricted to include only the most recent person-year observation in the HRS. The samples for these regressions only include those individuals that have reported smoking at some point in their lives. Later we will also consider non-smokers. To obtain a baseline association between cigarette consumption and lung health risks, Column (1) of Table 2 presents an estimate of the relationship between the maximum reported number of cigarettes smoked per day and the incidence of major non-cancerous lung ill-
ness. These estimates suggest a positive and significant relationship, with one extra cigarette per day associated with an increase in the probability of lung illness of about 0.3 percentage points. In Column (2), we regress the lung illness indicator against the SNPS (without cigarettes). We find a large, statistically significant coefficient on rs16969968, but no significant relationship with our other two SNPs. The coefficient on rs16969968 is large and consistent with the direction of its association with smoking behavior. Each extra copy of the minor allele at rs16969968 (positively associated with smoking) is associated with a 3.4 percentage point increase in the probability of lung illness.

In Table 4, we repeat this analysis with two additional health indicators: the incidence of a major heart illness, and the diagnosis of cancer. Although the maximum number of cigarettes per day is associated with elevated risks for heart disease and cancer, we generally find small, statistically insignificant relationships between our SNPs of interest and these outcomes. For the cancer outcome, this is partially explained by the fact that we are unable to specifically isolate lung cancer. The HRS survey only asks if an individual has ever been diagnosed with any cancer, regardless of the type. Since smoking is less strongly associated with other cancers, the lack of a strong relationship is not terribly surprising.

In Table 5, we investigate the relationship between our SNPs and mortality. Specifically, estimate a linear probability model to explain death in the next year. We pool all person-year observations for ever-smokers from 2006 onwards. The year restriction is imposed because individual had to survive until 2006 in order to be genotyped. Reported standard errors are clustered at the person-level. In Column (1) of Table 5, we find statistically significant associations between for both rs16969968 and rs4950. Each extra copy of the minor allele at rs16969968 (positively associated with max cigarettes per day) is associated with a 0.4 percentage point increase in the probability of death. Each extra copy of the minor allele at rs4950 (positively associated with cessation) is associated with a 0.3 percentage point reduction in the probability of death.

3.2.1 Interpreting the Reduced Form Evidence: Questions and Puzzles

One challenge in interpreting the gene-health associations presented here is the possibility that our SNPs work through channels other than smoking. For example, if rs16969968 affects both smoking and some other biological process related to lung health or mortality (e.g. fragility of lung
tissue), it becomes difficult to properly model the causal chain running through genes, smoking, and health. If our SNPs operate through non-smoking channels, presumably these channels would produce differences in the health outcomes of never-smokers with these SNPs. In Tables 6-7, we re-estimate our basic health and mortality specifications using the sample of genotyped never-smokers. We fail to find any statistically significant relationships between our SNPs and these outcomes among never-smokers. This strongly suggests that the gene-health associations documented earlier are driven by differences in smoking behavior.

The collection of reduced form evidence presented here suggests a complicated set of relationships between individual SNPs, smoking behavior, and health. We find strong effects of rs16969968 on the peak quantity of cigarettes consumed, as well lung health and mortality. However, we find no association between rs16969968 and cessation behavior later in life (in the HRS). Conversely, we do not find strong effects of rs4950 on cigarette quantity or lung health, but we find effects on cessation and morality. How can we reconcile these patterns?

The magnitude of the association between rs16969968 and lung health is particularly noteworthy. Indeed, this far exceeds the magnitude we would predict by naively considering the relationship between rs16969968 and max. cigarettes per day, and the relationship between max. cigarettes per day and lung illness. Each additional copy of the minor allele at rs16969968 is expected to increase max. cigarettes by 1.3, which should increase the probability of lung illness by \( \frac{1}{3} \times 0.3 = 0.39 \) percentage points. Our point estimate for the effect of an additional copy of the minor allele is about 10 times as large.

How can we reconcile the modest effect of rs16969968 on maximum cigarettes per day with the large effect on lung health? One possibility rests on the insufficiency of a simple metric like max. cigarettes as a measurement of life-cycle smoking behavior. Since rs16969968 has no known relationship with the functioning of lung tissue, we argue that the association with health emerges because of the effect of this SNP on life-cycle smoking patterns. An individual’s health is a function of not only maximum smoking intensity, but also the total length of time that an individual sustained that maximum intensity. The lung health association might better reflect the total life-cycle effect of rs16969968 on cumulative smoking behavior than the observed associations between rs16969968 and maximum cigarettes. Indeed as shown in Figure 1, individuals who continuously smoke in the NLSY on average experience a substantial reduction in the quantity of cigarettes that
they smoke per day over their life-cycle. It is possible that a SNP like rs16969968 not only affects peak quantity, but the evolution of quantity over time.

