**Health Spending: Luxury or Necessity?**

Evaluating Health Care Reforms Using an Estimated Model of Health Production Function

**Keyvan Eslami**†  
**Seyed M. Karimi**‡

January 2019  
Latest Version

**Abstract:** Aggregate health expenditures have risen more rapidly than aggregate income over the last five decades in the United States. This time series evidence suggests that the income elasticity of health care is well above one. In a cross section of households at a point in time, health expenditures do not vary with household income. This cross sectional evidence suggests that this elasticity is roughly zero. We reconcile these two conflicting patterns in a life-cycle model with heterogeneity in productivity and underlying health—health status—where individuals allocate resources between consumption and health spending. In our model, income growth over time causes an individual with given health status to devote more resources to health care to extend her life, since, under standard functional forms, the value of being alive relative to the value of consumption rises with income. If income and health status are strongly correlated and health status and health spending are substitutes, higher-income individuals spend less on health, because the marginal value of health care in extending life is lower for individuals in good health.

We estimate a structural life-cycle model using data from the Medical Expenditure Panel Survey and a novel computational method. The key parameters to be estimated are the value of being alive and the elasticity of substitution between health spending and health status. We find that health status and health spending are highly substitutable, implying that the returns to health expenditures are higher for low-income individuals relative to individuals in high-income groups.

We use the model to compute the welfare effects of two policy proposals: an extension of the current post-retirement Medicare program—which targets all income groups relatively equally—to all ages; and an expansion of the current Medicaid program—which subsidizes low-income families—to include more services. Assuming both policies are financed through an increased income tax, we show that “Medicare for all” entails a large welfare loss for high-income individuals by channeling resources from a period when they are old and unhealthy—when returns are relatively high—to a period when they are young and healthy—when returns are relatively low. “Medicaid expansion,” while designed to deliver the same benefits to low-income individuals, does so at a substantially lower cost to high-income groups because of its targeted nature.

**Keywords:** Health Spending, Health Status, Health Production Function, Indirect Inference, Statistical Value of Life.
1. Introduction

The rising share of health spending relative to income in the past half century in the United States—and in other developed countries—has led many economists to assert that health care is a luxury good, with an income elasticity well above one. However, in any cross section during this period, the income elasticity of health spending has been roughly zero, with no statistically significant difference between different income groups in health spending.

We develop a life-cycle model to reconcile these two patterns. In our framework, individuals are heterogeneous in their income and underlying health status (or health capital), and must allocate resources between health and non-health consumption. While consumption directly determines individuals’ utility, health spending and health status have an indirect effect on lifetime utility. In particular, health expenditures and health status determine individuals’ chance of mortality: higher health spending or health status means that the individual can enjoy consumption over a longer life span.

For a given health status, the growth in income leads to increases in health spending and consumption over time. In our framework, under standard functional forms, the increase in consumption implies a decline in marginal utility when normalized by average utility. The simultaneous rise in health spending also increases the marginal product of health spending relative to its average product. The resulting fall in the elasticity of utility with respect to

* We are grateful to V. V. Chari, Larry E. Jones, and Christopher Phelan, without whose help this project could not exist. Anmol Bhandari, Ellen McGrattan, and Charles Phelps deserve special thanks for their feedback and support. We are also grateful to Mariacristina De Nardi, Thomas J. Holmes, Naoki Aizawa, Elena Pastorino, Eric French, Brian Albrecht, and the participants of the Public Workshop at the University of Minnesota. The authors acknowledge the Minnesota Supercomputing Institute (MSI) at the University of Minnesota for providing resources that contributed to the results reported within this paper. You can find the latest version of this paper at keyvaneslami.com.

† Keyvan Eslami is a Ph.D. candidate at the University of Minnesota and a research analyst at the Federal Reserve Bank of Minneapolis. The views expressed in this paper are those of the author and not necessarily those of the Federal Reserve Bank of Minneapolis or the Federal Reserve System. The author can be reached at eslam010@umn.edu.

‡ Seyed M. Karimi is a lecturer at the University of Washington, Department of Politics, Philosophy, and Public Affairs. The author can be reached at skarimi2@uw.edu.
consumption relative to the elasticity of extending life with respect to health expenditures leads to an income elasticity of health care that is well above one in the time series.

In spite of the evidence supporting the cross-effects between underlying health and the productivity of health care, these effects have been largely ignored in the empirical and theoretical literature. By incorporating this consideration into our framework, we show that a strong correlation between health status and market productivity—an assumption that is supported by extensive literature—leads high-income individuals to allocate fewer resources to health care. This occurs because, in the presence of substitutability between health spending and health status, health expenditures are less effective in extending the life of healthier and wealthier people.

By addressing the patterns of health spending over time and in the cross section, our framework is consistent with the luxury-good channel that has been proposed as a cause for the rise in health spending over time. However, it also indicates that this channel is effective only to the extent that the pace of technological and income growth in the economy exceeds the growth rate of underlying health status of an average individual.

In addition, our framework has important insights for the literature that emphasizes the role of health technology as the main reason for the rise in health share. While we incorporate technological innovations as a contributing factor, our model implies that they cannot be the only cause for the rise in health spending. The reason is—at least from the perspective of a standard macroeconomic model—technological change entails a relative price change. The inelasticity of health spending in the cross section with respect to income suggests that the income effects of technological change are far more significant to allow for the observed dramatic rise in health spending over time, solely because of substitution effects.1

Many attempts have been made to estimate the relation between various measures of health outcome and health care utilization. Most of these attempts, however, ignore the possibility of cross-effects between the underlying health status and health spending. Even if that were not the case, one obstacle is finding an accurate measure of health status that

1. A similar argument applies to the role of health care policy over time. In the cross section, at least before retirement and except for the very bottom of the income distribution, the United States health care policy encourages more spending by higher-income individuals.
can convincingly address endogeneity.

