On the Identification of the Effect of Smoking on Mortality

Jérôme Adda*and Valérie Lechene†

January 29, 2004

Abstract

This paper considers the identification of the effect of tobacco on mortality. If individuals select into smoking according to some unobserved health characteristic, then estimates of the effect of tobacco on health that do not account for this are biased. We show that using information on mortality, morbidity and smoking, it is possible to control for this selection effect and obtain consistent estimates of the effect of smoking on mortality. We implement our method on Swedish data. We show that there is selection into smoking, and considerable dispersion around the average effect, so that health policies that aim at decreasing smoking prevalence and quantities smoked might have less effect in terms of average number of years of life gained than previously estimated. We also empirically show that selection into smoking has increased over the last fifty years with the availability of information on the dangers of smoking, so that future studies comparing smokers and non smokers will spuriously reveal a worsening effect of tobacco on health if they fail to control for selection.

JEL number: I12

Keywords: Health, Duration, Smoking, Selection, Mortality, Life Expectancy, Causality

*University College London and IFS. email: j.adda@ucl.ac.uk

†Wadham College, University of Oxford and IFS. email: valerie.lechene@economics.ox.ac.uk. We are grateful for comments by Orazio Attanasio, Alan Beggs, Andrew Chesher, Christian Dustmann, John Flemming, Pierre-Yves Geoffard, Michael Grossman, Hide Ichimura, Michael Marmot, Paul Schultz and to seminar participants at CEPR health workshop, CEMFI, CEU, DELTA, ESPE, Minneapolis Fed, SED, Tinbergen Institute, iHEA, ESEM, UCL and Copenhagen.

Contents

1 Introduction **3**

2 A review of different methodological approaches to the identification of the effect of tobacco on health **6**
2.1 A brief history of the tobacco-health controversy 6
2.2 Medical approach 10
2.3 Epidemiologic approach 11
2.4 Economic approach 13
2.5 Standard approaches to measurement in the presence of endogeneity . 15

3 Identification using medical and epidemiological information to construct a proxy for tobacco-free morbidity **18**

4 The Data **20**
4.1 Morbidity 22
4.2 Tobacco-Free Morbidity Scores 23
4.3 Other Risky Behaviors 25
4.4 Comparison with Previous Studies 26

5 Selection into smoking **26**
5.1 Selection into smoking status 27
5.2 Selection into smoking intensity 32
5.3 Selection into smoking duration 33
5.4 Further Robustness Check: Changes in tobacco-free morbidity 34

6 The effect of tobacco on mortality **35**

7 Compensating Foregone Earnings due to Tobacco Related Deaths **39**

8 Conclusion **42**

A Appendix **44**
A.1 Excerpt from "Cautions Against the Immoderate Use of Snuff and the Effects It Must Produce When This Way Taken into the Body", John Hill, 1761. 44
A.2 Excerpt on Causality from the 1964 Report to the Surgeon General on Tobacco and Health 44

1 Introduction

Smokers die on average younger than non smokers. This statement, usually substantiated by contrasting the life expectancy of smokers and that of non smokers adjusting for some individual characteristics, has formed, since the 1960s, the basis of government policies designed to curb smoking on the grounds of the detrimental effect of smoking on health. But the effect of tobacco on health can be inferred by comparing smokers and non smokers only if smoking is a random choice so that individuals do not self select into smoking on the basis of their health.

Although there is a large epidemiologic literature devoted to measuring the effect of tobacco on health[1], it shows limited concern for this question of selection into smoking. This is probably because epidemiologists consider smoking as exogenous. A different view point has been adopted by the economic literature, which considers smoking as a choice, following the health investment literature, pioneered by Grossman (1972) or the rational addiction literature (Becker and Murphy (1988)). In this view, smoking may depend on unobserved individual characteristics - such as the discount factor- that also influence health through other channels. If this is the case, then traditional epidemiological studies may yield misleading results as to the effect of tobacco on health. Rather surprisingly, the economic literature has not pushed this point much further, with the notable exceptions of Rosenzweig and Schultz (1983) and Evans and Ringel (1999), who consider health outcomes conditional on smoking, allowing for endogeneity. Both show that endogeneity is important and should be accounted for.

[1] Studies dates back at least to the nineteen twenties (Broders (1920), Lombard and Doering (1928) or Lickint (1935)). The seminal papers in the nineteen fifties and sixties, include the work of Doll and Hill (1950) and Wynder and Graham (1950) (see also Doll and Hill (1954), Doll and Hill (1956) and Hammond (1966). More recent estimates of the effect of tobacco on health include for instance Phillips et al. (1996) or Peto et al. (2000).

If there is selection into smoking so that individuals with poorer health are more likely to be smokers - which would lead to overstate the effect of tobacco on health - it would have enormous economic consequences. Public policies, even if successful in decreasing smoking prevalence, may not achieve large gains in life expectancy or large economies in health expenditure. Smokers who quit or reduce their consumption of cigarettes would certainly face a lower risk of tobacco related diseases, such as lung cancers, but would not necessarily see their life expectancy increase by large margins. This would also bear on legal compensations, which may have to be revised downwards, as part of the observed difference in average life expectancies of smokers and non smokers could be due to selection.

According to the World Health Organization, there are currently 1.25 billion smokers in the world; among those, there are each year 4 million deaths from tobacco-related diseases and it is forecast that there will be 10 million such deaths yearly by 2030; in other words tobacco causes more deaths than malaria, tuberculosis and major childhood conditions combined. A crucial policy question is whether preventing all these deaths will lead to substantial gains in life expectancy. To quantify the costs due to the anticipated death of smokers, it is necessary to measure the effect of tobacco on health, a task at which we have made here an attempt.

In this paper, we consider the identification of the effect of tobacco on health allowing for the possibility of selection into smoking. While most of the literature has used data on mortality and smoking behavior, we argue that identification is usually impossible using such data alone. Our approach is to combine data on mortality and smoking behaviour with detailed information on individual morbidity, together with medical and epidemiological knowledge on morbidity. Our identification strategy is novel but very simple. We use the additional information from the medical and epidemiological sciences to construct a proxy for the underlying health of the individual, where the underlying health is defined as being their health had they not smoked.. This is very similar in spirit to using test scores to proxy for ability in wage equations.

We use an extensive data set, where 38000 Swedish individuals are followed for up to eighteen years, recording their smoking behavior, other risky behaviors, mortality, a range of morbidity indicators and information on individual and family characteristics such as education, occupation and family income.

We present evidence of selection into smoking. We show that smokers come from a population with poorer underlying health, even when conditioning on a number of observed characteristics. We also show that the effect of tobacco is lower for individuals with poorer underlying health (and hence with lower life expectancy as a non smoker) than for individuals with better underlying health. This implies that the gains from reducing smoking are not as large as they would be thought to be without accounting for selection into smoking.

We also show that there is a strong cohort effect. The selection effect is important for the cohorts who started smoking when the information on the effect of tobacco on health was widely publicized, but not so much for previous cohorts. This suggests that the results obtained in the past by epidemiological studies are not far off the mark for the generations considered but that future studies comparing smokers and non smokers will spuriously reveal a worsening effect of tobacco on health if they fail to control for selection.

Finally, combining data on tobacco-free health, survival and income, we evaluate the amount of foregone earnings from tobacco related death. We evaluate these amounts for different groups of individuals, according to their sex, education level and underlying health. We show that foregone earnings depend crucially on the underlying health of the individual

We begin in Section 2 with a brief history of the tobacco health controversy. The remainder of the section presents a methodological review of the medical, epidemiologic and economic approaches to the identification of the effect of tobacco on health and mortality. Section 3 presents our strategy for the identification of selection into smoking and the measurement of the effect of tobacco on mortality, using a proxy for the individual's underlying health. Section 4 presents the data and

discusses the construction of the tobacco-free morbidity indices we use as proxy for underlying health. Section 5 presents evidence of selection into smoking. Section 6 presents estimates of the effect of tobacco on mortality controlling for selection. Section 7 evaluates the compensation for foregone earnings after a tobacco related death. Section 8 concludes.

2 A review of different methodological approaches to the identification of the effect of tobacco on health

The question of the effect of tobacco on health has generated passion since the introduction of tobacco to European societies at the end of the 15th century. Scientific contributions to this debate have come from the fields of medicine, epidemiology and economics. Before presenting their respective methodological contributions, we start this section by a historical review of the tobacco-health controversy, whose interest, if any, resides in the account of the evolution of the perception of the effect of tobacco on health since the 15th century, from beneficial to detrimental. It is also of interest to note that, in parallel to this evolution, the understanding that selection into smoking might cause endogeneity problems in the measurement and interpretation of the observed correlation was established as early as 1761. We then turn to the presentation of recent contributions of medicine, epidemiology and economics to the measurement of the effect of tobacco on health. Finally, we discuss various methods for controlling for the endogeneity of smoking in mortality.

2.1 A brief history of the tobacco-health controversy

From the moment of its introduction to the Old World in 1492 up to this day, tobacco has been thought to have effects on health, either beneficial or detrimental. From the end of the 15th century until the mid 17th century, tobacco is presented as a panacea, having curative properties against headaches, fever, rheumatism, nau-

sea, skin complaints. It is also supposed to be useful in helping childbirth, and remains thought to have properties against respiratory illnesses until the early 20th century.[2]

It is difficult to find evidence about what led to these surprising views to be held about tobacco; views that are all the more surprising that evidence against tobacco started to accumulate very early on. In 1670, the pioneering Dutch anatomist Kerckring describes his autopsies of heavy smokers: "The tongue of the cadaver is black and gives off an odor of poison; the trachea is coated with soot, like a cooking pot; the lungs are dried-out and almost friable. The corpse gives the overall impression that someone had lit a fire among the organs." A century later, the English physician John Hill publishes what is probably the first clinical study of the effects of tobacco on health, "Cautions against the Immoderate Use of Snuff", (Hill (1761)). Hill observes that the consumption of snuff appears to be associated with cancers of the nose. However, rather remarkably, he also warns that the observed association might also arise in the absence of a causal effect (see the appendix for an excerpt from the original text). Similar studies are conducted in other countries throughout the eighteenth and nineteenth centuries, but without causing any major stir. However, from the 1920s onwards, attitudes in the scientific community start to change, as epidemiological studies accumulate which show the existence of a statistical correlation between cancer and smoking.

In an influential article published in Science, Pearl (1938) shows that smokers do not live as long as non-smokers. By 1944, the American Cancer Society begins to warn about possible ill effects of smoking, although it admits that "no definite evidence exists" linking smoking and lung cancer. In a 1952 article in the Reader's Digest, Norr (1952) brings awareness to the larger public of the medical and epidemiological research on tobacco. Coincidentally or not, after this publication, consumption starts to fall for the first time in history, and governments are called

[2] During the 1665 Great Plague, smoking was made compulsory at Eton in a doomed effort to ward off infection. Failure to smoke resulted in a whipping. ("The diary of Samuel Pepys" (1633-1703)).

upon by public opinion to provide the means of establishing whether the allegations about the dangers of cigarettes smoking are true.

In the nineteen fifties, many governments appoint committees to review the evidence on smoking and health[3]. This leads to the publication in the US of the 1964 Report to the Surgeon General on Smoking and Health which concludes that, regarding age specific mortality, although there is evidence of a statistical association, "the total number of excess deaths causally related to cigarette smoking in the U.S. population cannot be accurately estimated. In view of the continuing and mounting evidence from many sources, it is the judgment of the Committee that cigarette smoking contributes substantially to mortality from certain specific diseases and to the overall death rate." (Chapter 4, p 31). The studies on which the Committee based its conclusions found that the excess death rate[4] is increasing in the quantity of cigarettes smoked; is highest at the earlier ages (40-50) and declines thereafter. However, later in the report, the strength of the causality is further qualified as it is concluded regarding smoking that: "This does not rule out physiological factors, especially in respect to habituation, nor the existence of predisposing constitutional or hereditary factors" (Chapter 14, p. 377). But whilst allowing for the possibility of selection into smoking, the Committee is not in a position to conclude as to its scale. Indeed, "the available evidence suggests the existence of some morphological differences between smokers and non-smokers, but is too meager to permit a conclusion" (Chapter 15, p. 387). In the UK, the 1962 Report to the Royal College of Physicians reached similar conclusions regarding excess death of smokers.