The operation of dynamic behavioral channels can also potentially explain the set of associations observed for rs4950. For example, it appears that rs4950 affects the ease of cessation. It is possible that this association is independent of the behavioral channels that affect the maximum quantity consumed (e.g. one’s preference for nicotine).

Rationalizing the collected associations between genes, smoking, and health requires developing a unified dynamic model

4 Model

Here we develop a dynamic structural model of life-cycle smoking behavior. A sizable existing literature uses the theory of rational addiction (7) to organize the empirical analysis of smoking. ? and Becker et al. (1994) present evidence in favor of the model’s prediction that both past and future cigarette prices should affect current consumption. (See Chaloupka & Warner (2000) for a survey). ? also finds indirect evidence that less educated and younger individuals are more myopic because their contemporaneous cigarette demand is less related to future consumption and prices. ? also find evidence for state dependence in cigarette consumption, consistent with the notion of habit-formation present in the Becker-Murphy model.

A smaller, fully structural literature jointly models smoking decisions along with health and mortality processes. This approach allows for a rigorous quantification of how health risks (or beliefs about health risks) alter the incentives to smoke over the life-cycle. ? develop and estimate one of the first structural models of smoking, health, and mortality in a sample of mature adults from the Health and Retirement Study (HRS). They find evidence in favor of forward looking behavior and support for habit formation in the form of substantial quitting costs. Darden (2013) develops and estimates a structural model of smoking decisions and focuses on the role of individual (Bayesian) learning about the health risks of smoking. He finds evidence that smokers quit in response to the onset of chronic illnesses, but are less likely to respond to new information about individual health markers such as blood pressure and high-density lipoprotein. Our model builds on the basic framework present in ? and ?.
4.1 Choice Set and Addiction Stocks

We model smoking as a discrete choice. Each period, individuals choose one of $J + 1$ levels of smoking: $\{c_0, c_1, \ldots, c_J\}$, where $c_0 = 0$ represents the non-smoking option, and more generally $c_j$ represents the quantity of cigarettes consumed per day under option $j$. We allow smokers to choose one of four intensities: $\{0, 5, 20, 30\}$. Let $C_{it}$ represent individual $i$’s cigarette consumption in period $t$.

We assume that smoking is associated with two kinds of persistent effects. First, current cigarette consumption fuels an addiction to nicotine that makes it difficult to reduce cigarette consumption in the future. The intensity of this addiction is captured by the addiction stock $S_{it}^a$. We assume that this evolves according to the following law of motion:

$$S_{it}^a = [1 - (\delta_{a1} + \delta_{a2} CT_{it})] S_{it}^a + \zeta_a C_{it}$$ (2)

Here $(\delta_{a1} + \delta_{a2} CT_{it})$ represents the rate at which the addiction stock depreciates, and $\zeta_a$ represents the rate at which current smoking contributes to the growth of the addiction stock. The variable $CT_{it}$ is an indicator for a “cold turkey” quit, which we define as a transition from moderate or heavy smoking to no smoking: $CT_{it} = 1\{C_{it} = 0 : and : C_{it-1} > 5\}$. When $\delta_{a2} > 0$, this corresponds to a situation in which heavy smokers find it easier to shake a habit if they completely eliminate cigarette consumption, rather than cutting back and trying to manage their addiction at a lower level of intensity.

In addition to fueling a behavioral habit, smoking may also have a persistent effect on an individual’s health. We assume that such effects are related to a separate stock, $S_{it}^h$, which reflects the latent potential for past smoking to induce negative health events. We refer to this as the smoking health stock, and it evolves according to the following law of motion:

$$S_{it}^h = (1 - \delta_h) S_{it}^h + \zeta_h C_{it}$$ (3)

Here $\delta_h$ represents the annual depreciation rate for the health stock, and $\zeta_h$ represents the rate at which cigarette consumption builds this stock.
4.2 The Health Process