To quantify our model, we take another approach to estimate this relationship. Instead of constructing a measure of health status, we use the insights from the model to infer the structural parameters of the model from variations in income over time and across individuals. This is done using the Medical Expenditure Panel Survey (MEPS) data, by adopting a simulation-based estimation method, and by employing a novel computational technique to solve the model—namely, the Markov chain approximation method.

Our results suggest that health status and health spending are relatively strong substitutes, though this substitutability declines with age. We use these results to compute the cost of saving a statistical life. While these costs are comparable to the estimated values of statistical life in the literature for a median agent at different ages, they are considerably higher for the top earners in our sample.

Our findings have important implications for health care policy. To show this, we use our estimates to compute the welfare implications of two policy proposals for different income groups: (i) an extension of the post-retirement United States health care policy—which subsidizes health spending at all income levels, though at different rates—to all ages; and, (ii) an expansion of the pre-retirement policy—which targets and subsidizes lower-income individuals—to deliver the same level of welfare to the low-income households as the first policy reform, leaving the high-income households as before. With a slight abuse of terminology, and for lack of better terms, we will refer to these proposals as Medicare for all and Medicaid expansion, respectively. Each policy is financed through an increased income tax rate.

Our simulations show that Medicare for all has a large and positive welfare impact at the bottom of the income distribution. Nonetheless, the welfare gains diminish quickly because of the increased income tax rate, disappearing entirely at the 17th income percentile. The impact is negative and considerable at the top of the income distribution. In comparison, the positive impact of Medicaid expansion become zero at the 14th percentile.

Importantly, the negative impact of Medicaid expansion is significantly smaller for the

---

2. After all, neither Medicaid nor Medicare are the only government insurance programs in the United States. Nevertheless, they are the largest of their kind, before and after retirement, respectively.
top income groups, compared to Medicare for all. The intuition, based on our model, is that, while Medicare for all subsidizes the health expenditures of high-income individuals when they are young, it does so at the expense of considerably higher income taxes: a 6 percentage point increase in income tax rate for Medicare for all compared to 0.8 percentage points for Medicaid expansion. The increased health spending, however, does little to increase the probability of survival for this group, as suggested by the considerable cost of saving a life for them: individuals in this group are healthy, especially when young, and have no urgent needs for health care. Nevertheless, the increased income tax causes them to allocate fewer resources to health spending when they are older and have more health care needs. Our simulations show that, in total, life expectancy declines for this group of people after the policy implementation.

In what follows, after providing a brief literature review, in Section 2, we lay out the full economy and characterize its equilibrium. This is the model that we will eventually bring to the data under standard assumptions for the functional forms. Using a simplified version of this economy, we discuss the primary mechanisms that enable the model to account for the different patterns of health spending in the cross section and over time in Section 3, and how these mechanisms can be used to infer the structural parameters of the full model. In Section 4, we explain the details of our quantitative method, before presenting our results in Section 5. We discuss the implications of these results for health care policy in the United States in Section 6. Section 7 concludes.

A Review of the Literature

The rapid rise in the share of health expenditures relative to income and the downward-sloping Engel curve in the cross section in the past five decades have separately been documented by many researchers before us, both in the United States and in other developed countries. Examples include Hall and Jones (2007), Ales, Hosseini, and Jones (2014), Ozkan (2014), French and Kelly (2016), Dickman et al. (2016), and Dickman, Himmelstein, and
Woolhandler (2017), among others. However, this study is the first attempt to address both observations simultaneously.

From a modeling perspective, our paper is another step in a long line of literature going back to Grossman (1972)’s seminal work in introducing health capital as an important determinant in individuals’ utility. It is closely related to papers such as Ehrlich and Chuma (1990), Fonseca et al. (2009), Scholz and Seshadri (2011), Hugonnier, Pelgrin, and St-Amour (2013), Ozkan (2014), and Ales, Hosseini, and Jones (2014), who model individuals’ life-cycle health spending. Foremost, this paper builds on Hall and Jones (2007)’s idea that changes in individuals’ valuations of the quality versus quantity of life is an important driving force in the observed rise in health expenditures over time. Our study extends this framework to address a flat Engel curve in the cross section, in addition to a rising share of health spending over time. It does so by introducing heterogeneity into a decentralized economy and by relaxing Hall and Jones’s assumption that the cross-elasticity of health outcomes with respect to health spending and health status—that is, “other factors” in Hall and Jones—is zero.

From an empirical standpoint, this paper is related to a vast literature that measures the relation between various measures of health outcome and health care utilization—namely, a health production function—such as Newhouse and Friedlander (1980), Brook et al. (1983), Finkelstein et al. (2012), and Baicker et al. (2013). Nevertheless, it departs from this strand of literature in two important ways. First, we explicitly allow for the cross-elasticity of health outcomes with respect to health spending and underlying health to be non-zero. Brook et al. (1983) is among the very few papers in this literature that consider such possibilities. Second, while most of this literature uses standard estimation techniques to measure the impact of health care on outcomes, we take an indirect approach. We use a structural model to estimate the health production function through the use of the indirect inference method.

3. In another paper, the authors document the cross sectional differences in health care spending among different income groups based on the type of services (Eslami and Karimi 2018a).
4. See Freeman et al. (2008) and Levy and Meltzer (2008) for excellent reviews.
5. See Eslami and Karimi (2018b) for an example in which these cross-effects are explicitly incorporated into an instrumental variable (IV) estimation.
and an auxiliary model.

From the standpoint of its empirical methodology, this paper uses the structural estimation method proposed by Smith (1990, 1993) and developed further by Gourieroux, Monfort, and Renault (1993). It is closely but indirectly related to studies such as Guvenen and Smith (2010) that, instead of using simplifying assumptions for the sake of empirical tractability, take an indirect approach toward statistical inference.