[3] The first task of these committees was to determine what is the nature of evidence which can lead to a diagnosis of causality. The discussion of the 1964 Report on Smoking and Health to the Surgeon General on this topic is methodologically still largely in force. We reproduce excerpts from the Report's discussion on this in the Appendix.

[4] The excess death rate is the difference between actual and expected number of deaths of smokers, where the latter is constructed from the number of deaths of non smokers, for a specific cause of death, and sometimes adjusted for age, gender and the quantities of cigarettes smoked.

The US, UK and other similar reports have formed the basis for the public health policies against smoking which have been put in place since the mid nineteen sixties. From our point of view, these reports are particularly interesting as they do not rule out the possibility of intrinsic differences between smokers and non smokers; intrinsic differences which may lead to selection into smoking and explain part or all of the observed correlation between smoking and ill health.

After the publication of these reports, the tobacco health controversy develops into its current form, with four types of players: the tobacco industry, health associations, government agencies and scientists from the fields of statistics, medicine and epidemiology. The interests of the first two groups are clear. The tobacco industry deploys numerous strategies to avoid its wares being banned, whilst the health associations lobby governments for regulation of the tobacco industry and the implementation of legal and other incentives aimed at reducing or outright banning of tobacco consumption. Governments have to satisfy various objectives, which we will not discuss here. Finally, the scientific community continues to refine the methods used to explore the links between smoking, mortality and morbidity.

Among the scientific community, a debate concerning the interpretation of the correlation between tobacco and mortality has involved eminent scientists. The most prominent opponent of the causal interpretation of the statistical association observed between tobacco and mortality was the statistician R.A. Fisher (1957a and b, 1958a, b and c), who argued that however strong or many they are, measures of statistical association cannot, by their nature and on their own, support a causal interpretation. Fisher contested the leap in interpretation from a correlation to a causality and called for further investigation. He argued that the correlation between mortality, morbidity and smoking could be due to causality in either direction or to the influence of a third factor. Given all that has been written about Fisher's position, we feel it is important to stress that nowhere in his writings have we found a defense of the position that tobacco is not harmful to health, but rather a position of principle which led him to issue cautionary words against interpreting any correlation

as a causation. The argument of the proponents of the causal interpretation is essentially that if there is no causal effect, the statistical association would not be either so strong or so universal.

Our view is that the findings of the medical and epidemiological sciences concerning the causal effect of tobacco regarding a large number of diseases are uncontestable. However, it remains nonetheless crucial to investigate the question of selection and its consequences of the measurement of the effect of tobacco on health.

We now turn to the recent scientific contributions on the measurement of the effect of tobacco on health. We will argue that the existing medical and epidemiologic approaches to the identification of the effect of tobacco on health are based upon an extreme assumption, so that the estimates obtained under this assumption do not constitute good approximations of the effect of tobacco on health. Indeed, both the epidemiologic approach and the medical approach boil down to comparing outcomes for smokers and non smokers, conditionally on some characteristics, but do not explicitly allow for selection.

2.2 Medical approach

The medical approach is based on either clinical studies, autopsy studies or animal experiments. The method of both clinical studies and autopsies is to observe smokers and non smokers and compare either health outcomes in the case of clinical studies, or organs, cells and tissues in the case of autopsies. The conclusion from both type of studies is that damages to body functions, organs, cells and tissues occur more frequently and severely in smokers.

For obvious reasons, it is not possible to proceed to random experiments involving humans to assess the effect of smoking on health. However, animal experiments (on mice, rabbits, dogs and monkeys) have been conducted since the mid 1920s[5]. The method of these experiments is to expose animals to tobacco smoke and tars,

[5] A very famous series of animal experiments was conducted by Oscar Auerbach, in the 1970s, involving beagles (Auerbach and Garfinkel (1970)).

and to the various chemical compounds they contain, and to assess their effect by comparison with control groups, as well as perform clinical studies of the treated populations. These experiments lead to the conclusion that "several of the compounds of tobacco are carcinogenic, and that other substances in tobacco and smoke, though not carcinogenic themselves, promote cancer production or lower the threshold to a known carcinogen". (1964 Report to the Surgeon General)

The conclusion from medical studies is that causation is established for animals, and that the similarity in damage caused by tobacco in animals and damage found in humans makes a very strong case for causation in humans as well. From our understanding, this is the closest to causation that medical studies have reached at this point.

2.3 Epidemiologic approach

The epidemiologic approach, based on population studies, is direct and consists in estimating some measure of statistical association between smoking and either mortality or specific health problems, by comparing outcomes for smokers and non smokers. It is best described using the framework of binary models, which relate the probability of death (whether in general or from a specific cause) or the probability of developing a specific disease, to individual characteristics and smoking (see for instance Doll et al. (1994), Lam et al. (2001)):

$$Y_{it} = \begin{cases} 1 \\ 0 \end{cases} \qquad \text{if } X_{it}\beta + \alpha S_i + u_{it} \gtreqless 0 \tag{1}$$

where Y_{it} is a binary variable indicating whether individual i is alive at date t (alternatively, whether individual i has developed a certain disease at date t), X_{it} is a set of variables thought to influence the probability of death (or of developing the disease), S_i is a variable capturing smoking behavior and u_{it} is a random shock. Depending on the context, the smoking variable is a variable indicating whether the individual smokes (or has smoked) or the quantity of cigarettes smoked (in

total or habitually).[6] In a setup such as equation (1), the effect of tobacco on the health outcome of interest (e.g. mortality), α, is identified by comparing estimated probabilities for smokers and non smokers, conditional on X. This is only valid to the extent that smokers and non smokers are randomly drawn from the same population in terms of underlying risks. In other words, if smoking is exogenous, the effect of smoking on mortality, α, is identified by such a setting. But if mortality depends on a characteristic ε_i^*, unobserved by the researcher (but potentially known to the individual), such that smoking depends on ε_i^* ($cov(\varepsilon_i^*, S_i) \neq 0$), then u_{it} can be written as $\varepsilon_i^* + v_{it}$, and the effect of smoking on the outcome Y_{it} is not identified by equation (1). In the case where there are unobserved factors which influence both smoking and the health outcomes, we have a classic problem of endogeneity and it is no longer possible to consistently estimate the effect of tobacco on life expectancy without additional information.

Concerns that a limited number of controls in X_i might not capture all the characteristics which influence both mortality and smoking has led Sterling and Weinkam (1990) and Smith and Shipley (1991) to advocate for controlling for as many observable characteristics as possible. The study which controls for the largest set of individual characteristics is Thun et al. (2000) which controls for education level, race, marital status, diet and alcohol consumption using a large cohort study from the American Cancer Society prospective study. They find that the effect of smoking is decreased when conditioning on observed characteristics, although by a small amount. This and other similar studies show that smokers are more likely to die from tobacco related diseases than non smokers. What these studies do not address is the question of the net gain of not smoking in terms of overall life expectancy, especially if selection is present.

Other studies focus on life expectancy and relate the duration to death to individual characteristics (Phillips et al. (1996), for instance).

[6] Statistics commonly reported in epidemiology are the excess death of smokers and the mortality ratio (relative death rates of smokers and non smokers for specific diseases). Both statistics can be obtained with binary models.

Finally, a considerable number of epidemiological studies compare smokers and non smokers' outcomes for specific causes of death.[7] They have consistently associated smoking with a variety of diseases, including lung cancer, cancer of the larynx or chronic obstructive pulmonary diseases. These studies show that reducing tobacco consumption will greatly reduce death from *these* causes, which certainly indicates a causal effect of tobacco on health, but does not rule out selection. In other words, the question remains of whether smokers are more likely to have shorter life expectancy independently from smoking.

Table A summarizes some of this research and presents estimates of the impact of tobacco on overall life expectancy. The results indicate that the hazard for dying could be between two and five times higher for smokers as compared to non smokers, at all durations. Sterling and Weinkam (1990) and Smith and Shipley (1991) try to reduce confounding effects by controlling for occupation and find that the effect of smoking is then reduced. As we can see in Table 1, the range of controls is usually rather limited.

Recent epidemiologic research, also based on population studies, measures the association between genetic make-up and human personality (eg Munafò et al. (2003)). It is found that certain genetic make-ups may be associated with personality traits, which in turn may influence life-style choices such as smoking. We will come back below in section 3 on how these arguments fit with our own approach.

2.4 Economic approach

Economists have been interested in smoking as a choice, that is from the angle of rationality. The question then is how can rational individuals chose to consume a good which has detrimental effects on health? One answer is given in the rational addiction model of Becker and Murphy (1988). Interestingly, the rational addiction

[7] For instance, a number of studies have used the Framingham study, an important epidemiological study where the population of a village has been followed for over fifty years so as to study morbidity and mortality among its members.

Table 1: Selected Previous Literature, Effect of Tobacco on Overall Life Expectancy

Authors	Data set	Method	Controls	Estimates
Doll et al. (1994)	Cohort of 34439 male British doctors	Comparing hazard rates of smokers and non smokers	sex, age, occupation	hazard rate 2 to 3 times for smokers
Shaw et al. (2000)	Cohort of 34439 male British doctors	Comparing hazard rates of smokers and non smokers	sex, age, occupation	Each cigarettes reduces life expectancy by 11 minutes
Phillips et al. (1996)	7735 middle aged British men	Duration to death, Cox proportional hazard model	marital status, alcohol consumption, body mass index	hazard rate of 2.41 for smokers compared to non smokers
Rogers and Powell-Griner (1991)	US National Health Interview and National Mortality Followback	Comparing life expectancy of smokers and non smokers	age, sex	Heavy smokers have a 25% shorter life expectancy compared to non smokers
Smith and Shipley (1991)	Whitehall Study of civil servants aged 40-64, UK	Comparing probability of death for smokers and non smokers.	age, occupation	Smokers have a two fold probability of death.
Hummer et al. (1998)	US 1986 National Mortality Followback Survey	Compare hazard rates of smoker and non smoker	sex	Hazard up to 3 to 5 higher for smokers versus non smokers

model eschews the question of why certain individuals smoke whilst some others do not (the selection question), but assumes that individuals are born with a level of stock of the addictive good. It is the interaction of this level of stock, of preference parameters, and of the environment (prices, income) which determines whether a given individual smokes or not, as the result of intertemporal optimisation. Because smoking is viewed as a choice, the consumption of tobacco must respond to standard economic incentives such as income and mainly prices. The focus of economic studies of tobacco consumption has therefore been to evaluate the effect of changes in prices on consumption (for instance Becker et al. (1994), Chaloupka (1991) and DeCicca et al. (2001) for measurement of price elasticities of consumption of tobacco).

To our knowledge, only two papers address the question of selection into smoking, by Rosenzweig and Schultz (1983), and Evans and Ringel (1999). Both papers study health outcomes, precisely birth weight of babies, conditional on smoking of the mother. Both allow for endogeneity of smoking and instrument it with prices (or taxes). Evans and Ringel (1999) find evidence of endogeneity, hence of selection, whilst the evidence from Rosenzweig and Schultz (1983) is more mixed. Both use prices as well as a number of individual and local characteristics as instruments, but they explain little in the variation in smoking, which might explain why the evidence in terms of endogeneity is mixed.

2.5 Standard approaches to measurement in the presence of endogeneity

To fix ideas, let us consider an illustrative version of the endogeneity problem, where individual i's age at death, T_i is related to individual characteristics X_i, to smoking behaviour S_i, to ε_i^*, an unobserved characteristic of individual i such that $cov(S_i, \varepsilon_i^*) \neq 0$, and to a random shock u_i:

$$T_i = \beta X_i + \alpha S_i + \underbrace{\varepsilon_i^* + u_i}_{\text{unobserved}} \tag{2}$$

The economic and econometric literatures have considered several possible ways to deal with endogeneity problems such as outlined above.