We assume that individuals can enter into two kinds of bad health states: chronic, non-cancerous conditions related to the lungs, and all other conditions. Even though both sets of health events play important roles in influencing health behaviors and mortality, the medical literature suggests that smoking most directly affects the pulmonary system. Furthermore, we would like to explain the reduced form patterns that we observe between our SNPs of interest and lung health, so we treat such illness as a separate category. Let $B^S_{it} \in \{0, 1\}$ indicate that individual $i$ experiences a bad health state related to the lungs in period $t$, and let and $B^O_{it} \in \{0, 1\}$ indicate a bad health state related to other conditions. Since our data on lung illness indicate whether an individual has ever experienced a chronic lung condition, we model lung illness as an absorbing state, so $B^S_{it+1} = 1$ if $B^S_{it} = 1$. For an individual who has never been diagnosed with a chronic lung condition, we model the joint distribution of $B^S_{it}$ and $B^O_{it}$ through a bivariate probit specification. Let $b^S_{it}$ and $b^O_{it}$ be continuous indices reflecting an individual’s propensity to fall into the various bad-health states. We assume that:

\begin{align*}
    b^S_{it} &= \beta_0^S + \beta_{age}^S Age + \beta_{age^2}^S Age^2/100 + \beta_{S_{h}}^S S_{h_{it}} + \beta_{ageh}^S Age S_{h_{it}} + \epsilon^S_{it} \quad (4) \\
    b^O_{it} &= \beta_0^O + \beta_{age}^O Age + \beta_{age^2}^O Age^2/100 + \beta_{S_{h}}^O S_{h_{it}} + \beta_{ageh}^O Age S_{h_{it}} + \beta_{bo}^O B^O_{i;t-1} + \beta_{boa}^O Age B^O_{i;t-1} + \epsilon^O_{it} \quad (5)
\end{align*}

Here $\{\epsilon^S_{it}, \epsilon^O_{it}\} \sim N(0, \Sigma)$, $\Sigma$ is a variance-covariance matrix, and we allow $\sigma_{12} \neq 0$. Then $B^S_{it} = 1$ if $b^S_{it} > 0$, and $B^S_{it} = 0$. Similarly, $B^O_{it} = 1$ if $b^O_{it} > 0$, and $B^O_{it} = 0$. If $B^S_{it-1} = 1$, then the process determining $B^O_{it}$ collapses to the single-equation probit specified by Equation 5.

At the beginning of a period, before $B^S_{it}$ and $B^O_{it}$ are determined, an individual dies with probability:

$$
\pi^D_{it} = \Phi(\beta^d X^d_{it})
$$

Here $X^d_{it}$ is a vector of regressors that includes a vector of age controls, $B^S_{it-1}, B^O_{it-1}$, and $S^h_{it}$. Let $\pi^S_{it} = (1 - \pi^D_{it})$ represent the survival probability.
4.3 Period Utility

Let $Z_{it}$ refer to the set of state variables for individual $i$’s decision problem in period $t$, excluding transitory shocks to utility. This vector contains the addiction and smoking health stocks, as well as the current health states, last period’s smoking choice, income in the current period ($y_{it}$), an indicator for whether or not the individual has ever smoked ($ES_{it}$), and a reduction cost draw ($\alpha_{3it}$): $Z_{it} = \{S_{it}^a, S_{it}^h, B_{it}^O, B_{it}^S, C_{it-1}, ES_{it}, y_{it}, \alpha_{3it}, t\}$. The period utility associated with choosing option $j$ from the choice set is given by $\bar{u}_j(Z_{it}, e_{jit}) = u_j(Z_{it}) + \epsilon_{jit}$. That is, period utility for choice $j$ is the sum of a component that depends on the state variables $u_j$, and a random shock, $\epsilon_{jit}$, with the state-dependent component specified as:

$$
\begin{align*}
    u_j(Z_{it}) &= \alpha_0 \ln(1 + c_j) + \alpha_1 \ln(y_{it} - e(\tau_{it}, c_j)) \\
    &\quad + [\alpha_5 + \alpha_6 Age_{it} + \alpha_7 Age_{it}^2] 1\{c_j > 0 \text{ and } PS_{it} = 0\} \\
    &\quad + [\alpha_{3it} + \alpha_4 S_{it}^a] 1\{c_j < C_{it-1}\}
\end{align*}
$$

(7)