Finally, this paper is indirectly related to a literature that studies the relationship between health outcomes—such as self-reported health status or longevity—and income and other socio-economic factors. Some of the important works in this literature are Adler et al. (1994), Backlund, Sorlie, and Johnson (1996), Ettner (1996), Deaton and Paxson (1998), Adler and Ostrove (1999), and Kawachi and Kennedy (1999). As noted by Smith (1999), this relationship is complex and multilateral, and its study calls for the use of theoretic models. Our paper is an example of such models.

2. The Full Model

Time is continuous and infinite, denoted by $t$. At each date $t$, a new cohort of individuals enters the economy. The individuals’ age, denoted by $a$, is $a = a$ upon entry. Agents are identified by their entry cohort and live up to $\bar{a} = a + T$.

Individuals of a single cohort are heterogeneous in terms of their initial health status, $h_0 \in \mathcal{H} \subset \mathbb{R}_+$. We will denote the initial distribution of health status in cohort $t_0$ by the

7. We allow $a$ to be non-zero mainly to be consistent with the existing literature on mortality at certain ages in our quantitative exercise.
8. At each date $t$, the individual’s age and cohort of entry are related according to $t_0 = t - (a - a)$.
9. In an extension of this model, we allow individuals of a cohort $t_0$ to be heterogeneous in terms of an idiosyncratic productivity shock, $\nu_0 := \nu (t_0)$, distributed according to $\Phi (\cdot, t_0)$. These shocks are assumed to affect income, as will be discussed later on. While this extension is conceptually important, especially to examine the predictions of the model for temporary income shocks, the inclusion of $\nu$ is extremely costly from a numerical perspective. In addition, we lack reliable data to discipline these shocks. As a result, in our quantitative exercise, we limit ourselves to a reasonable range for $\theta_{\nu}$ and $\sigma_{\nu}$. Our estimates do not reflect
measure $\Gamma (\cdot, t_0)$ over $\mathcal{H}$. An individual’s health status at time $t$ evolves according to a geometric Brownian motion, as

$$d \ln (h(t)) = g(h(t), a) \cdot dt + \sigma_h \cdot d\omega_h(t),$$

(1)

where $\omega_h(\cdot)$ is a Brownian motion. We will refer to $g(\cdot)$ in (1) as the depreciation function, even though there is no assumption in (1) to indicate that health status cannot accumulate over time.

Individuals retire at age $a^R \in (a, \bar{a})$. Before retirement and at time $t$, an individual with health status $h(t)$ earns a flow of income given by $y(h(t), a, t)$. After retirement, income is a constant function of income at the age of retirement,

$$\phi\left(y\left(h\left(t^R\right), a^R, t^R\right), t^R\right),$$

where $t^R := t - (a - a^R)$ is the time of retirement (following Guvenen and Smith, 2010). With a slight abuse of notation, we summarize these by an income equation of the following significant changes as a consequence of their addition to the model. As a result, instead of modifying the benchmark model to incorporate them, we will only briefly mention, in the footnotes that follow, the major modifications that are needed to incorporate $\nu$.

10. Note that $\Gamma (\cdot)$ need not be a probability measure. If so, it implicitly incorporates the variations in the birthrate over time.

11. Let us assume that $(\Omega, \mathcal{F}, P)$ is a probability space with a filtration $\{\mathcal{F}_t, t \in [0, \infty)\}$ defined on it. For the sake of consistency, by a stochastic process we henceforth mean a set of random variables, $x : [0, \infty) \to \mathbb{R}^k$, defined over this probability space, such that for each $t \in [0, \infty)$, $x(t)$ is $\mathcal{F}_t$-measurable.

By a Brownian motion $\omega(\cdot)$ we refer to a $\mathcal{F}_t$-Wiener process. Naturally, a Wiener process is assumed to have continuous sample paths; that is for each outcome in $\Omega$, $\omega(t)$ is a continuous function of $t$, for all $t \in [0, \infty)$.

12. This formulation of health shocks is consistent with Deaton and Paxson (1998).
form: \( y(h(t), h^R(t), a, t) = \begin{cases} \ y(h(t), a, t) & \text{if } a \in [a, a^R], \\ \phi(y(h^R(t), a^R, t^R), t^R) & \text{if } a \in [a^R, \bar{a}] \end{cases} \) \( (3) \)

where \( h^R(t) \) is the health status at the time of retirement. We will say more about this in the sections that follow. Importantly, this formulation allows for individuals’ income profiles to change over time.

Individuals can save their income or allocate it between health and non-health spending—\( m \) and \( c \), respectively. Individuals’ flow utility from consumption \( c \) is specified by the utility function \( u(c) \). Agents discount the future at rate \( \rho \), and we normalize their utility upon death to zero, \( V^d = 0 \).

At each age, individuals face an endogenous chance of mortality. We model mortality as the first jump of a Poisson process with intensity \( 1/\chi \). (See Hugonnier, Pelgrin, and St-Amour 2013 for a detailed discussion.) The variable \( \chi \) depends on individuals’ health status and health spending. Specifically, at any date \( t \), given agents’ health status, \( h(t) \), and health spending, \( m \), \( \chi \) at age \( a \) is characterized by a health production function as

\[ \chi = f(m, h(t), a, t). \]  \( (4) \)

Note that the production of health at any age can change with technological innovations.

Markets are incomplete in the sense that individuals can only save in a risk-free saving

---

13. With idiosyncratic productivity shocks, pre-retirement income is assumed to be a function of \( \nu(t) \), besides health status. Productivity shocks are assumed to evolve according to an Ornstein-Uhlenbeck process of the form

\[ d\nu(t) = -\theta_{\nu} \cdot \nu(t) \cdot dt + \sigma_{\nu} \cdot d\omega_{\nu}(t), \]  \( (2) \)

where \( \omega_{\nu}(\cdot) \) is a Brownian motion.