A natural solution, as exploited by Rosenzweig and Schultz (1983) and Evans and Ringel (1999) is to use an instrumental variable approach. One would need to find an instrument, correlated with smoking and uncorrelated with the unobservables in equation (2). This is easier said than done. Any individual characteristic could arguably figure as a control variable in the regression model (2). Indeed, epidemiologists have argued that education levels, occupation, income, stress... have a direct effect on health and mortality, whilst economists would argue that they have an effect on smoking. Another candidate as an instrument could be prices, to the extent that they influence smoking behavior. [8] However, in a setting such as equation (2), what is usually thought to influence the outcome is a measure of smoking over the entire life cycle. Using time series variation in prices would not be satisfactory, as prices would mainly pick up cohort effects. Younger cohorts would have faced higher prices than older ones. However, at any point in time, mortality is directly explained by cohort effects. Finally, spatial variations in prices are not very big and it has been argued that these are endogenous too.

The announcement of a link between smoking and health in the nineteen sixties could be seen as an exogenous event, but it would also be linked with the date of birth. Moreover, the medical literature had started incriminating smoking well before the announcement of the Surgeon General in the US in 1964 and the Royal College of Physicians in the UK in 1962, so it might be possible that more educated individuals had already curbed their smoking behavior.

All in all, it is difficult to think of a good instrument for smoking patterns over the life cycle.

A second way, which is often used in economics is to impose more structure on the problem. We could augment the model with a structural model which would

[8] But DeCicca et al. (2001) find a limited effect of prices at least on young individuals.

specify how S_i and ε_i^* are related. For instance, we could consider a model where the individual maximizes the expected utility of smoking and of longevity. Such a model would lead to a pair of behavioural equations, linking smoking behaviour and health behaviour to individual characteristics, both observed and unobserved, and the environment. Identification would then result from the structure of the model. It is worth considering the identification of such a model more closely. Suppose that the model restricts smoking to be affected only by one unobserved characteristic ε_i^* and by observed characteristics X_i: $S_i = S_i(\varepsilon_i^*, X_i)$. Conditional on X, observed smoking patterns would give us a clear signal of the unobserved characteristic. By inverting the relationship, we could express ε_i^* as a function of smoking and other observable characteristics. We could then go back to equation (2) to get a consistent estimate of the effect of tobacco, as we would now "observe" ε_i^*. However, it is doubtful that, conditional on X, all the heterogeneity in smoking patterns can be explained by a unique heterogeneity term also affecting mortality. Indeed, variation in smoking could also arise from unobserved taste for tobacco, η_i, so that the model would be $S_i = S(\varepsilon_i^*, \eta_i, X_i)$. In this case, conditional on X, smoking is a blurred signal of differences in both ε^* and η. We could no longer invert this relationship to extract information on ε_i^*, unless we know the joint distribution of both ε_i^* and η_i. As this is rather unlikely, a structural model would be, in this case, rather useless as an identification tool.

A third identification approach would be to use panel data on morbidity, assuming a fixed effect common to mortality and morbidity and potentially correlated with smoking. Conditional on functional forms, the fixed effect can be predicted in a two step procedure involving the estimation of the morbidity equation in first differences. Unfortunately, this method is only valid to the extent that one knows the correct specification which relates smoking to morbidity. Any mispecification would bias the estimate of ε_i^* and thus of the relationship between smoking and mortality.

We will now expose the very simple but novel strategy we follow to control for selection into smoking and measure the effect of tobacco on mortality consistently.

3 Identification using medical and epidemiological information to construct a proxy for tobacco-free morbidity

The previous sections illustrate the difficulty of identifying the effect of tobacco on mortality using the tools of epidemiology or of economics alone. We propose an identification strategy which borrows from both fields of study, as well as from medicine. Examining equation (2), we see that there would be no problem of endogeneity if we could observe ε_i^*, that is if we could observe common influences on smoking and mortality, or if we could observe a proxy for these influences. Suppose that ε_i is a proxy for ε_i^*, so that:

$$\varepsilon_i^* = \delta_0 + \delta_1 \varepsilon_i + v_i$$

If δ_1, which measures the partial correlation between the proxy and the variable of interest, is different from zero, then we can obtain consistent estimates of the effect of tobacco on mortality by plugging the proxy ε_i into the equation of interest, under the following assumptions:

$$\begin{aligned} E[u_i|X_i] &= E[u_i|S_i] = E[u_i|\varepsilon_i^*] = E[u_i|\varepsilon_i] = 0 \\ E[\varepsilon_i^*|X_i, S_i, \varepsilon_i] &= E[\varepsilon_i^*|\varepsilon_i] = \delta_0 + \delta_1 \varepsilon_i \end{aligned}$$

The first line of assumptions concerns the equation of interest and states that the errors u_i in that equation are uncorrelated with the conditioning variables X_i and S_i , the unobserved variable ε_i^* and the proxy ε_i. The second line of assumptions concerns the equation for the proxy and states that the errors from that equation, v_i, are uncorrelated with the conditioning variables of the model.

In our context, ε_i^* represents unobserved characteristics of the individual which influence both mortality and smoking but are not caused by tobacco. Fisher proposed that these characteristics could be the individual's genotype. The 1964 Report to the Surgeon General concludes that it is not possible "to rule out the existence of predisposing constitutional or hereditary factors", which could influence both smoking and mortality. Finally, recent research by Munafo et al. (2003) suggests that there is indeed a link between genes and behaviour. Our view is that the relevant individual characteristic is probably broader than the genotype. Indeed, it is possible that not only the genotype, but also randomness or accidents as well as life-style choices determine our state of health, hence our mortality and could influence whether we smoke or not (without being caused by smoking). For instance, individuals with the same genotype, such as identical twins do not have the same phenotype and will not have identical health trajectories. Some of the differences will be determined by "accidents" or random events, some other by choices. We will come back to the treatment of life style choices in our method in detail in section 4.3. For now, let us note that it is sometimes said that smoking and drinking are complementary activities (but also sometimes that they are substitutes). In the former case, individuals who drink have a greater probability to be smokers as well. But it is not possible to determine whether there is a causal effect from one to the other or whether they are jointly determined.

In practice, we do not observe individuals' genotype, or all relevant accidents and life style choices, but we observe information which may be linked to the genotypes and to the rest of the elements affecting an individuals' fundamental health. Indeed, we observe individuals' health, and we propose to use part of the information contained in the variation in health between individuals as a proxy for ε_i^*. More precisely, we use variation in health which is not caused by smoking. Suppose we can construct an indicator of morbidity that is correlated with mortality but not caused by smoking. This morbidity indicator will constitute a valid proxy for ε_i^* if it satisfies the conditions detailed above. We need some characteristic of the individual which

causes mortality but is not caused by smoking. Whilst there is disagreement among the medical and epidemiological communities about the extent of the detrimental effect of tobacco on health, there is agreement on which illnesses are caused and, to some extent, on which illnesses are not caused by tobacco. Illnesses can therefore be organized into three categories: illnesses that are caused by tobacco (e.g. lung cancer, chronic obstructive pulmonary diseases), illnesses that may be caused by tobacco but may also occur independently from it (cardio vascular diseases, stomach ulcer, type II diabetes...), and finally illnesses that occur independently from the individual's consumption of tobacco (congenital diseases, urinary infections, back pain...).

Using medical and epidemiological information, we can determine, for any given medical condition, which category it falls in. On the basis of this information, we can construct an index of the part of morbidity which is independent from tobacco. The part of an individual's morbidity which occurs independently from tobacco is what we use as a proxy for ε_i^*. It is a valid proxy if it is correlated with life expectancy, but not caused by smoking. Note that out identification strategy is valid if the development of illnesses that are not caused by tobacco is the same for smokers and non smokers.

A similar methodology is often used in labor economics: proxies for ability such as test scores or IQ scores are often used for ability in wage equations.

We now turn to the presentation of our data set and on the construction of the tobacco-free morbidity index we use as a proxy.

4 The Data

We use data from the Swedish Survey of Living Condition, "Undersökningen av LevnadsFörhållanden" (ULF). Approximately 6000 individuals, representative of the whole population, are surveyed each year. The ULF reports information on quantities smoked, smoking history, education, occupation, family composition, in-

come as well as many health measures. The data set has been merged with the Record of Deaths in 1999, so that we observe whether a given individual is alive up to the end of 1998, and if not, the date and cause of death. We use the 1980-81, 1988-89 and 1996-97 cross sections, as in these years the survey has a special section on health. In total, the data set includes 38986 individuals and we observe 6593 deaths. Within this large dataset, Statistic Sweden has constructed a smaller panel data set which follows individuals for two or three interviews (about 5000 individuals which we use for robustness checks).

Table 2 displays the characteristics of the data set and of the Swedish population. About half of the individuals in the sample are or have been smoker. Men are more likely to be or to have been smokers. The smoking prevalence is around 25%, with similar proportion for men and women. However, at young ages (under age 30), women are more likely to be smoking. The number of cigarettes consumed per day is low compared to other countries (15.5 in the UK, 24 in the US [9]).

Regarding smoking behaviour, we observe the quantities smoked, the duration of the smoking habit, and for some individuals the age at which they start smoking. However, individuals are not asked complete histories, but rather they answer questions from which it is possible to construct histories under the assumption that they have smoked continuously since they started smoking (until they have quit if they have). This is a drawback of this data, in that it does not allow the analysis of multiple smoking spells.

The survey records traditional individual outcomes and characteristics, such as education, occupation, family composition, or income. It is important to note that other risk taking behaviour, such as consumption of alcohol or of snuss (a variety of chewing tobacco) are recorded, as well as risky occupations.

We first present the morbidity information content of the data, before turning to the construction of the tobacco-free health scores.

[9]Sources, UK: British Household Panel Survey, 1995, US: World Health Organization, 2000.

4.1 Morbidity

The data set contains an extensive set of health questions, including self-assessed health, body mass index, hospital visits, ability to run, walk or climb stairs. The survey also asked extensive information on any specific health problems which were coded with the International Classification of Diseases (ICD 8 and 9) by nurses. Each individual can report up to six different health problems. In addition to all this information, we also have information on the severity of the disease (coded in 4 modalities) and an indication on when this problem started, so we can distinguish acute from chronic problems. These health problems range from relatively minor problems such as back pain or skin problems to life threatening such as specific cancers, ischemic heart problems or diabetes. In total, there are 155 variables to describe the health of an individual.

To summarize the information contained in this large number of variables, we construct a general morbidity index, using principal components analysis. We use indicators of general health, of perceived state of health relative to one's cohort, an indicator of the existence of long term illness, indicators for the range of body mass index in 3 modalities, indicators of whether the individual can run, walk up a flight of stairs, and board a bus. We also use information on the presence of heart conditions, of insomnia, anxiety, of taking antibiotics, of coughing, having a skin condition, having been to the hospital in the past two weeks, of being diabetic, having a neoplasm, hypertension, asthma, ischemic problems, cerebral problems, problems with arteries, veins, pulmonary obstructive diseases, stomach illness, hernia, cirrhoses, etc...

The morbidity index is increasing with age. Its variance is also increasing with age until around 85 years old, after which it decreases. However, there is considerable heterogeneity even at young ages. [10] The index is evidently correlated with smoking

[10] From the panel dimension, health appears to be very persistent through time. Individuals in poorer health in one period are very likely to be in poor health eight years later. In fact, health appears to be a random walk at the individual level.

as we have included all observed conditions, some of them being directly caused by smoking. We turn next to the construction of several tobacco-free morbidity scores, which will be used to proxy for the omitted common determinant of health and smoking.