Here the $\alpha_0 \ln(1 + c_j)$ term represents the part of period utility that an individual receives from smoking at level $c_j$, and $\alpha_1 \ln(y_{it} - e(\tau_{it}, c_j))$ represents utility from all other consumption goods besides cigarettes. Income in period $t$ is given by $y_{it}$, and $e(\tau_{it}, c_j)$ represents expenditure on cigarettes, which depends on both the level of cigarette consumption, $c_j$, as well as $pc_{it}$, the cigarette price that individual $i$ faces in period $t$. The term $[\alpha_{3it} + \alpha_4 S_{it}^a] 1\{c_j < C_{it-1}\}$ captures the disutility that an individual receives from reducing their cigarette consumption between two periods. Notice that this is allowed to vary with the addiction stock, $S_{it}^a$. Individuals with a more intense addiction may experience greater costs of quitting or reducing cigarette consumption. The reduction cost also depends on a stochastic component $\alpha_{3it}$. Finally, the $\epsilon_{jit}$ terms are shocks to each choice that are assumed to be i.i.d. across individuals, choices, and time periods and are drawn from a Type I extreme value distribution.

4.4 Stochastic Reduction Costs and Learning

The stochastic component of reduction costs, $\alpha_{3it}$, is assumed to be drawn from a discrete distribution with two points of support, $\{\alpha_3, \overline{\alpha_3}\}$, with associated probabilities $P_{\alpha_3} = \{\phi_i, (1-\phi_i)\}$. 

15
That is, each period, any individual can either receive a low cost draw, $\alpha_3$, or a high cost draw $\alpha_3$. We have assumed that the probabilities $P_{\alpha_3}^i$ are heterogeneous in the population, so that some people are more likely than others to draw a low cost each period. We place additional structure on this heterogeneity by assuming that individuals can be of two cost types: easy and hard. That is, $P_{\alpha_3}^i$ takes one of two values: $P_{\alpha_3}^{\text{easy}}$, or $P_{\alpha_3}^{\text{hard}}$, with $\phi^{\text{easy}} < \phi^{\text{hard}}$. Let $\pi^{\text{easy}}$ refer to the probability that an individual is an easy-cost type.

We assume that individuals are initially unaware of their exact cost type, and only come to learn the true value of $P_{\alpha_3}^i$ through experiential learning. Let $B_{\alpha_3}^{it}$ refer to individual $i$'s belief in time period $t$ about the cost draw probabilities. If an individual is informed about their own type, then $B_{\alpha_3}^{it} = P_{\alpha_3}^i$. However, if an individual is uninformed, then $B_{\alpha_3}^{it} = \pi^{\text{easy}} P_{\alpha_3}^{\text{easy}} + (1 - \pi^{\text{easy}}) P_{\alpha_3}^{\text{hard}}$. An individual who is not informed about their own type assumes that their probability of receiving the low-cost draw is equivalent to the average probability that a smoker receives the low-cost draw in the population. Let $\text{Info}_{it}$ be a dummy variable that indicates whether or not an individual is informed in time period $t$. We assume that individuals start out life uninformed ($\text{Info}_{i0} = 0$). However, after every period in which an individual smokes, there is some probability $\pi^{\text{learn}}$ that an individual learns their true type.

We have proposed a rather crude learning mechanism. A rational, Bayesian agent would update their belief $B_{\alpha_3}^{it}$ on the basis of the observed sequence of past cost draws $\{\alpha_{3is}\}_{s=0}^{t-1}$. However Bayesian learning dynamics would complicate their numerical solution of the model by necessitating an additional state variable - the history of past cost draws. To avoid excessive computational burden, we make the starker assumption that individuals randomly transition from the uniformed to the informed state.

### 4.5 Decision Problem

The individual’s decision problem can be expressed as:

$$\max_{j \in 0, 1, 2, 3} \left( u_j(Z_{it}, \epsilon_{jit}) + \beta \pi^{S}_{it+1}(Z_{it}, S_{it+1}^h) E \left[ V_{t+1}(Z_{it+1}) \mid B_{\alpha_3}^{it} \right] \right)$$

(8)

Here we recognize that the probability of dying between periods $t$ and $t+1$ depends on the period $t$ state variables, $Z_{it}$, as well as the updated smoking health stock $S_{it+1}^h$. Also, $V_{t+1}(Z_{it+1})$ is the
value of the decision problem in period \( t + 1 \) given the state vector \( Z_{it+1} \). The expectation of \( V_{t+1}(Z_{it+1}) \) is taken with respect to the random state vector \( Z_{it+1} \) and the vector of shocks \( \epsilon_{it} \) conditional on survival, \( Z_{it} \), and choice \( j \). Note also that the expectation depends on the current set of beliefs about the reduction cost distribution, \( B_{it}^{it} \).