14. As we will discuss in the next section, this interpretation of the health production function is very narrow. While, in theory, one can interpret \( f(\cdot) \) as a determinant of the marginal utility of consumption and its level, in practice we need an interpretation that allows for the quantitative identification of \( f(\cdot) \). That is why we restrict ourselves to the current definition of the health production function as the determinant of the survival rate.
technology with a fixed rate of return \( r \). No borrowing is allowed, and upon death, individuals’ savings are destroyed.\(^{15}\) We will denote individuals’ asset (physical capital) holdings by \( k(t) \) and assume that the initial asset holdings are zero for all the individuals of each cohort, \( k_0 = 0. \(^{16}\)

Policy in this economy is characterized by an income tax and a subsidy on health expenditures. In particular, at each date \( t \), the government is assumed to tax income at the fixed rate \( \tau(t) \). Depending on their income level and age, individuals face a subsidy rate of \( s(y, a) \) on their health spending. Consolidating policy as a single rate of subsidy—which depends on income and age—allows us to capture in a stylized way the relatively complicated and segmented health care policy in the United States.\(^{17,18}\)

**Individual’s Problem** At any given time \( t \), besides her age \( a \), an individual’s state consists of her health status, \( h(t) \), health status at retirement, \( h^R(t) \), asset holdings, \( k(t) \), and mortality, \( \iota(t) \in \{0, 1\} \)—where \( \iota(t) = 1 \) indicates death.\(^{19}\)

It is worth emphasizing that individuals’ income after retirement is a function of their health status at the age of retirement. To be able to restrict our attention to feedback control rules (to be discussed in a moment), including health status at retirement as an individual

\(^{15}\) We can think of this environment as an economy with international lenders who confiscate agents’ deposits upon death. Without altruistic motives, neither of these settings has an impact on our results.

Alternatively, one can assume that international capital markets are competitive. As a result, the equilibrium rate of return is the break-even rate, determined endogenously as a function of the distribution of the mortality rate. An endogenous rate of return ensures that the capital accounts will balance in the equilibrium.

\(^{16}\) This assumption is consistent with the no-bequest assumption made earlier.

\(^{17}\) The health care policy in the United States is complicated, and a discussion of all of its different facets calls for a separate study. However, as we will discuss in more detail later on, a single rate of subsidy on health expenditures does a relatively good job in consolidating this complicated system for our purposes.

\(^{18}\) An important provision in the United States tax code—that is missing from our model—is the deductibility of employer provided health insurance from income tax. This policy encourages higher income individuals to spend more on health (Chari and Eslami 2016), and its absence in our model potentially leads to an underestimation of the substitutability of health spending and health status.

\(^{19}\) We find it constructive to think of \( t \) as an aggregate state variable and \( a \) as an individual state. In addition, this distinction helps with the notational brevity. In practice, we use cohort of entry and age as the aggregate states for a cohort of individuals.
state is best. Of course, for the individual state to be adapted to the same filtration as $\omega_h (\cdot)$, $h^R (\cdot)$ cannot be anticipative. To avoid this, we will assume $h^R (t) = h (t)$ when $t < t^R$, and $h^R (t) = h (t^R)$ when $t \geq t^R$. Formally,

$$
\frac{d \ln (h (t))}{dt} = g^R (h (t), a) \cdot dt + \sigma^R_h (a) \cdot d\omega_h (t),
$$

(5)

where

$$
g^R (h (t), a) = \begin{cases} 
g (h (t), a) & \text{if } a \in [a, a^R), \\
0 & \text{if } a \in [a^R, a] \end{cases},
$$

(6)

and

$$
\sigma^R_h (a) = \begin{cases} 
\sigma_h & \text{if } a \in [a, a^R), \\
0 & \text{if } a \in [a^R, a] \end{cases}.
$$

(7)

At each date $t$, we are going to summarize the individual’s states in an individual state vector of the form

$$
x (t) := [a, k (t), h (t), h^R (t), \iota (t)]',
$$

and denote the domain of $x$ by $X$. We will reserve $x_0$ for the individual’s initial state:

$$
x_0 := [a, k_0 = 0, h_0, h_0, \iota = 0]^'.
$$

Then, for individuals of cohort $t_0$, $x (t_0) = x_0$ almost surely.

If we let $U := \mathbb{R}^2_+ \ni (c, m)$, an individual control $u (\cdot) = (c (\cdot), m (\cdot))$ is a $U$-valued stochastic process that is admissible with respect to $\omega_h$.\(^{21}\) In this paper, we are going to

20. In our numerical exercise, we divide the individual’s problem into two periods: before and after retirement. This eliminates one of the state variables (namely, $h^R$) before the age of retirement, decreasing the computational burden to some extent.

21. The stochastic process $u (\cdot)$ is said to be admissible with respect to $\omega_h (\cdot)$ if there exists a filtration, $\mathcal{F}_t$, defined over the probability space $(\Omega, \mathcal{F}, P)$ such that $u (\cdot)$ is $\mathcal{F}_t$-adapted and $\omega_h (\cdot)$ is a $\mathcal{F}_t$-Wiener process. If so, $u (\cdot)$ is called non-anticipative with respect to $\omega_h (\cdot)$.

Adapting this definition to incorporate a vector-valued Wiener process (for when productivity shocks are present) is straightforward.
focus on pure Markov controls\textsuperscript{22} of the form

\[
  u : \mathcal{X} \times [0, \infty) \to \mathcal{U}. \textsuperscript{23}
\]

Then, at any given date \( t \), the law of motion of \( k \) under a feedback rule \( u = (c, m) \) is

\[
  \dot{k}(t) = r \cdot k(t) + \left[ 1 - \tau(t) \right] \cdot y(h(t), h_R(t), a, t) \\
  - c(x(t), t) - \left[ 1 - s \left( y(h(t), h_R(t), a, t), a \right) \right] m(x(t), t) \\
  =: q(x(t), t, u). \quad (8)
\]

No-borrowing constrained is modeled as a reflecting barrier at \( k = 0 \).