4.2 Tobacco-Free Morbidity Scores

As described in section (3), the identification strategy we employ combines simple econometric techniques with medical and epidemiological knowledge. The latter is used to isolate medical conditions of which it is known that they are not caused by tobacco. Our proxy for tobacco-free morbidity will be constructed using variability in diseases that are not caused by tobacco. To construct the tobacco-free morbidity index, we disregard a number of diseases which have been linked to tobacco consumption. These include a number of cancers (eg cancers of the lung or of the oral cavity), all cardiovascular diseases (including ischemic heart disease and hypertension), respiratory diseases and diseases of the oesophagus (which includes stomach ulcers). We also disregard general health measures such as self-assessed health, body mass index and a number of variables describing the ability to walk or climb stairs, which could be caused by smoking.

To establish whether a disease should be included or excluded from the proxy, for each morbidity indicator in our data set, we checked in the medical and epidemiological literature whether the disease has been linked to smoking. While it is easy to exclude well researched diseases such as cancers and cardiovascular problems, it is more difficult to classify more particular ones. For some diseases, it may be that no link is known because the medical profession has not yet established a link between smoking and morbidity or mortality. Furthermore, drawing the line between diseases is also made more difficult given the frequent confusion in the literature between correlation and causation. These reasons could lead us to either include or exclude too many diseases from the tobacco-free morbidity score. The latter case would result in a loss of power for the proxy. With fewer diseases, the

health score is less likely to contain any tobacco related diseases, but it will perform more poorly as a proxy for the unobserved general health. The former case, where diseases caused by tobacco are included in the tobacco free morbidity scores, would lead to there remaining some bias in the estimated effect of tobacco on mortality. There is therefore a trade-off between a potential bias due to the definition of our health score and its power.

In the absence of an entirely clear line of demarcation between types of diseases, we construct three different tobacco-free morbidity scores, each with fewer and fewer diseases included. The first score contains 29 health conditions, the second 19 and the third score contains only one health condition, namely adult height adjusted for sex. A list of the morbidity indicators used to construct the three scores are listed in Table 3. We believe that by using three scores, the last of which is obviously not caused by tobacco [11], we are able to deflect the criticism that our results are due to remaining endogeneity.

We construct the three health scores, (labelled as Score 1, 2 and 3) by using the factor analysis discussed above and selecting only the relevant diseases. To check whether the health scores are correlated with subsequent mortality, we estimate the effect of the health scores on the duration to death using a Cox proportional hazard model. We rank the individual's tobacco-free morbidity within age groups (using 10 years bands) and we classify individuals who are in the lowest 25% quantile as being in good health. Similarly, we classify individuals in the upper 25% quantile as being in poor health. The hazard ratio for poor health compared to good health is equal to 1.24, [1.15, 1.33] for Score 1, 1.16 [1.08, 1.25] for Score 2 and 1.09 [1.01, 1.17] for Score 3 (95% confidence intervals in brackets). All of the three health scores predict mortality, although not surprisingly, the effect is stronger the more health

[11] While maternal smoking leads to low birth weight, the rate of growth of these children in subsequent years compensates the initial handicap, so that, at puberty, there is no impact of maternal smoking, see for instance Ong et al. (2002). From a purely technical point of view, note that if low birth weight leads to shorter adult height, height could nonetheless be used as a proxy for ε_i^* provided it is note caused by the individual's smoking.

conditions are included.

Without loss of generality, each of the health scores have been normalized between 0 (for the individual with best health) and 100 (worst health).[12]

4.3 Other Risky Behaviors

Health can be affected by risky behaviors other than smoking. For instance, smokers are also more prone to be heavy drinkers or to drive without a seat-belt (Hersch (1996)). Should our proxy for the individual's underlying health capture the effects on health of other risky behaviors? It depends on the question we want to address, whether we want to evaluate the medical effect of tobacco on health or the total effect on health of a smoking ban. In the first case, we want to compare the health of a smoker to the health of a non smoker with similar characteristics, including other risky behaviors such as drinking. The presence of morbidity indicators caused by other risky behavior is therefore not a problem.

In the latter case, the use of morbidity indicators related to other risky behavior matters for the interpretation of the results. The interpretation depends on whether smoking and other risky behaviors are substitutes or complements. The literature on this subject gives mixed results. [13] Eradicating tobacco may lead the individual to either increase or decrease other risky behavior, and this may have an impact on the individual's health. If the tobacco-free morbidity scores are influenced by, say, drinking, we would like to contrast the health of a smoker to that of a non-smoker who either drinks more (substitute) or less (complement).

Given the lack of clear results in the literature, we adopt two strategies. First, we have tried to disregard morbidity indices which are too obviously related to other risky behavior (such as cirrhosis and diseases of the liver). Our last health score is

[12]Regarding Score 3, height adjusted for sex, it means that high values of the score correspond to short height (adjusted for sex) and vice-versa.

[13]Chaloupka (1999) finds that smoking and marijuana appears to be complements. Dee (1999) finds that smoking and drinking appears to be complements, while Decker and Schwartz (2000)) find that smoking and drinking are substitutes.

certainly immune from any effect of other risky behavior. Second, given that we have information on other risky behavior in our data set, we also use this information as control variables in our regressions.

4.4 Comparison with Previous Studies

Before turning to the measure of the effect of tobacco on mortality correcting for selection, we replicate, on the Swedish data, the studies detailed in Table A above, so as to have some point of comparison. We find that the hazard ratio, when controlling for age, sex, marital status, BMI, and alcohol consumption is 1.3, to compare with a hazard ratio of 2.4 in the study of Phillips, Wannamethee, Walker, Thomson, Davey-Smith (1996). Our results show that the estimated effect of tobacco is consistently smaller in the Swedish data than in the epidemiological studies we use as benchmark. We believe the hazard rate for smokers is much lower for the Swedish data because average quantities smoked are much lower in Sweden than in the US or in the UK.

We will now use the tobacco-free morbidity scores to establish that there is selection into smoking, in the sense that smokers are not drawn randomly from the distribution of tobacco-free morbidity scores. We will then turn to measuring the effect of tobacco on mortality controlling for selection.

5 Selection into smoking

We will now proceed to show that there is indeed selection into smoking in that individuals with worse tobacco-free morbidity tend to be associated more with smoking. To do this, if it is the case that there is no causal link from smoking to the tobacco-free morbidity status as measured by the three scores, it suffices to document the existence of a correlation between the tobacco-free morbidity scores and smoking. We will then discuss how the pattern of evidence in the data goes against the hypothesis that the tobacco free morbidity scores are contaminated by diseases caused by tobacco.

Smoking is captured in three dimensions: whether an individual is a smoker or not, the quantities smoked and finally the duration of the smoking habit. We will show that there is selection in these three dimensions, starting with whether individuals are or have ever been smokers. We will also provide robustness checks of our identifying assumption.

5.1 Selection into smoking status

The first piece of evidence we present, in Figure 1, is the relationship between smoking and the tobacco-free morbidity scores. Each curve presents the average (by age) of tobacco-free morbidity as a function of age respectively for smokers (current and formers) and for non smokers.

The values taken by the morbidity scores are increasing with age, indicating a worsening of tobacco-free morbidity with age, for both smokers and non smokers. Recall that the third morbidity index (Score 3) is the opposite of height adjusted for gender, so the fact that it increases with age merely indicates that younger generations are taller. As the other two morbidity scores include height for sex, the increases in the first two scores with age are partly due to differences in height across generations. However, the increase in their values with age also reflects the fact that, independently from smoking, health tends to deteriorate with age.

Turning now to the correlation between smoking and tobacco-free morbidity, the first two graphs illustrate that, at all ages, smokers (current and former) are on average in poorer health than individuals who have never smoked. However, for Score 1 and Score 2, at around age 50, the difference in morbidity vanishes and smokers and non smokers appear to have essentially the same health. Note that a decreasing number of observations at later ages means that the confidence bands get larger after age 60 and probably explains the pattern for Score 3.

These results, which show a difference in terms of tobacco-free morbidity between smokers and non smokers, are crude averages which do not control for any other characteristics. We will present results obtained controlling for observed char-

acteristics below, but first we discuss why the difference in tobacco-free morbidity between smokers and non smokers becomes less significant at older ages.

To interpret the fact that the tobacco-free morbidity scores of smokers and non-smokers appear to converge at older ages, we need to separate the age effect from the cohort effect. To this end, we use the repeated cross section feature of our data set. To be able to compare the health of individuals at different ages, we consider their health quantile within an age group, rather than their health score directly, as health is age dependent (we use 10 years age bands). Next, we compute the average quantile in the general health distribution of a smoker by year of birth, from 1904 to 1977. Under the assumption that there is no selection on health, the median health quantile of smokers and non smokers should be equal to 0.5.

Figure 2 displays the median tobacco-free health quantile of smokers as a function of year of birth. Note that we have three cross sections, so we observe a same cohort up to three times at different ages.

The further from the 0.5 line is the average quantile for health for smokers, the more selection there is. As apparent from the figure, there is some evidence that smokers born around 1900 were in better health than non smokers. This could either be a healthy survivor effect or the fact that in the beginning of last century, mostly affluent and well-off people (who are also in better health) smoked. This effect becomes negligible between 1910 and 1940. From 1940 onwards, smokers are increasingly coming from a population with poorer health. [14] Among individuals born towards the end of the period of observation, the median smoker has a health in the 54% quantile of the health distribution. This pattern seems to indicate that the selection based on health started when information on the health effect of cigarettes was released. Those with the best health may have decided that smoking was not worth the risk, so that prevalence among this group decreases through time.

Next, we control for a number of observed factors. Table 4, Panel A presents the odds ratio corresponding to the probability of being a smoker (current or for-

[14] This effect is clearly apparent for both Score 1 and Score 2 and to a lesser extent for Score 3.