Following Rust (1987), if we assume that the shocks \( \epsilon_{it}^j \), are additively separable, satisfy the conditional independence assumption, and follow a Type I extreme value distribution, then the the expected value of the next period’s value function (conditional on survival) can be expressed as:

\[
E \left[ V_{t+1}(Z_{it+1}) \mid B_{it}^{it} \right] = E \left[ \ln \left( \sum_j \exp \left\{ \nu_{jt+1}(Z_{it+1}, B_{it}^{it}) \right\} \right) \right] \tag{9}
\]

Here \( \nu_j(Z_{it+1}, B_{it}^{it}) \) is the conditional value function associated with making choice \( j \) in time period \( t \). This is the expected value of making the choice, net of the \( \epsilon_{jt} \) shock:

\[
\nu_{jt}(Z_{it}, B_{it}^{it}) = u_j(Z_{it}) + \beta x_{it+1}^S(Z_{it}, S_{it+1}^h) E \left[ \ln \left( \sum_j \exp \left\{ \nu_{jt+1}(Z_{it+1}, B_{it}^{it}) \right\} \right) \right] \tag{10}
\]

In the terminal period \( t = T \), the conditional value functions reduce down to \( \nu_{jT}(Z_{iT}) = u_j(Z_{iT}, B_{iT}^{iT}) \).

The individual’s decision problem can thus be expressed as:

\[
\max_{j \in h, 1, 2, 3} \nu_j(Z_{it}, B_{it}^{it}) + \epsilon_{jt} \tag{11}
\]

The conditional choice probabilities associated with this optimization problem can be expressed as:

\[
Prob(j = j' \mid Z_{it}, B_{it}^{it}) = \frac{\exp(\nu_{j't}(Z_{it}, B_{it}^{it}))}{\sum_j \exp(\nu_{jt}(Z_{it}, B_{it}^{it}))} \tag{12}
\]

### 4.6 Genes and Parameter Heterogeneity

Thus far, we have not introduced any heterogeneity in the model parameters, specified the role of genes, or described our assumptions about the information that individuals possess about their genotype. Let \( G_i \) refer to individual \( i \)'s genotype. We allow for gene-induced heterogeneity in
two places in the model: 1) An individual’s taste for cigarettes (the $\alpha_0$ parameter), and 2) The probability that an individual is an easy cost type, $\pi^{easy}$. Specifically, we allow these parameters to vary with the number of copies of the minor alleles associated with rs16969968 and rs4950:

$$\alpha_0(G_i) = \alpha_{01} + \alpha_{02}rs_{169i} + \alpha_{03}rs_{4950i}$$

$$\pi^{easy}(G_i) = \pi_0^{easy} + \pi_{1}^{easy}rs_{169i} + \pi_{2}^{easy}rs_{4950i}$$
Table 1: Cross Sectional Characteristics in HRS

<table>
<thead>
<tr>
<th>Variable</th>
<th>Mean</th>
<th>Std. Dev.</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Age</td>
<td>73.73</td>
<td>7.63</td>
<td>8122</td>
</tr>
<tr>
<td>Male</td>
<td>0.43</td>
<td>0.50</td>
<td>8122</td>
</tr>
<tr>
<td>Ever Smoked</td>
<td>0.57</td>
<td>0.49</td>
<td>8042</td>
</tr>
<tr>
<td>Max. Cigs Per Day</td>
<td>25.67</td>
<td>17.90</td>
<td>4592</td>
</tr>
<tr>
<td>Ever Lung Illness</td>
<td>0.18</td>
<td>0.38</td>
<td>8122</td>
</tr>
<tr>
<td>Ever Heart Illness</td>
<td>0.38</td>
<td>0.48</td>
<td>8122</td>
</tr>
<tr>
<td>Ever Cancer</td>
<td>0.23</td>
<td>0.42</td>
<td>8122</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>SNP</th>
<th>Minor Allele Frequency</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs16969968</td>
<td>0  0.45  0.44  0.11</td>
</tr>
<tr>
<td>rs4950</td>
<td>0  0.60  0.35  0.05</td>
</tr>
<tr>
<td>rs680244</td>
<td>0  0.35  0.48  0.18</td>
</tr>
</tbody>
</table>

Notes: Cross-sectional statistics are calculated based on an individual’s last appearance in the HRS. That is, each observation represents the last year in which the individual was alive and responded to an HRS survey. This sample consists of all individuals who are caucasian, born between 1920-1949, and for whom we have non-missing genetic data.