For an individual of cohort \( t_0 \), given the initial state \( x_0 \), the evolution of individual state \( x \), under an admissible control \( u \) is given by the following controlled jump-diffusion process:

\[
  dx(t) = \begin{bmatrix}
  1 \\
  q(x(t), t, u) \\
  g(h(t), a) \\
  g^R(h(t), a) \\
  0
  \end{bmatrix} dt \\
  + \begin{bmatrix}
  0 & 0 & \sigma_h & \sigma^{R}(a) & 0 \\
  0 & 0 & 0 & 0 & 1
  \end{bmatrix}' d\omega_h(t) \\
  + \begin{bmatrix}
  0 & 0 & 0 & 0 & 1
  \end{bmatrix}' dJ(x, t, u) \\
  =: b(x, t, u) da + \Sigma(a) dw(t) + \Pi dJ(x, t, u), \quad (9)
\]

subject to \( x(t_0) = x_0 \) almost surely. In this equation, \( J(\cdot) \) is a jump process whose intensity \( 1/\chi \) is defined by (4).\textsuperscript{24} We will refer to \( b(\cdot) \) in (9) as the drift vector. \( D(x) := \Sigma(a) \Sigma'(a) / 2 \)

\textsuperscript{22} Pure Markov or feedback controls are the controls that are only functions of the current state and time. It is easy to see that such controls are admissible with respect to any Wiener process. One can show that restricting attention to the feedback class of controls is without any loss of generality for the problem at hand.

\textsuperscript{23} While \( u \) maps \( \mathcal{X} \times \mathbb{R}_+ \) to \( \mathcal{U} \), in practice only the fraction of the control process over individual’s lifetime is of interest to us.

\textsuperscript{24} More precisely, \( J(\cdot) \) is characterized by a Poisson random measure adapted to the same filtration as \( \omega_h \).
is known as the diffusion tensor.\textsuperscript{25}

Given an admissible control $u$, let $\varrho_{t_1}^u$ be the random variable characterizing the first jump of $J(x, t, u)$, conditioned on no jumps before time $t_1$:

$$\varrho_{t_1}^u := \inf \{ t : \iota(t) = 1 \mid \iota(t_1) = 0 \}.$$  

At age $a_1$ and starting from the state

$$x_1 := x(t_1) = [a_1, k_1, h_1, h_R^1, \iota(t_1) = 0]^\prime,$$

an individual’s expected discounted utility, under the admissible control $u$, is given by

$$W(x_1, t_1, u) := \mathbb{E}_{x_1}^u \left[ \int_{t_1}^{\bar{t} \wedge \varrho_{t_1}^u} e^{-\rho(t-t_1)} \cdot u(c(x(t), t)) \cdot dt + e^{-\rho(\bar{t} \wedge \varrho_{t_1}^u - t_1)} \cdot V^d \right],$$

where $\bar{t} := t_1 + T - (a_1 - a)$ and $\mathbb{E}_{x_1}^u [\cdot]$ represents the expectations with respect to the process governing $x$ (Equation (9)) under the feedback control $u$, assuming $x(t_1) = x_1$.

Under the assumption that $J(\cdot)$ is governed by a Poisson random measure whose intensity is given by $1/\chi$, the random variable $\varrho_{t_1}^u$ has exponential distribution with density (and $\omega_{\nu}$, when present).

\textsuperscript{25} In the presence of productivity shocks, $x(\cdot)$ has an additional term:

$$x(t) = [a, k(t), h(t), h_R(t), \nu(t), \iota(t)]^\prime.$$

Then, we need to modify the drift vector, diffusion tensor, and the vector of Brownian shocks as follows:

$$b(x, t, u) = \begin{bmatrix} 1 \\ q(x(t), t, u) \\ g(h(t), a) \\ g^R(h(t), a) \\ -\theta_{\nu}\nu(t) \\ 0 \end{bmatrix},$$

$$\Sigma(a) = \begin{bmatrix} 0 & 0 & \sigma_h & \sigma_h^R(a) & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \end{bmatrix}^\prime,$$

and

$$w(t) = \begin{bmatrix} \omega_h(t) \\ \omega_{\nu}(t) \end{bmatrix}.$$
\[ \exp \left( -\frac{\varrho}{\chi} \right) / \chi. \] When \( V^d = 0 \), Equation (13) can be simplified as

\[ W(\mathbf{x}_1, t_1, \mathbf{u}) = \mathbb{E}_{\mathbf{x}_1}^u \left[ \int_{t_1}^T e^{-\rho(t-t_1)} \cdot e^{-\int_{t_1}^t \frac{1}{\chi(\ell)} d\ell} \cdot u(\mathbf{c}(\mathbf{x}(t), t)) \cdot dt \right], \tag{14} \]

where

\[ \chi(t) = f(m(\mathbf{x}(t), t), h(t), a, t). \]

Starting from any individual state \( \mathbf{x}_1 \) at time \( t_1 \), an individual chooses an admissible control \( \mathbf{u} \) to maximize her expected discounted utility, given by (14). If we denote the individual’s value at \( \mathbf{x}_1 \) by \( V(\mathbf{x}_1, t_1) \), this value is given by

\[ V(\mathbf{x}_1, t_1) = \sup_{\mathbf{u}} W(\mathbf{x}_1, t_1, \mathbf{u}), \tag{15} \]

where the optimization is over the set of all feedback control rules.

Writing an individual’s lifetime utility as in Equation (14) allows us to dispense with \( \iota \) as an individual state.\(^{26}\) With some abuse of notation, we will use \( \mathbf{x} \) to denote the individual’s state vector, absent mortality, and let \( \mathcal{X} \) denote the corresponding (new) state-space. Then, under the assumption that the function \( V(\cdot) \) is smooth enough, one can show that the individual’s value function satisfies the partial differential equation known as Hamilton-Jacobi-Bellman (HJB) equation.\(^{27,28}\)

**Proposition 1** For any individual state \( \mathbf{x} \in \mathcal{X} \) at date \( t \), the individual’s value function

---

26. This also means we can think of health spending and health status, more broadly, as determinants of lifetime utility. Equation (13) does not allow for such broad interpretation. We will talk more about this in the following sections.