Figure 1: Tobacco-Free Morbidity (Score 1, 2 and 3) and Smoking

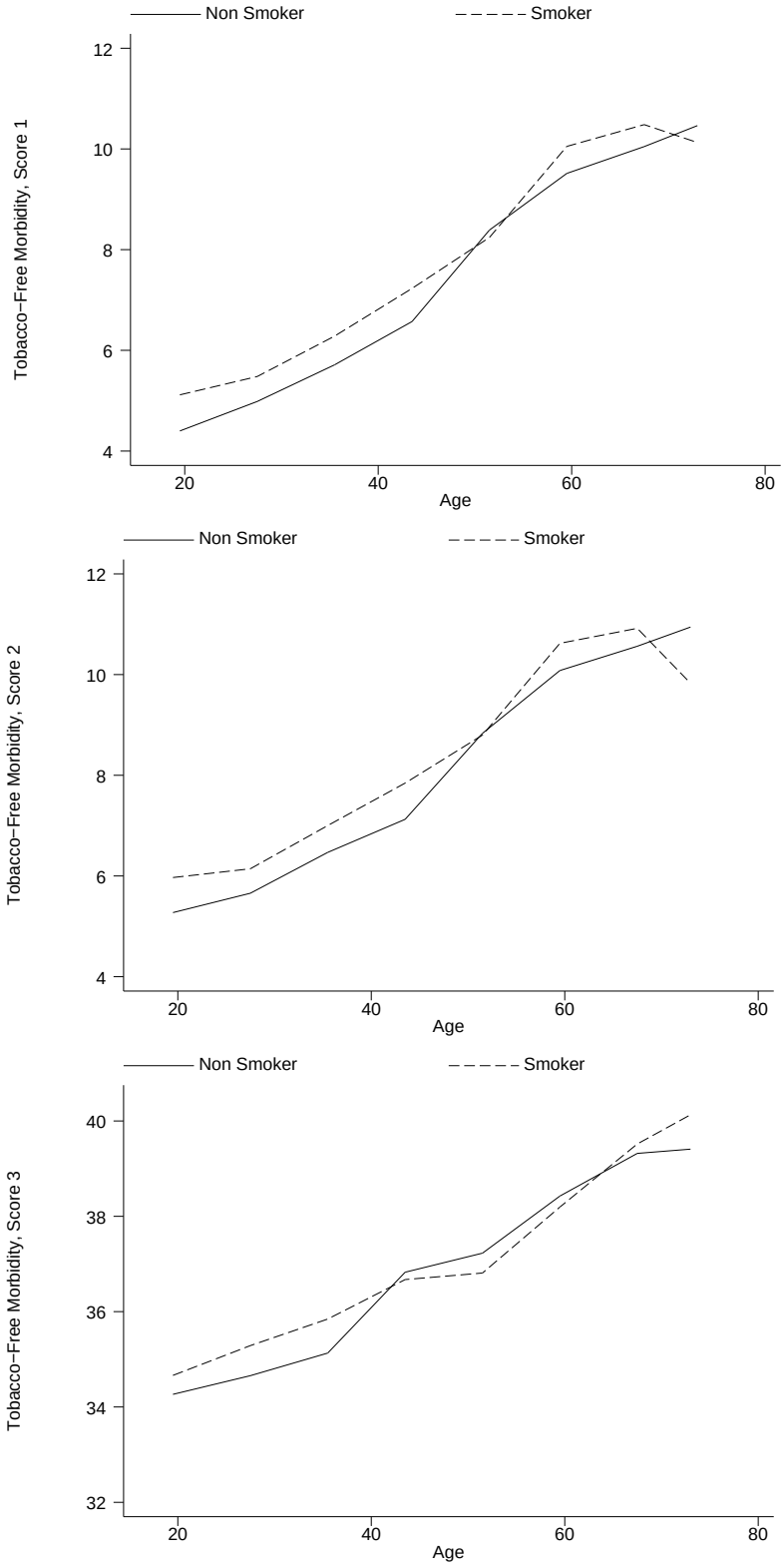
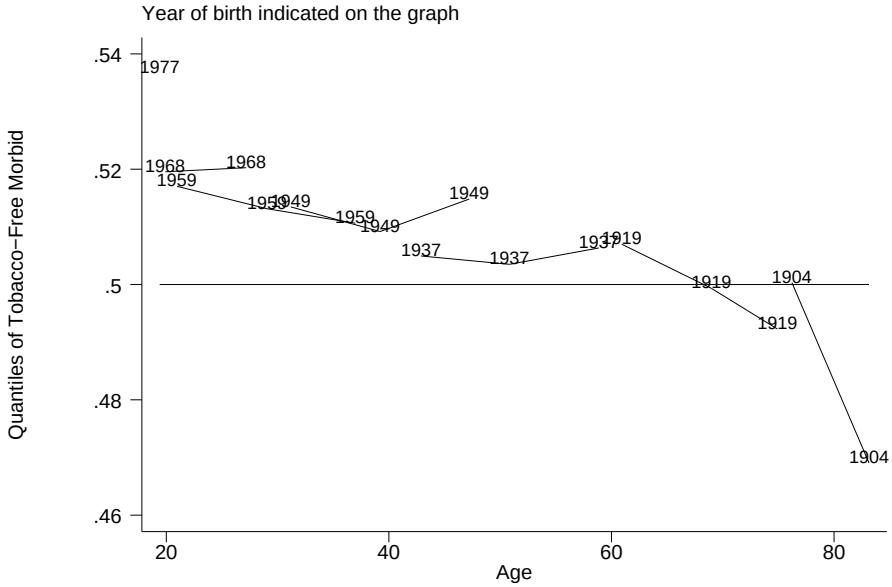


Figure 2: Average Quantiles of Health for Smokers (Current and Former)



mer), controlling for age, sex, education level, interview year effects, risk taking behaviors (on the job risk, snus consumption, alcohol consumption) and tobacco-free morbidity. The odds-ratio are calculated for the 25% of individuals with the best tobacco-free morbidity relative to the 25% of individuals with the worst tobacco-free morbidity (where the ranking was defined within age groups as above). We find essentially the same results as in Figure 1. Controlling for individual characteristics, at older ages, tobacco-free morbidity does not appear to predict the smoking status (the odds ratio is not different from 1 for a sub-sample of individuals aged more than sixty). However, when we look at younger individuals, the effect of being in poor health (in terms of tobacco-free morbidity) becomes significant and is quite strong. When we consider Score 1, and selecting individuals aged less than 25, an individual with poor tobacco-free morbidity is 1.36 times more likely to be a smoker than if he were in good health (in terms of tobacco free morbidity), everything else being equal. We get comparable effects when we use the second morbidity index (Score 2). For Score 3, the magnitude of the effect is reduced and the standard errors are larger, but the effect for the youngest individuals is still significant at the 5% level. This is somewhat expected as Score 3 contains much less information on the "health type" of an individual than the first two measures.

Overall, this provides evidence that there is a significant correlation between smoking and tobacco-free morbidity, except for older individuals, even when one controls for other risky behaviour. The fact that there is a correlation could be evidence of selection where individuals in worse health are more likely to engage in smoking, or it could be evidence that our health scores are not tobacco-free as we claimed initially. If we have included in the health score a morbidity variable which is in fact caused by tobacco, we would evidently find that smokers are more likely to be in worse health than non smokers. However, the pattern of the effect does not appear to be consistent with this latter explanation. If the tobacco-free morbidity indices did contain some illness related to tobacco, one would expect that the effect of health on smoking would be much stronger as age increases. Older

smokers would be more likely to develop tobacco-related diseases as they have been exposed to tobacco for a longer period. Not only do we not see this, but we find the opposite, a smaller correlation between smoking and the health indices for older individuals. Moreover, we find a correlation between tobacco-free health scores and smoking at young ages. If this were due to causal effects of smoking on health as encapsulated by our scores, it would mean that the diseases caused by tobacco which we would be capturing would have to develop fast enough to affect the youngest age group in a significant way (note that the median age in that group is only 20 years.) The idea that tobacco would affect health so rapidly goes against all received wisdom about the effect of tobacco on health, so that we believe it can be ruled out. We will provide further evidence that the tobacco-free morbidity scores do not appear to be contaminated by tobacco-related diseases in section 5.4. We interpret this as evidence that the tobacco free morbidity scores are indeed free of causal effect from tobacco. We also see the fact that older individuals are not affected as indicating a change in selection into smoking through the twentieth century.

We believe this evidence is very difficult to explain by means other than selection and moreover by an increasing selection for younger cohorts. Indeed selection appears to be absent for older individuals. This group was born between 1900 and 1936 and reached adulthood before the effect of tobacco on health was mass-publicized. On the contrary, the youngest group grew up in an environment where the risks were known. [15]

5.2 Selection into smoking intensity

We next look at the intensity of smoking for current smokers and its relation to the tobacco-free morbidity scores. We regress an indicator for heavy smoking (more than a pack a day), conditional on smoking, on tobacco-free morbidity indicators and the same controls as in the previous section. The results are presented in Table 4, Panel

[15] Viscusi (1990) and Kenkel (1991) show that smokers are aware of the risks associated with smoking, and sometimes over-estimate the risks.

B. The coefficients reported show odds-ratio for heavy smoking for the poor health group compared to the good health group. We find significant effects for individuals between 25 and 50, using the first two morbidity scores. Poor health individuals are about 1.3 times more likely to be heavy smokers.

At a young age, the number of cigarettes smoked may not be linked to the underlying morbidity of the individual, but healthier types of individuals reduce the quantities they are smoking as they age.

Panel C provides the results for the probability of smoking more than a pack a day, unconditional on smoking, for individuals between 25 and 50. These odds-ratio combines both the selection on the extensive and the intensive margin. For the first two health scores, the results are similar. Overall, a poor health individual is 1.8 times more likely to be a heavy smoker, compared to a good health individual. The effect is smaller when we use Score 3, at around 1.3, and is significant at the 10% confidence level.

5.3 Selection into smoking duration

We now turn to the duration of the smoking habit. There are two dimensions, the age at which individuals start smoking and the overall duration. Table 4, Panel D presents the relative effect of poor compared to good tobacco-free morbidity on the probability of starting smoking before age 15 (strictly). The table presents the odds-ratio, obtained from a logistic regression of the probability of starting smoking before age 15, controlling for sex, education level, other risk taking behaviors, interview year dummies and age. Using the first two scores, poor health individuals are about 1.5 times more likely to start smoking at an early age. For Score 3, this figure is about 1.2. In this regression, we have restricted the sample to young individuals who are observed smoking in order to be able to compute an accurate measure of the starting age.

Based on the information on the duration of the smoking habit, we estimate a Cox proportional hazard duration model to quitting. We include as regressors a

list of individual characteristics (sex, education, occupation, interview year effects, risk taking behavior other than smoking) as well as an indicator for the tobacco-free health group. The regression is stratified by sex, education level and occupation.

The results are presented in the form of hazard ratios in Table 4, Panel E. In all cases, the poor health individuals are less likely to give up smoking. The magnitude of the effect is about 10% and remarkably consistent across health measures.

Interestingly, the epidemiological literature often finds significant beneficial effects of quitting smoking (see for instance Doll and Hill (1956), Hammond and Horn (1958), Doll and Peto (1976), Kawachi et al. (1993), Kawachi et al. (1997), Hrubec and McLaughlin (1997)). The results presented above do not dispute the fact that quitting may result in lower rates of lung cancers or any other tobacco related diseases. But they indicate that the *overall* benefit of quitting smoking is probably somewhat lower than what has been indicated in the literature, given that these studies do not control for the underlying health of the individuals.

5.4 Further Robustness Check: Changes in tobacco-free morbidity

Finally, to provide a further robustness check of our methodology, we investigate the extent to which the changes in the tobacco-free morbidity scores are related to either smoking status, quantities or durations. Indeed, our identification strategy requires that the evolution of tobacco-free health be independent from smoking. If we found that the value of the indices increases with quantities or duration, one would be suspicious that one of the morbidity indicators used to construct the health scores might be causally related to tobacco. We therefore check that this is not the case. To this end, we use the panel data contained within our larger repeated cross-section data. We regress the change in the tobacco-free morbidity scores (eight years apart) on smoking status, quantities smoked and duration. The results are displayed in Table 5. We cannot use the third health score (adjusted adult height) as this is a fixed characteristic of the individual. For the first two health scores, we cannot

find any evidence that the health of smokers deteriorates faster than for those who have never smoked, even for individuals older than age 40. We find similar evidence when we investigate the role of smoking intensity. Finally, the duration of the habit appears to be uncorrelated with changes in health scores. We conclude from these results that our morbidity indicators are picking up health problems not related to tobacco. The difference in tobacco-free health *levels* between smokers and non smokers are therefore likely to be the consequence of selection.

6 The effect of tobacco on mortality

In this section, we evaluate the effect of smoking on mortality controlling for selection.

We have shown in the previous section that smokers are more likely to be drawn from a population with worse tobacco-free health. Given that health is correlated with subsequent mortality, the fact that there is selection implies that comparing the life expectancy of smokers to that of non smokers will not give the correct effect of smoking on mortality. This is true even when conditioning on usual observed characteristics such as sex, education levels and even other risk taking behaviors. The correct way to proceed is to compare the life expectancy of individuals, smokers and non smokers, who would have the same life expectancy if they did not smoke. This is what we propose to do using the tobacco-free morbidity scores, which are constructed to capture the fundamental health of the individual, independently from smoking.

However, before we do this, we have to recall that we have also shown that selection is mostly present for younger generations who have been aware of the risks they faced when choosing to smoke and started in the nineteen fifties and nineteen sixties. This has two important consequences for the measurement of the effect of tobacco on health. Firstly, note that studies which investigate the effect of smoking on mortality rely mainly on elderly individuals for identification (as individuals who

die are essentially drawn from the eldest cohorts of both smokers and non smokers), and we have seen that this is a population for which there is a minimal selection bias. This means that previous epidemiology studies probably do not miss much by ignoring selection on the basis of underlying health. The second consequence is that, as time passes, such studies will obtain different results and hence conclude to a worsening effect of tobacco on health, when what is happening is increased selection. Indeed, from 2010-2020 onwards, the generations born in the nineteen fifties and nineteen sixties will start to face an increased likelihood of death and studies that use data on mortality and smoking alone will spuriously reveal a worsening effect of tobacco, as these studies will increasingly compare poor health smokers to non smokers in better health.