Table 2: SNPs and Smoking Behavior (Smokers)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Max. Cigs. Per Day at Peak (1)</th>
<th>Current Smoke (4)</th>
<th>Quitting (2-year) (6)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(2)</td>
<td>(3)</td>
<td>(5)</td>
</tr>
<tr>
<td></td>
<td>(2-year)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs16969968</td>
<td>1.316*** (0.390)</td>
<td>1.381*** (0.484)</td>
<td>0.000 (0.010)</td>
</tr>
<tr>
<td>rs4950</td>
<td>-0.721* (0.434)</td>
<td>-0.713 (0.434)</td>
<td>-0.023** (0.009)</td>
</tr>
<tr>
<td>rs680244</td>
<td>-0.655* (0.368)</td>
<td>0.108 (0.457)</td>
<td>0.004 (0.010)</td>
</tr>
</tbody>
</table>

| Obs.    | 4592                            | 4592                | 4592                   | 4621 | 849 |
| R²      | 0.097                           | 0.095               | 0.097                  | 0.067| 0.098 |

19
### Table 3: SNPs and Lung Health (Smokers)

<table>
<thead>
<tr>
<th>rs16969968</th>
<th>Lung Illness (1)</th>
<th>Lung Illness (2)</th>
<th>Lung Illness (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.034***</td>
<td>0.029**</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td></td>
</tr>
<tr>
<td>rs4950</td>
<td>0.005</td>
<td>0.006</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.011)</td>
<td></td>
</tr>
<tr>
<td>rs680244</td>
<td>0.009</td>
<td>0.009</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.011)</td>
<td>(0.011)</td>
<td></td>
</tr>
<tr>
<td>MaxCigsPerDay</td>
<td>0.003***</td>
<td>0.003***</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td></td>
</tr>
<tr>
<td>Obs.</td>
<td>4592</td>
<td>4622</td>
<td>4592</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.044</td>
<td>0.027</td>
<td>0.046</td>
</tr>
</tbody>
</table>

### Table 4: SNPs, Heart Illness, and Cancer (Smokers)

<table>
<thead>
<tr>
<th>rs16969968</th>
<th>Heart Illness (4)</th>
<th>Heart Illness (5)</th>
<th>Heart Illness (6)</th>
<th>Cancer (7)</th>
<th>Cancer (8)</th>
<th>Cancer (9)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>-0.016</td>
<td>-0.018</td>
<td>-0.005</td>
<td>-0.006</td>
<td>(0.012)</td>
<td>(0.012)</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.014)</td>
<td>(0.012)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs4950</td>
<td>0.021*</td>
<td>0.022*</td>
<td>0.013</td>
<td>0.015</td>
<td>(0.011)</td>
<td>(0.011)</td>
</tr>
<tr>
<td></td>
<td>(0.012)</td>
<td>(0.012)</td>
<td>(0.011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rs680244</td>
<td>-0.000</td>
<td>0.001</td>
<td>-0.013</td>
<td>-0.011</td>
<td>(0.011)</td>
<td>(0.011)</td>
</tr>
<tr>
<td></td>
<td>(0.013)</td>
<td>(0.013)</td>
<td>(0.011)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Max Cigs. Per Day</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>0.002***</td>
<td>(0.000)</td>
<td>(0.000)</td>
</tr>
<tr>
<td></td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td>(0.000)</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Obs.</td>
<td>4592</td>
<td>4622</td>
<td>4592</td>
<td>4592</td>
<td>4622</td>
<td>4592</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.070</td>
<td>0.068</td>
<td>0.072</td>
<td>0.045</td>
<td>0.041</td>
<td>0.045</td>
</tr>
</tbody>
</table>