27. See the supplementary appendix for a heuristic derivation of Equation (16).

28. Even under standard functional forms for the utility and health production functions, we cannot be sure that the optimization problem on the right hand side of the HJB equation is concave. Nevertheless, in our numerical results of Section 4, the problem always seems to have an interior solution and the resulting value function is concave and differentiable everywhere. Even in the absence of such well behaved solutions, one can argue that the individual’s value is the viscosity solution of Equation (16).
solves the Hamilton-Jacobi-Bellman equation,
\[
- \frac{\partial V(x(t), t)}{\partial t} = \sup_{(c,m) \in U} \left\{ u(c) - \left[ \rho + \frac{1}{f(m, h(t), a, t)} \right] V(x(t), t) \right. \\
\left. + \frac{\partial V(x(t), t)}{\partial a} + \left[ rk(t) + [1 - \tau(t)] y(h(t), h^R(t), a, t) \right] \frac{\partial V(x(t), t)}{\partial k} \\
- c - \left[ 1 - s(y(h(t), h^R(t), a, t), a) \right] m \frac{\partial V(x(t), t)}{\partial k} \right. \\
\left. + g(h(t), a) \frac{\partial V(x(t), t)}{\partial \ln(h)} + g^R(h(t), a) \frac{\partial V(x(t), t)}{\partial \ln(h^R)} \right. \\
\left. + \frac{1}{2} \sigma_h^2 \frac{\partial^2 V(x(t), t)}{[\partial \ln(h)]^2} + \frac{1}{2} \left[ \sigma_h^R(a) \right]^2 \frac{\partial^2 V(x(t), t)}{[\partial \ln(h^R)]^2} \right\}, \tag{16}
\]

subject to the boundary value \( V(x, t) = V^d \) when \( a \geq \bar{a} \) and the smooth pasting condition,
\[
\left. \frac{\partial V(x, t)}{\partial k} \right|_{k=0} = 0. \tag{17}
\]

In addition, under the assumption that an optimal admissible control exists such that
\[
V(x, t) = W(x, t, \tilde{u}), \tag{18}
\]
then \( \tilde{u}(t) = (\tilde{c}, \tilde{m}) \) is a solution to the optimization problem on the right hand side of \( (16) \).

To characterize the distribution of individual states, let \( p(x, t, u) \) denote the probability of being alive and in state \( x \) at time \( t \), under the admissible control \( u \). The dynamic of \( p(\cdot) \) is determined by the stochastic process governing the individual state, Equation \( (9) \), under the feedback rule \( u \). One can show \( p(\cdot) \) evolves according to a partial differential equation known as the Kolmogorov's forward (KF) equation (or Fokker-Plank equation), as stated in the following proposition.

**Proposition 2** Given the diffusion process governing \( x \)—Equation \( (9) \)—starting from any

29. Except for the probability of jumps, Equation \( (19) \) is a standard Fokker-Plank equation. Heuristically, given the Poisson random measure governing the jumps, the probability of mortality in each infinitesimal
initial distribution of individual states at time \( t_1 \), namely \( p_1 (\cdot) \) over \( \mathcal{X} \), the probability of being alive and in state \( x \) at time \( t \) is a solution to the Kolmogorov’s forward equation, given by

\[
\frac{\partial p(x, t, u)}{\partial t} = - \sum_{i=1}^{4} \frac{\partial}{\partial x_i} [b_i (x, t, u) \cdot p(x, t, u)]
+ \sum_{i=1}^{4} \sum_{j=1}^{4} \frac{\partial^2}{\partial x_i \partial x_j} [D_{i,j} (x) \cdot p(x, t, u)]
- \frac{1}{f (m(x(t), t), h(t), a, t)} \cdot p(x, t, u), \quad (19)
\]

subject to the boundary condition

\[
p(x_1, t_1, u) = p_1 (x_1), \quad \forall x_1 \in \mathcal{X}.
\quad (20)
\]

In Equation (9), \( b_i \) and \( D_{i,j} \)’s are the components of the drift vector and diffusion tensor associated with \( x \).\(^{30}\)

Using Proposition 2, one can derive the probability of moving from state \( x_1 \) at date \( t_1 \) to \( x_2 \) at \( t_2 \)—under the feedback rule \( u \)—by finding the solution to KF equation subject to the boundary condition \( p_1 (x) = \delta (x_1) \) at time \( t_1 \), where \( \delta (x_1) \) is the Dirac delta function with unit point mass at \( x_1 \). For the future use, let us denote this transition probability by \( \vartheta (x_1, t_1, x_2, t_2, u) \).

interval of length \( dt \) is given by

\[
dt / f (m(x(t), t), h(t), a, t) + o(dt).
\]

When \( dt \to 0 \), the change in the measure of individuals who are alive and in state \( x \) in \( t + dt \) should be adjusted to incorporate the fraction of people who die during \( dt \). This is the intuition behind the last term in (19). The rigorous derivation, however, is rather cumbersome. Interested reader can refer to Hanson (2007).

30. It is implicitly assumed \( x \) is such that \( a \leq \bar{a} \).
The Evolution of Physical Capital  The distribution of states among individuals of each cohort, together with the feedback rule \( u(\cdot) = (c(\cdot), m(\cdot)) \), determine the evolution of aggregate (average) physical capital in the economy as follows:

\[
\dot{K}(t) = \int_{t-T}^{t} \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[ rk + [1 - \tau(t)] y(h, h^R, a, t) - c(x, t)\right]
- [1 - s(y(h, h^R, a, t), a)] m(x, t) ]
\times \vartheta(x_0, \ell, dx, t, u) \Gamma(dh_0, \ell) d\ell, \tag{21}
\]

where \( x_0 = [a, 0, h_0, h_0]' \) and \( x = [a, k, h, h^R] \).