We can easily document this using our data. We estimate a model of duration to death, where we condition for a set of individual characteristics, smoking and non tobacco-related morbidity. The results are displayed in Table 6. We obtained them from a Cox proportional model, stratified by sex and education level, controlling or not for tobacco-free health. The first line of the table shows the hazard ratio for smokers as compared to non smokers (1.31) and for heavy smokers as compared to non smokers (1.85). The second line of the table shows the same hazard ratio, when stratifying on tobacco-free health as well. The values of the hazard ratio do not change, respectively at 1.30 and 1.83. As expected, given that the individuals who are observed dying are largely drawn from the older cohorts for whom there is little or no selection into smoking, the hazard ratio is unchanged when we control for health. We are not in a position to document that there is a selection on underlying health for the younger generations using the model of duration to death, as there are too few deaths among the cohorts concerned at this point. However, we have already documented above that there is selection for these cohorts, which makes the point.

These remarks are important as they determine how we can, using data where essentially those who die did not select into smoking, obtain a consistent measure of

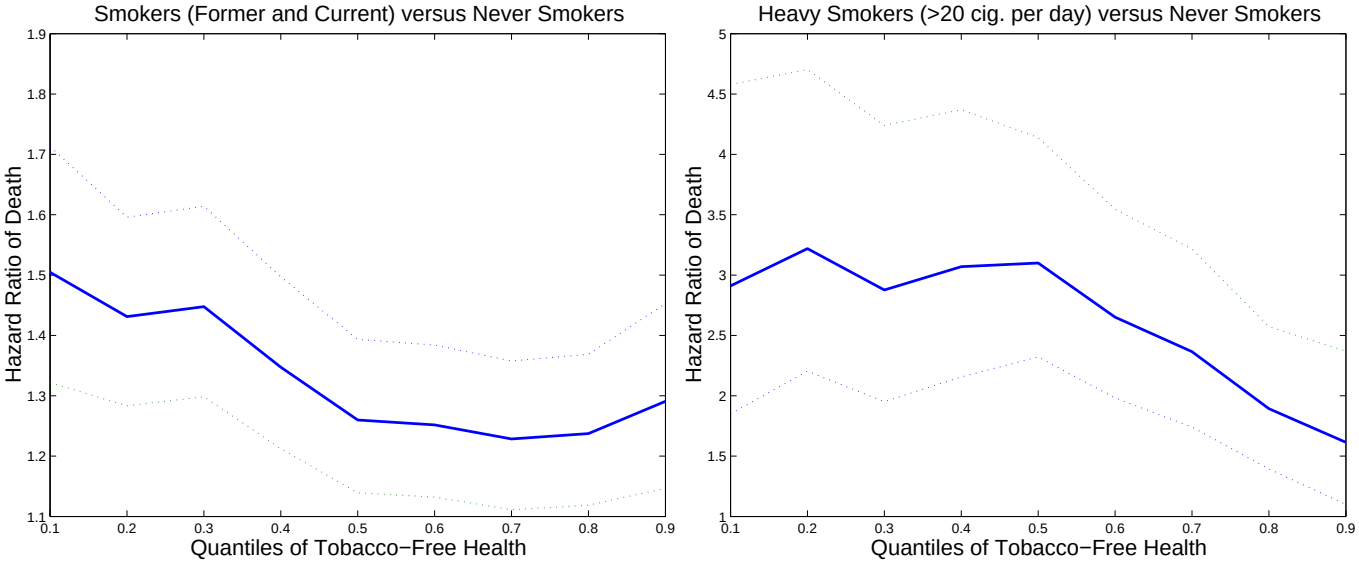
the effect of tobacco on mortality for selected groups. Indeed, a stratified model such as estimated corresponds to the assumption that the baseline hazard is unique to each stratum, but that otherwise, different observations have identical coefficients:

$$\lambda(t_i) = \lambda_{0k}(t_i) \exp(X_i\beta)$$

for observation i in strata k. It is possible that such a model is too constraining for the data, and that with the emergence of selection and the change in the composition of the smoking population, it is necessary to allow for both the baseline hazard and other coefficients to differ between types of individuals. We estimate a model where we allow the baseline hazard to differ between strata defined on sex and education and other coefficients to differ between health groups defined as before. Table 5, lines 3 and 4 show the hazard ratio for poor health smokers compared to poor health non smokers, which show that at each duration, the former have a 1.25 higher chance of dying than the latter (for heavy smokers, the figure is almost 1.6). Turning to good health smokers, the hazard ratio is 1.45 for smokers and 1.84 for heavy smokers. Finally, we estimate a last model, where we only stratify on sex and allow all coefficients to differ by health groups and education level. The results are not very different from those obtained above. Note that here, we are able to capture differences in the parameters for good and poor health individuals because we are exploiting data from all age groups, so that even though there are few deaths among the younger cohorts, and little selection among the older cohorts, put together, there is enough variation that differences can be made apparent using a Cox proportional hazard model. Figure 3 plots the hazard ratio of death as a function of the quantile of tobacco free health. The first panel displays the hazard for smokers (current and former) compared to never smokers. The hazard ratio of death is equal to 1.5 for the healthiest individuals and decreases at about 1.2 for poor health individuals. The second panel displays the hazard ratio for heavy smokers (more than 20 cigarettes per day) compared to never smokers. The hazard ratio varies between 3 for the healthiest to about 1.5 for those with the poorest health.

It is interesting to convert the numbers obtained in terms of hazard ratio into

Figure 3: Heterogeneity in the Effect of Tobacco on Mortality



numbers of years of life lost. Poor health smokers lose on average 3 years, as compared to poor health non smokers, and 5.5 years if they are poor health heavy smokers. For individuals whose health is good independently from smoking, the losses are greater: 4.5 years from being a smoker, and 6.8 from being a heavy smoker. The loss from smoking is greater for individuals whose life expectancy is greater if they do not smoke. Note that this result is not an artefact of our data or our method. Indeed, recall that the 1964 Report to the Surgeon General on Tobacco and Health collates results from a large number of studies which report that the excess death rate of smokers is highest at younger ages (around 40 to 50 years of age). This is in agreement with the pattern we find where the loss is greater for long life expectancy individuals than for short life expectancy individuals. Individuals with a high ε_i^* (and thus with a long life expectancy) might loose more from smoking as they may die prematurely from, say, lung cancer. On the other hand, those with a short life expectancy even as a non smoker may loose less from smoking. If this is the case, it means that the selection effect will have some consequences on policies which try to reduce smoking prevalence. The effect of tobacco on mortality estimated on a population born at the beginning of the twentieth century will be misleading to predict the benefit of not smoking for a younger population. The real gain from not smoking will be declining over time due to the increased selection.

7 Compensating Foregone Earnings due to Tobacco Related Deaths

The previous sections showed the importance of health selection into smoking and its implication for the evaluation of the effect of tobacco on health and mortality. One cannot easily compare smokers and non smokers, even when controlling for observed characteristics such as gender, education or occupation. We investigate here the impact of health selection on monetary compensations as a result of a tobacco related death.

We calculate the amount of foregone earnings as a result of a premature death.

We do not attempt to evaluate the general value of life or the willingness to pay for an extended life-time. We combine information on survival and earnings and we compare smokers and non smokers, controlling for some observed characteristics such as gender and education level and more importantly for their underlying health as captured by our tobacco-free morbidity indicators.

We assume that we can approximate the survival and earnings patterns of smokers by the survival and earnings patterns of non smokers with similar characteristics if we include their underlying health. We calculate the net present value of earnings, weighted by the probability of survival for non smokers at all ages between 20 and 100. We interpret this amount as foregone earnings for smokers as a result of an early death.

For any possible age at death, we calculate the amount of compensation equal to the sum of future forgone earnings, weighted by the age and group specific probability of survival. We split the population into 18 separate group according to their gender, education level (in three modalities) and underlying health as measured by our tobacco-free morbidity Score 1 (in three modalities).

Let t be the age at death. Let $y_{i,t+l}$ represent the average income of an individual of age $t+l$ who is alive and in group $i = 1, \ldots, 18$. Let s_{it+l} be the conditional survival probability from age $t+l$ to $t+l+1$ of an individual in group i who has never smoked. We define the compensation as:

$$Comp_{it} = \sum_{l=0}^{\infty} (1+r)^{-l} y_{i,t+l} s_{i,t+l} \tag{3}$$

where r is the real interest rate. All income figures have been converted into US dollars, as of year 2000 for ease of reading. We fixed the interest rate to 3 percent annually. Figure 4 plots the envelope of compensations across all groups and for all possible age at death between 20 and 100. The amount of compensation is overall decreasing with age. It varies from 0.8 million dollars to almost zero at the oldest ages. Moreover, at any given age at death, there is considerable heterogeneity in the amount of compensations required. This is due to variability across groups in survival probabilities and in average level of earnings.

Figure 4: Average Compensation for Loss of Income As a Function of Age at Death, Millions of US Dollars

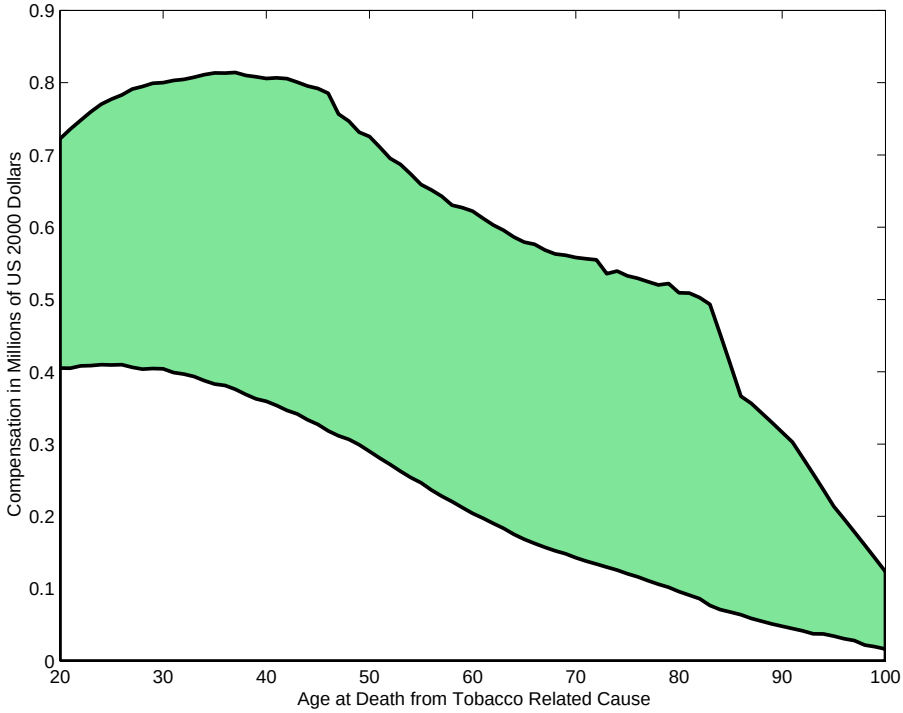


Table 7 displays the compensation figures by sex, education level and tobacco-free health groups at a hypothetical age at death of 55. All figures are expressed in millions of US dollars. We also provide in brackets the compensation based on the 20% and 80% quantiles of life cycle-earnings. For individuals with a low education, the figure is at about 0.1 million US dollars. This number is fairly similar for men and women and vary slightly with the underlying health of the individual. For medium educated individuals, and especially for men, the underlying health matters, and the compensations varies from 0.1 to about 0.2 million dollars. There is also a gender difference, reflecting the fact that men earn on average higher wages.

The highest compensation figure is for men with a high education and in good (tobacco-free) health. This is due to the fact that this group has the highest earnings and a low probability of death for non smokers. Going from the lowest health group to the highest doubles the average compensation estimate, from 0.14 to about 0.27 million dollars.