### Table 5: SNPs and Mortality (Smokers)

<table>
<thead>
<tr>
<th>rs16969968</th>
<th>Mortality (1)</th>
<th>Mortality (2)</th>
<th>Mortality (3)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>0.004***</td>
<td>0.004</td>
<td>0.005**</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>rs4950</td>
<td>-0.003*</td>
<td>-0.002</td>
<td>-0.005***</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>rs680244</td>
<td>0.003</td>
<td>0.004</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Obs.</td>
<td>17525</td>
<td>11046</td>
<td>6479</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.018</td>
<td>0.015</td>
<td>0.009</td>
</tr>
</tbody>
</table>
### Table 6: SNPs and Health (Non-Smokers)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Lung Illness</th>
<th>Heart Illness</th>
<th>Cancer</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs16969968</td>
<td>0.008</td>
<td>0.007</td>
<td>-0.002</td>
</tr>
<tr>
<td></td>
<td>(0.010)</td>
<td>(0.015)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>rs4950</td>
<td>0.013</td>
<td>0.008</td>
<td>-0.009</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.013)</td>
<td>(0.012)</td>
</tr>
<tr>
<td>rs680244</td>
<td>-0.004</td>
<td>0.008</td>
<td>-0.016</td>
</tr>
<tr>
<td></td>
<td>(0.009)</td>
<td>(0.014)</td>
<td>(0.013)</td>
</tr>
<tr>
<td>Obs.</td>
<td>3420</td>
<td>3420</td>
<td>3420</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.031</td>
<td>0.064</td>
<td>0.040</td>
</tr>
</tbody>
</table>

### Table 7: SNPs and Mortality (Non-Smokers)

<table>
<thead>
<tr>
<th>SNP</th>
<th>Born 1920-1949</th>
<th>Born 1920-1939</th>
<th>Birth 1940-1949</th>
</tr>
</thead>
<tbody>
<tr>
<td>rs16969968</td>
<td>0.001</td>
<td>0.001</td>
<td>0.000</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>rs4950</td>
<td>0.000</td>
<td>0.001</td>
<td>-0.000</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.002)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>rs680244</td>
<td>0.002</td>
<td>0.002</td>
<td>0.001</td>
</tr>
<tr>
<td></td>
<td>(0.002)</td>
<td>(0.003)</td>
<td>(0.002)</td>
</tr>
<tr>
<td>Obs.</td>
<td>13071</td>
<td>8355</td>
<td>4716</td>
</tr>
<tr>
<td>$R^2$</td>
<td>0.026</td>
<td>0.026</td>
<td>0.013</td>
</tr>
</tbody>
</table>
5 Appendix

We pool all waves of the HRS (the primary HRS surveys which occur biennially from 1992-2010), along with the 1993 and 1995 AHEAD surveys. We also merge the HRS Cross-Wave Tracker file, which includes data on mortality events, as well as variables from RAND’s cleaned HRS sample. Merging all of these files, we start with a sample of 37,239 unique individuals. We construct our final sample by making the following sample restrictions sequentially:

- Drop 1 individual who is only in the RAND data (produces sample of 37,239 individuals).
- Drop 118 individuals with missing year of birth data (produces sample of 37,121 individuals)
- Drop 6,297 individuals with missing gender data (produces sample of 30,824 individuals)
- Drop 63 individuals with inconsistent smoking data. That is, individuals who report never smoking, but report smoking positive quantities of cigarettes when asked about smoking intensity. (produces sample of 30,761 individuals)
- Drop 18,659 individuals with missing genetic data (produces sample of 12,102 individuals)
- Keep only individuals born between 1920 and 1950 (produces sample of 9,592 individuals).

These sample restrictions leave us with a final sample of 80,360 person-year observations on 9,592 unique individuals

5.1 Moments used in Estimation

A total of 73 moments are used in the estimation, which we itemize below. Unless otherwise noted, we condition on age by considering means across the following seven age groups: (55-59,60-64,65-69,70-74,75-79,80-84,85-89);

- **Initiation Moments**: average age at start, fraction of starts occurring before age 15, and the fraction of starts occurring after age 30. (3 moments)

- **Smoking Extensive Margin**: Fraction of individuals who are smoking by age group. (7 more to 10 moments)
• **Smoking Intensive Margin**: Fraction of smokers choosing categories 1 and 3, by age group (14 more to 24)

• **Quitting**: Fraction of smokers quitting two years later, by age group (7 more to 31)

• **Death Rates**: Fraction of individuals that die each year, by age group. These are also computed conditional on current (binary) smoking status. (14 more to 45 moments)

• **Lung Illness**: Fraction of individuals that have ever reported a major lung illness, by age group. These are also computed conditional on current (binary) smoking status. (14 more to 59 Moments)

• **Bad Health Status**: Fraction of individuals that are currently in bad health, by age group. These are also computed conditional on current (binary) smoking status (14 more to 73 moments)