Government's Budget  Government is assumed to run a period-by-period balanced budget. For a given feedback rule, we can write government’s budget constraint in date \( t \) as

\[
\int_{t-T}^{t} \int_{\mathcal{X}} \int_{\mathcal{Y}} \left[ s(y(h, h^R, a, t), a) \cdot m(x, t) - \tau(t) \cdot y(h, h^R, a, t) \right]
\times \vartheta(x_0, \ell, dx, t, u) \Gamma(dh_0, \ell) d\ell = 0. \tag{22}
\]

Recursive Equilibrium  Without a supply sector and with an exogenous rate of return, the notion of equilibrium in this economy is rather mechanical. Nevertheless, we formalize this notion in Definition 1 for the sake of completeness.

**Definition 1**  A recursive equilibrium of the economy of Section 2 consists of a value functions \( \hat{V} \), a corresponding admissible control \( \hat{u} \), and a probability kernel \( \hat{\vartheta} \), such that, given the policies \( \tau \) and \( s \),

(i) for each \( x \in \mathcal{X} \) and \( t \in [0, \infty) \), \( \hat{V}(x, t) \) solves the Hamilton-Jacobi-Bellman equation and \( \hat{u} \) is the corresponding optimal feedback rule;

(ii) for any \( x_1, x_2 \in \mathcal{X} \) and \( t_1, t_2 \in [0, \infty) \), \( \hat{\vartheta}(x_1, t_1, x_2, t_2, \hat{u}) \) is the solution to Kolmogorov’s forward equation under the boundary condition \( p_1(x) = \delta(x_1) \) at date \( t_1 \);

(iii) average physical capital, evolving according to (21) under the admissible control, satisfies \( K(t) \geq 0 \), for all \( t \); and
(iv) government runs a balanced budget under the admissible control.

If it exists, the recursive equilibrium of this economy is fully characterized by the HJB and KF equations. Nevertheless, the partial differential equation governing individuals’ value functions and optimal controls is too complicated to be solved analytically. For this reason, we propose a quantitative method to solve the HJB and KF equations numerically. These solutions, then, can be used to make inferences about the important structural parameters of the economy.

Before doing so, we find it useful to discuss the mechanisms in this economy that will deliver a declining schedule for health spending among income groups in the cross section, while implying a sharply upward sloping Engel curve in the time series. To do so, in the next section, we will simplify the full model by abstracting from the aging of the agents. This simplification will help us write the individuals’ problem as a stationary one whose solution is considerably easier to find. We will use this simplified model to discuss the main mechanisms of the model and how they will help us identify the parameters of interest in our estimation exercise.

3. A Simplified Economy

Consider the economy of Section 2 and assume individuals of a given cohort $t_0$ can live forever without retiring—that is $a^R, T \to \infty$. In addition, suppose individuals’ initial health status remains constant while alive—so that $\sigma_h = 0$, $g(\cdot) = 0$—and, for simplicity, they weight the future the same way they value today, $\rho = 0$. Also, to abstract from the saving decisions, let us assume $r \to -\infty$. Moreover, suppose $y(h,a,t_0) = y(h,t_0)$ and $f(m,h,a,t_0) = f(m,h)$ for all $a$, so that the income equation and production of health are independent of age. To be able to focus only on the important mechanisms of the model, let us simplify the economy even more by abstracting from the effects of policy and

31. One can come up with government policies under which no such equilibria exist.

32. The assumption that the health production function is independent of time eliminates the possibility of technological progress. In this section, we can dispense with this simplification. However, it is an important part of the full model.
assuming \( \tau (t) = s(y, a) = 0 \).

Under these assumptions, the individual’s state—absent her state of mortality, as assumed in Section 2—is going to remain constant over time. As a result, the problem of an individual of cohort \( t_0 \) with health status \( h_0 \) can be written simply as

\[
\max_{c(\cdot), m(\cdot)} \int_{t=t_0}^{\infty} e^{-\int_{t_0}^{t} \frac{1}{f(m(t), h_0)} dt} u(c(t)) \, dt \quad \text{s.t.} \quad c(t) + m(t) = y(h_0, t_0). \tag{23}
\]

It is easy to see that, with a constant state vector, the individual chooses the same level of consumption and health spending at all dates (assuming such an optimum is unique). Hence, the solution to (23) coincides with that of the following static problem:

\[
\max_{c, m} f(m, h_0) u(c) \quad \text{s.t.} \quad c + m = y(h_0, t_0). \tag{24}
\]

By writing the individual’s problem as (24), we can see that health production has a broader interpretation than the determinant of longevity. While one certainly can construe the first term in the objective function of Problem 24 as individual’s quantity of life (as Hall and Jones 2007 note, in contrast to the quality of life, determined by the second term in (24)), Problem (24) allows for a broader interpretation of health as a factor determining the marginal utility of consumption. These readings include the standard argument regarding the dependency of utility on the state of health (see Finkelstein, Luttmer, and Notowidigdo 2013, as an example) and health as a determinant of life-years adjusted for the burden of diseases (see Eslami and Karimi 2018b for a discussion).

In addition to shedding light on the meaning of health and health production, an advantage of writing the individual’s problem as a static one is that it allows for comparative static exercises that can clarify the channels through which health spending displays the characteristics of a luxury good over time and an absolute necessity in the cross section.
A Luxury over Time  Note that the optimal share of health spending in income in Problem (24) is given by the following condition:

\[
\left( \frac{s^*}{1 - s^*} \right) = \frac{\partial f (m^*, h_0)}{\partial m} \frac{m^*}{c^*} \frac{\partial u (c^*)}{\partial c} \frac{c^*}{m^* c^*},
\]  

(25)
where asterisks specify the optimal values. If we denote the elasticity of utility function with respect to consumption at the optimum by \( \varepsilon^u_c \) and elasticity of health production function with respect to health spending at the optimum by \( \varepsilon^f_m \), the following lemma formalizes the conditions under which \( s \) increases with income, all else equal.