Conditioning on the tobacco-free health measures does matter quite a lot for individuals with a high or medium education. This is due to the heterogenous effect of tobacco on life expectancy documented in section 6.

8 Conclusion

This paper considers the effect of tobacco on mortality allowing smoking to be endogenous. If smoking and underlying health are correlated, most estimates found in the literature are biased. We discuss the identification of the effect of tobacco allowing for endogeneity and we propose a way to get a consistent estimate of this effect under weaker assumptions than are usually made in this literature. Our approach is to use a proxy for the unobservable element which causes the endogeneity bias. We use extensive data on date of death and morbidity, together with a model of duration to death to obtain estimates of the effect of tobacco on health which correct for selection. Our main findings are:

- There is evidence of selection into smoking. Everything else being equal, smokers come from a population in poorer health independently from smoking than non smokers. In other words, individuals with shorter potential life expectancy smoke more than individuals with longer potential life expectancy.

- The effect of smoking on life expectancy differs by types of individuals, with individuals with longer potential expectancy having more to loose in terms of years of life by smoking. The variation in terms of years of life lost is quite important, going from three to almost eight years. XXX

- This implies that the gains from reducing smoking are not as large as they would be thought to be without accounting for selection into smoking, given that health influences potential life expectancy.

- There is a strong cohort effect. The selection effect is important for the cohorts who started smoking when the information on the effect of tobacco on health was widely publicized, but not so much for previous cohorts.

- The existence of the cohort effect means that the results obtained in the past by epidemiological studies are not far off the mark for the generations considered but that future studies comparing smokers and non smokers will spuriously reveal a worsening effect of tobacco on health if they fail to control for selection.

- Finally, combining data on tobacco-free health, survival and income, we evaluate the amount of foregone earnings from tobacco related death and we show that foregone earnings depend crucially on the underlying health of the individual.

A number of factors could explain a correlation between smoking choices and mortality, above the sheer medical effect. For instance, both mortality and smoking decision could be influenced by other factors such as stress, neighborhood effects or social norms. It is also possible that smoking and mortality are linked through a trade-off between smoking and longer life expectancy. In this trade-off, individuals

with longer potential life expectancy might have incentives to smoke less. Finally, smokers and non smokers may have different discount factors. Whatever the reasons, it is important to try to separate out the true effect of tobacco from the selection effect, which is what we do here. In a companion paper, Adda and Lechene (2004), we examine the question of the structural mechanisms which can lead to the observed evidence.

A Appendix

A.1 Excerpt from "Cautions Against the Immoderate Use of Snuff and the Effects It Must Produce When This Way Taken into the Body", John Hill, 1761.

Excerpt from the first clinical study of the effect of tobacco on health, "Cautions Against the Immoderate Use of Snuff and the Effects It Must Produce When This Way Taken into the Body", by John Hill, 1761, London: R. Baldwin & J. Jackson: "Whether or not polypuses, which attend Snuff-takers, are absolutely caused by that custom; or whether the principles of the disorder were there before, and Snuff only irritated the parts, and hastened the mischief, I shall not pretend to determine: but even supposing the latter only to be the case, the damage is certainly more than the indulgence is worth. No man should venture upon Snuff, who is not sure that he is not so far liable to a cancer: and no man can be sure of that."

A.2 Excerpt on Causality from the 1964 Report to the Surgeon General on Tobacco and Health

The 1964 Surgeon General Report on Smoking and Health marks a turning point in the history of tobacco. By appointing a committee charged with the task of gathering evidence to the effect of establishing whether there is a detrimental effect of tobacco on health, the Surgeon General effectively brings the scholarly debate into the public arena. The introduction to the report states that: "Few medical questions have stirred such public interest or created more scientific debate than the tobacco-health controversy. The interrelationships of smoking and health undoubtedly are complex. The subject does not lend itself to easy answers. Nevertheless, it has been increasingly apparent that answers must be found." The report goes on to state that the reason why answers must be found is "to act in accordance with that evidence for the benefit of the people of the United States."

The committee is composed of 10 scientific members, who conduct an investigation, to which participate over 170 individuals and a dozen of organizations among which, interestingly, one can find most of the major American tobacco companies: the American Tobacco Company, Brown & Williamson Tobacco Corp., R.J.Reynolds Tobacco Co, Liggett & Myers, P.Lorillard Co, Phillip Morris Inc.

After examining all evidence available to date, the committee produces a report which summarises the evidence of the relationship of smoking to health, not only in terms of lung cancer, but in terms of a large number of other conditions and mortality. The report is also very careful to exposit what criteria have been used to assess the evidence.

As evidence, the committee uses "(1) retrospective studies which deal with data from the personal histories and medical and mortality records of human individuals in groups; and (2) prospective studies, in which men and women are chosen randomly or from some special group, such as a profession, and are followed from the time of their entry into the

study for an indefinite period, or until they die or are lost on account of other events."

The committee reviews more than 6000 articles published in some 1200 journals, plus reports, and other articles from various sources.Finally, tobacco companies are invited to submit statements to the committee.

The committee's mandate is summarised by the statement: "If it be shown that an association exists, then the question is asked: Does the association have a causal significance?" The key to the debate, then and now, is the establishment of a causal effect. The committee's position on the manner in which to establish causality is as follows:

"Statistical methods cannot establish proof of a causal relationship in an association. The causal significance of an association is a matter of judgment which goes beyond any statement of statistical probability. To judge or evaluate the causal significance of the association between the attribute or agent and the disease, or effect upon health, a number of criteria must be utilized, no one of which is an all-sufficient basis for judgment. These criteria include: a) The consistency of the association, b) The strength of the association, c) The specificity of the association, d) The temporal relationship of the association, e) The coherence of the association.

On causality:

"Without summarizing the more important concepts of causality that have determined human attitudes and actions from the days even before Aristotle, through the continuing era of observation and experiment, to the statistical certainties of the present atomic age, the point of view of the Committee with regard to causality and to the language used in this respect in this report may be stated briefly as follows:

1. The situation of smoking in relation to the health of mankind includes a host (variable man) and a complex agent (tobacco and its products, particularly those formed by combustion in smoking). The probe of this inquiry is into the effect, or non-effect, of components of the agent upon the tissues, organs, and various qualities of the host which might: a) improve his well-being, b) let him proceed normally, or c) injure his health in one way or another. To obtain information on these points the Committee did its best, with extensive aid, to examine all available sources of information in publications and reports and through consultation with well informed persons.

2. When a relationship or an association between smoking, or other uses of tobacco, and some condition in the host was noted, the significance of the association was assessed.

3. The characterization of the assessment called for a specific term. The chief terms considered were factor, determinant, and cause. The Committee agreed that while a factor could be a source of variation, not all sources of variation are causes. It is recognized that often the coexistence of several factors is required for the occurrence of a disease, and that one of the factors may play a determinant role, i.e., without it the other factors (as genetic susceptibility) are impotent. Hormones in breast cancer can play such a determinant role. The word cause is the one in general usage in connection with matters considered in this study, and it is capable of conveying the notion of a significant, effectual, relationship between an agent and an associated disorder or disease in the host.

4. It should be said at once, however, that no member of this Committee used the word cause in an absolute sense in the area of this study. Although various disciplines and fields of scientific knowledge were represented among the membership, all members shared

a common conception of the multiple etiology of biological processes. No member was so naive as to insist upon mono-etiology in pathological processes or in vital phenomena. All were thoroughly aware of the fact that there are series of events in occurrences and developments in these fields, and that the end results are the net effect of many actions and counteractions.

5. Granted that these complexities were recognized, it is to be noted clearly that the Committees considered decision to use the words a cause, or a major cause, or a significant cause, or a causal association in certain conclusions about smoking and health affirms their conviction."

Regarding age specific mortality, the Report concludes that although there is evidence of a statistical association, "The total number of excess deaths causally related to cigarette smoking in the U.S. population cannot be accurately estimated. In view of the continuing and mounting evidence from many sources, it is the judgment of the Committee that cigarette smoking contributes substantially to mortality from certain specific diseases and to the overall death rate." (Chapter 4) The available evidence suggests the existence of some morphological differences between smokers and non-smokers, but is too meager to permit a conclusion (Chapter 15, p. 387). This does not rule out physiological factors, especially in respect to habituation, nor the existence of predisposing constitutional or hereditary factors (Chapter 14, p. 377).

References

Auerbach, O. and L. Garfinkel (1970). "Effect Of Cigarette Smoking On Dogs. II. Pulmonary Neoplasms." *Archives of Environmental Health*, 21, 754–768.

Becker, G. S., M. Grossman, and K. M. Murphy (1994). "An Empirical Analysis of Cigarette Addiction." *American Economic Review*, 84(3), 396–418.

Becker, G. S. and K. M. Murphy (1988). "A Theory of Rational Addiction." *Journal of Political Economy*, 96(4), 675–699.

Broders, A. C. (1920). "Squamous-cell Epithelioma of the Lip." *Journal of the American Medical Association*, 74, 656–664.

Chaloupka, F. (1991). "Rational Addictive Behavior and Cigarette Smoking." *Journal of Political Economy*, 99(4), 722–742.

Chaloupka, F. (1999). "Do Higher Cigarette Prices Encourage Youth to Use Marijuana?" NBER Working Paper 6939.

DeCicca, P., D. Kenkel, and A. Mathios (2001). "Putting out the Fires: Will Higher Taxes Reduce the Onset of Youth Smoking?" *Journal of Political Economy*.

Decker, S. and A. Schwartz (2000). "Cigarettes and Alcohol: Substitutes or Complements?" NBER Working Paper 7535.

Dee, T. (1999). "The Complementarity of Teen Smoking and Drinking." *Journal of Health Economics*, 18(6), 769–793.

Doll, R. and A. B. Hill (1950). "Smoking and Carcinoma of the Lung. Preliminary report." *British Medical Journal*, ii, 739–748.

Doll, R. and A. B. Hill (1954). "The Mortality of Doctors in Relation to their Smoking Habits. A Preliminary Report." *British Medical Journal*, i, 1451–1455.

Doll, R. and A. B. Hill (1956). "Lung Cancer and Other Causes of Death in Relation to Smoking. A Second Report on the Mortality of British Doctors." *British Medical Journal*, ii, 1071–1076.

Doll, R. and R. Peto (1976). "Mortality in Relation to Smoking: 20 years' Observations on Male British Doctors." *British Medical Journal*, ii, 1525–1536.

Doll, R., R. Peto, K. Wheatley, R. Gray, and I. Sutherland (1994). "Mortality in Relation to Smoking: 40 years' Observations on Male British Doctors." *British Medical Journal*, 309(6959), 901–911.

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

Fisher, R. A. (1957a). "Dangers of cigarette smoking." *Brit med J*, 2, 43.

Fisher, R. A. (1957b). "Dangers of cigarette smoking." *Brit med J*, 2, 297–298.

Fisher, R. A. (1958a). "Cancer and smoking." *Nature*, 182, 596.

Fisher, R. A. (1958b). "Lung Cancer and Cigarettes?" *Nature*, 182, 108.

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

Hammond, E. C. (1966). "Smoking in Relation to the Death Rates of one Million Men and Women." *Natl Cancer Inst Monogr*, 19, 127–204.

Hammond, E. C. and D. Horn (1958). "Smoking And Death Rates. Part I. Total Mortality. Part II. Death Rates By Cause." *Journal of the American Medical Association*, 166, 1159–1172.

Hersch, J. (1996). "Smoking, Seat Belts and Other Risky Consumer Decisions: Differences by Gender and Race." *Managerial and Decision Economics*, 17, 471–481.

Hill, J. (1761). *Cautions Against the Immoderate Use of Snuff and the Effects It Must Produce When This Way Taken into the Body*. R. Baldwin and J. Jackson, London.

Hrubec, Z. and J. K. McLaughlin (1997). "Former Cigarette Smoking and Mortality Among Veterans: A 26-Year Followup, 1954 to 1980." In *Monograph 8: Changes in Cigarette-Related Disease Risks and Their Implications for Prevention and Control*, volume 8 of *Smoking and Tobacco Control Monographs*, chapter 7, pages 501–529. National Cancer Institute.