### 5.2 Identification

Here we sketch the intuition behind how the model parameters are identified from the selected moments used in the estimation. We begin with the parameters related to period utility:

\[
\begin{align*}
    u_j(Z_{it}) &= \alpha_0 \ln(1 + c_j) + \alpha_1 \ln(y_t - \varepsilon(\tau_{it}, c_j)) \\
    &+ [\alpha_3 + \alpha_4 S^a_{it}] 1\{c_j < C_{it-1}\} \\
    &+ [\alpha_5 + \alpha_6 Age_{it} + \alpha_7 Age_{it}^2] 1\{c_j > 0 \text{ and } PS_{it} = 0\}
\end{align*}
\]

(13)

The cross-sectional fraction of individuals choosing different smoking intensities pins down the coefficient on the log of cigarettes, \(\alpha_0\). Currently nothing in the model pins down the marginal utility of non-smoking consumption, \(\alpha_1\). However, this can be pinned down either by smoking intensities conditional on income, or by looking at cross-sectional variation in smoking intensities by state (cigarette tax rates differ across states). The reduction cost parameters \(\alpha_3\) and \(\alpha_4\) are identified by the quitting moments, as well as the changing distribution of intensities across age.
The effect of the addiction stock on reduction costs could be more sharply identified by considering quitting moments conditional on past smoking behavior (here we could use maximum cigarettes per day, or look at average age at quitting for those individuals who are not smoking at age 50). The initiation parameters \( \alpha_5, \alpha_6, \alpha_7 \), are pinned down by the three initiation parameters.

Next, turn to the health and mortality parameters:

\[
\begin{align*}
 b_{it}^{S_t} &= \beta_0^S + \beta_{age}^S Age + \beta_{age2}^S Age^2/100 + \beta_h^S S_{it}^h + \beta_{ageh}^S Age S_{it}^h + \epsilon_{it}^S \\
 b_{it}^{O_t} &= \beta_0^O + \beta_{age}^O Age + \beta_{age2}^O Age^2/100 + \beta_h^O S_{it}^h + \beta_{ageh}^O Age S_{it}^h + \beta_{bo}^O B_{i,t-1}^O + \epsilon_{it}^O \\
 \pi_{it}^D &= \Phi(b^d X_{it}^d) 
\end{align*}
\]

Here \( X_{it}^d \) is a vector of regressors that includes a vector of age controls, \( B_{i,t-1}^S, B_{i,t-1}^O, \) and \( S_{it}^h \). Let \( \pi_{it}^S = (1 - \pi_{it}^D) \) represent the survival probability.

Here the constant and the age coefficients in the S disease index (\( \beta_0^S, \beta_{age}^S, \) and \( \beta_{age2}^S \)) are identified from the unconditional age profile in lung illness. The coefficients on the smoking health stock and the interaction of the smoking health stock with age (\( \beta_h^S, \) and \( \beta_{ageh}^S \)) are pinned down by the age profile in lung illness among smokers. Lung illness conditional on past intensity or on longer lags (including max cigarettes per day) would provide more identifying information for these parameters. Similarly, the age profile in the O-illness index is identified from the bad health moments conditional on age, and the coefficients on the health stock are identified by bad health status conditional on age. The persistence parameter \( \beta_{bo}^0 \) would be most clearly identified by bad health rates conditional on current bad health (need to add that in). The death parameters are identified in a similar way, but I need to add in death conditional on past bad health and lung-illness.

Next, we turn to the addiction stock parameters:

\[
\begin{align*}
 S_{it}^a &= [1 - (\delta_{a1} + \delta_{a2} C_{it})] S_{it}^a + \zeta_a C_{it} \\
 S_{it}^h &= (1 - \delta_h) S_{it}^h + \zeta_h C_{it}
\end{align*}
\]
The parameters related to the evolution of the addiction habit stock are pinned down by persistence in smoking behavior. The relationship between maximum cigarettes per day and current smoking would provide one set of possible moments that would identify \( \delta_{a1} \). The cold-turkey parameter \( \delta_{a2} \) be be directly identified off of the fraction of quits that are cold-turkey (coming from higher smoking intensities). The depreciation of the health stock parameter \( \delta_{h} \) is identified off of the differential influence of recent smoking v.s. past smoking. Health and mortality moments conditional on ever smoking or on maximum cigarettes per day can pin this down.

Finally, we are allowing for heterogeneity in the taste for smoking, \( \alpha_0 \). Working on this ... quits by age are one way to pin this down. This is the standard argument that negative duration dependence is a function of the degree of heterogeneity in hazard rates.

References