Hummer, R. A., C. B. Nam, and R. G. Rogers (1998). "Mortality Differentials Associated with Cigarette Smoking in the USA." *Population Research and Policy Review*, 17(3), 285–304.

Kawachi, I., G. A. Colditz, M. J. Stampfer, W. C. Willet, J. E. Manson, B. Rosner, D. J. Hunter, C. H. Hennekens, and F. E. Speizer (1993). "Smoking Cessation In Relation To Total Mortality Rates In Women. A Prospective Cohort Study." *Annals of Internal Medicine*, 119, 992–1000.

Kawachi, I., G. A. Colditz, M. J. Stampfer, W. C. Willett, J. E. Manson, B. Rosner, D. J. Hunter, C. H. Hennekens, and F. E. Speizer (1997). "Smoking Cessation and Decreased Risks Of Total Mortality, Stroke, and Coronary Heart Disease Incidence Among Women: A Prospective Cohort Study." In *Monograph 8: Changes in Cigarette-Related Disease Risks and Their Implications for Prevention and Control*, edited by D. M. Burns, L. Garfinkel, and J. M. Samet, volume 8 of *Smoking and Tobacco Control Monographs*, chapter 8, pages 531–565. National Cancer Institute.

Kenkel, D. S. (1991). "Health Behavior, Health Knowledge, and Schooling." *Journal of Political Economy*, 99(2), 287–305.

Lam, T. H., S. Y. Ho, A. J. Hedley, K. H. Mak, and R. Peto (2001). "Mortality and smoking in Hong Kong: case-control study of all adult deaths in 1998." *British Medical Journal*, 323, 1–6.

Lickint, F. (1935). "Der Bronchialkrebs der Raucher." *Munch Med Wschr*, 82, 122–124.

Lombard, H. L. and C. R. Doering (1928). "Classics in Oncology. Cancer Studies in Massachusetts Habits, Characteristics and Environment of Individuals with and without Cancer." *New England Journal of Medicine*, 198, 481–487.

Munafò, M. R., T. G. Clark, L. R. Moore, E. Payne, R. Walton, and J. Flint (2003). "Genetic Polymorphisms and Personality in Healthy Adults: A systematic review and meta-analysis." *Molecular Psychiatry*, 8(5), 471–484.

Norr, R. (1952). "Cancer by the carton." *The Reader's Digest*, 61, 7–8.

Ong, K., M. Preece, P. Emmett, M. Ahmed, and D. Dunger (2002). "Size At Birth And Early Childhood Growth In Relation To Maternal Smoking, Parity And Infant Breast-Feeding: Longitudinal Birth Cohort Study And Analysis." *Pediatric Research*, 52(6), 863–867.

Pearl, R. (1938). "Tobacco Smoking and Longevity." *Science*, 87, 2253–4.

Peto, R., S. Darby, H. Deo, P. Silcocks, E. Whitley, and R. Doll (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.

Phillips, A., G. Wannamethee, M. Walker, A. Thomson, and G. D. Smith (1996). "Life expectancy in men who have never smoked and those who have smoked continuously: 15 year follow up of large cohort of middle aged British men." *British Medical Journal*, 313, 907–908.

Rogers, R. G. and E. Powell-Griner (1991). "Life expectancies of cigarette smokers and nonsmokers in the United States." *Social Science and Medicine*, 32(10), 1151–1159.

Rosenzweig, M. and P. Schultz (1983). "Estimating a Household Production Function: Heterogeneity, the Demand for Health Inputs, and Their Effects on Birth Weight." *JPE*, 91(5), 723–746.

Shaw, M., R. Mitchell, and D. Dorling (2000). "Time for a Smoke? One Cigarette Reduces your Life by 11 Minutes." *British Medical Journal*, 320(7226), 53–54.

Smith, G. D. and M. J. Shipley (1991). "Confounding of occupation and smoking: its magnitude and consequences." *Social Science and Medicine*, 32(11), 1297–1300.

Sterling, T. and J. Weinkam (1990). "The Confounding of Occupation and Smoking and its Consequences." *Social Science and Medecine*, 30(4), 457–467.

Thun, M. J., L. F. Apicella, and S. J. Henley (2000). "Smoking Vs Other Risk Factors as the Cause of Smoking-Attributable Deaths: Confounding in the Courtroom." *Journal of the American Medical Association*, 284(6), 706–712.

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

Wynder, E. and E. Graham (1950). "Tobacco Smoking as a Possible Etiologic Factor in Bronchogenic Carcinoma." *Journal of the American Medical Association*, 143, 329–336.

Table 2: Descriptive Statistics

Variable	Men	Women	Total
Sample size	19176	20402	39578
Proportion in the sample	0.48	0.52	1
Smokers and former smokers	0.58	0.42	0.50
Smoking prevalence	0.26	0.24	0.25
Smoking prevalence under age 30	0.22	0.30	0.26
Smoking prevalence over age 50	0.25	0.14	0.19
Average smoking duration for former smokers (years)	17.91	11.73	15.56
Average quantities of cigarettes per day for smokers	14.71	12.67	13.71
Number of deaths by 1998	3291	3302	6593
Age	46.32	48.52	47.45
Proportion with low education	0.41	0.45	0.43
medium education	0.39	0.37	0.38
high education	0.19	0.18	0.19

Table 3: Variables Used to Construct the Tobacco-free Health Scores

Description (ICD9 code)	Score 1	Score 2	Score 3	Cases
adjusted adult height	X	X	X	39578
antibiotic prescription	X			1130
poliomyelitis (40-45)	X	X		55
herpes (53-55)	X	X		32
other infectious and parasitic diseases (1-139)	X			130
malignant neoplasm (140-240)[a]	X			851
endocrine, nutritional and metabolic diseases, and immunity disorders, excluding diabetes (240-280)	X			921
diabetes, type 1 (250)	X	X		136
diseases of the blood and blood-forming organs (280-290)	X			258
mental disorders (290-320)	X	X		1031
diseases of the nervous system and sense organs (320-390)	X			3288
pneumoconioses due to external agents (500-509)	X			23
hernia of abdominal cavity (550-554)	X	X		153
noninfective enteritis and colitis (555-560)	X	X		137
appendicitis, other diseases of intestines (540-544, 560-570)	X	X		194
other diseases of digestive system (570-580)	X			202
calculus (592-595)	X	X		181
urinary tract infection (599-600)	X	X		126
diseases of male genital organs (600-610)	X	X		184
inflammatory disease of female pelvic organs and other disorders of female genital tract (614-616)	X	X		94
amenorrhea (627)	X	X		60
menopausal and postmenopausal disorders (627)	X	X		140
hematocele (629)	X	X		35
psoriasis (696)	X			267
diseases of the musculoskeletal system (710-740)	X	X		6496
headache (784)	X			195
senility (797)	X	X		147
accidents (excluding fire due to smoking) (800-999)	X	X		2127

[a] excluding neoplasm of: lip, oral cavity pharynx (140-149); esophagus (150); pancreas (157); larynx (161); trachea, lung, bronchus (162); cervix uteri (180); urinary bladder (188);kidney, other urinary (189)

Table 4: Selection Into Smoking: Effect of Poor Tobacco-free Health versus Good Health.

	Score 1	Score 2	Score 3
Panel A, Odds ratio, Probability of Being a Smokers (current and former)			
Age>60	1.05 [0.84, 1.31]	1.02 [0.82, 1.27]	0.88 [0.70,1.10]
Age<50	1.19** [1.05, 1.34]	1.17** [1.03, 1.32]	1.07 [0.95,1.22]
Age<35	1.21** [1.03, 1.42]	1.20** [1.02, 1.41]	1.17* [0.99,1.36]
Age<25	1.36** [1.09, 1.67]	1.40** [1.12, 1.73]	1.25** [1.00,1.55]
Panel B, Odds ratio, Probability of Heavy Smoking, Conditional on Smoking.			
Age>60	0.88 [0.58 , 1.34]	0.82 [0.54 , 1.25]	0.78 [0.52, 1.17]
Age<50	1.30** [1.09 , 1.55]	1.30** [1.09 , 1.55]	1.07 [0.89, 1.27]
Age<35	1.27** [1.00 , 1.61]	1.37** [1.07 , 1.73]	0.99 [0.77, 1.25]
Age<25	0.99 [0.68 , 1.45]	1.10 [0.75 , 1.60]	0.83 [0.56, 1.21]
Panel C, Odds ratio, Probability of Heavy-Smoking, Unconditional on Smoking.			
Age$\in [25, 50]$	1.79** [1.31, 2.43]	1.82** [1.34, 2.46]	1.29* [0.96, 1.72]
Panel D, Odds ratio, Probability of Starting Smoking Before Age 15.			
Age<25	1.48** [1.07, 2.05]	1.47** [1.06, 2.05]	1.20 [0.86,1.69]
Panel E, Hazard ratio, Duration to Quitting.			
All Ages	0.87** [0.83, 0.92]	0.88**[0.82, 0.93]	0.88** [0.78,0.99]
Age<35	0.91** [0.86, 0.96]	0.92** [0.86, 0.98]	0.94 [0.84,1.07]

Note: * and ** denotes significance at the 10% and 5% level. For Panel A and B, the coefficients are obtained by a logistic regression of an indicator for ever smoker or for heavy smoking (more than a pack a day). For Panel C, the coefficients were obtained from a stratified Cox duration regression. Controls are age dummies, sex, education level, interview year effects, alcohol consumption, snus consumption and risky occupation. Robust standard errors were computed. 95% confidence intervals in brackets.

Table 5: Changes in Tobacco-free Morbidity and Smoking

	Score 1	Score 2
Smokers (current and former) compared to Never Smokers		
All Ages	0.18 (0.17)	0.15 (0.18)
Age>40	0.27 (0.23)	0.19 (0.24)
Effect of Quantities Smoked		
All Ages	0.019 (.014)	0.02 (0.015)
Age>40	0.026 (.019)	0.027 (0.021)
Effect of Duration of Habit, Conditional on Ever Smoker		
All Ages	0.008 (.011)	0.011 (0.012)
Age>40	.003 (.012)	0.005 (0.012)

Note: Robust standard errors were computed. Regressions control for age, sex and education levels.

Table 6: Smoking and Life Expectancy.

Sample	Controls	Hazard Ratio [Loss in Years]	
		Ever Smoker	Heavy Smoker
All	sex, educ.	1.31** [3.4]	1.85** [6.9]
All	sex, educ., health	1.30** [3.4]	1.83** [6.9]
Poor Health	sex, educ.	1.24** [3.0]	1.58** [5.5]
Good Health	sex, educ.	1.45** [4.5]	1.84** [6.8]
Poor Health, low Ed.	sex	1.18** [3.0]	1.27** [4.2]
Good Health, High Ed.	sex	1.44** [4.4]	2.04** [7.8]

Note: ** denotes significance at the 5% level. Heavy Smoker defined as smoking at least a pack a day. Loss (in brackets) is defined as the difference in life expectancy.

Table 7: Compensation for Foregone Earnings at Age 55 (Millions of US Dollars)

Tobacco-Free Health Type	Low Ed	Medium Ed	High Ed
		Men	
Good	.1165 [.085, .14]	.1770 [.13, .21]	.2705 [.20, .34]
Average	.0997 [.074, .11]	.1273 [.08, .16]	.2437 [.19, .29]
Bad	.0908 [.062, .11]	.1052 [.07, .13]	.1380 [.10, .16]
		Women	
Good	.0924 [.06, .11]	.1268 [.08, .17]	.1805 [.13, .21]
Average	.0905 [.06, .11]	.1347 [.08, .18]	.1900 [.14, .22]
Bad	.0784 [.05, .09]	.1009 [.06, .13]	.1651 [.12, .20]

Note: Numbers in brackets are computed using the 20 and 80% quantiles of life-cycle earnings.