**Lemma 3** For a fixed level of health status and for large enough income, the optimal share of health spending in income increases if, and only if, \( \varepsilon^f_m / \varepsilon^u_c \) falls with income in the optimum.

If we think of (24) as the problem of an individual allocating resources between health and non-health consumption in a given period, Lemma 3 characterizes a standard luxury-good channel for health spending: As income increases (say, between two periods), if the marginal utility of consumption normalized by its average utility, falls relative to the marginal product of health spending normalized by its average product, the individual is better off dedicating more resources to health spending.\(^{33}\)

Under the condition of Lemma 3, assuming \( \partial y (h, t) / \partial t > 0 \) and \( \Gamma (h, t) = \Gamma (h) \) for all \( h \in \mathcal{H} \), as time passes and new cohorts enter the economy, the share of average health spending in average income increases:

\[
\frac{\int_{\mathcal{H}} m^* (h_0, t_1) \Gamma (dh_0)}{\int_{\mathcal{H}} y (h_0, t_1) \Gamma (dh_0)} > \frac{\int_{\mathcal{H}} m^* (h_0, t_2) \Gamma (dh_0)}{\int_{\mathcal{H}} y (h_0, t_2) \Gamma (dh_0)},
\]  

(26)
if \( t_1 < t_2 \).\(^{34}\)

33. Replacing the objective function in (24) by a function of the form \( U (c_1, c_2) \) does not change this argument by much: For \( c_1 \) to be a luxury good, the marginal utility of \( c_2 \) must fall rapidly, compared to the marginal utility of \( c_1 \).

34. As we are going to discuss, the assumption that \( \Gamma (\cdot, t) \) remains constant over time is not pivotal to this result: As long as the increase in the average health status in the economy does not dominate the increase in average income, Inequality (26) will hold.
A Necessity in the Cross Section  Among the individuals of a single cohort, however, income and health status move simultaneously. In general, we are inclined to believe that an individual with a higher initial health status, $\bar{h}_0$, has higher income than an individual with a low level of health status, $h_0$.$^{35}$ Under the assumption of Lemma 3, as a result of this income differential, a high-income individual tends to dedicate a higher share of her income to health spending.

Besides this indirect effect of health status on health spending through income, differences in health status have a direct impact on health spending through their effect on the marginal product of health spending: If marginal product of health spending in extending life falls as a result of an increase in health status, individuals with better health tend to allocate less resources to health spending.

The total effect of an increase in health status on health spending depends on the relative importance of the direct and indirect effects, as formalized by the following lemma. $\varepsilon_f^{fm}$ and $\varepsilon_c^{uc}$ in Lemma 4 are the elasticity of marginal product of $m$ with respect to health status and the elasticity of marginal utility of consumption with respect to $c$, at the optimum.$^{36}$

**Lemma 4**  In a given cohort $t_0$, $\partial m^* (h, t_0) / \partial h > 0$ if, and only if,

$$
\left[ \frac{\partial y (h, t_0)}{\partial h} \right] \left[ \frac{\varepsilon_c^{uc} - \varepsilon_c^{uc}}{c^* (h, t_0)} \right] - \left( \frac{\varepsilon_f^{fm} - \varepsilon_f^{fm}}{h} \right) > 0. \tag{27}
$$

The first term on the left hand side of Inequality (27) captures the indirect effect of health status on health spending through income, whereas the second term characterizes its direct effect through its impact on health production function.

Note that the direct effect of health status on spending depends directly on the cross-elasticity of the health production function with respect to health spending and health status: The higher (and more positive) this cross-elasticity, the higher the chance that the sec-

---

35. See Section 1 for a brief review of the literature that documents this relationship for different indices of health status.

36. The proof of Lemma 4 calls for differentiating the first order conditions of Problem (24) and rearranging the resulting equation.
ond term in (27) dominates the first term, leading to a declining schedule of health spending as a function of income, in the cross section.

Moreover, if a small difference in income among the individuals of a single cohort is associated with a large difference in health status (that is there exists a strong correlation between income and health status in the cross section), the probability that the first term on the left-hand side of (27) is dominated by the second term is higher.

Lemmas 3 and 4 illustrate the main mechanisms behind a steep and upward sloping Engel curve over time and a downward sloping curve in the cross section: In the cross section, if an observed increase in income in the data is associated with a large increase in individuals’ underlying health, under the condition of Lemma 4, we can expect the health spending to decline. On the other hand, as long as the rise in income over time is not dominated by an improvement in the general health status in the economy, a standard luxury-good argument implies that we can expect the share of health spending to rise.37

Before summarizing these arguments for two standard functional forms of particular interest for the utility and health production, we are going to briefly explain the relation between the simplified model of this section and the full economy of Section 2.

Relation to the Full Model  Even though it is hard to extend Lemmas 3 and 4 analytically to the full model, their logic still applies to the complete economy of Section 2. This can be seen by comparing the optimality condition of Problem (24),

\[
\frac{\partial f(m, h_0)}{\partial m} \frac{f(m, h_0)}{u(c)} = \frac{u'(c)}{u(c)},
\]

37. Lemma 4 clarifies what we mean by "domination" in this context: As long as the income growth over time is not accompanied with an increase in health status that violates Inequality (27), it leads to an increase in health spending. In the next section, we will specify the conditions under which this increase actually leads to a rise in the health spending as a share of income, for the functional forms of interest.
to that for the optimal feedback rule, which given the value function $V(\cdot)$, the state vector $x$, and time $t$, can be written as

$$\frac{\partial f(m, h, a, t)}{\partial m} \frac{f'(c)}{f(m, h, a, t)} = \left[1 - s\left(y(h, h^{R}, a, t), a\right)\right] \frac{u'(c)}{V(x, t) / f(m, h, a, t)}.$$  \hspace{1cm} (29)

If we could approximate $V(x, t) / f(m, h, a, t)$ by $u(c)$, then the same logic as the simplified model would directly carry over to the full model. Though such an approximation is not accurate, mostly due to the depreciation of health status through life, our numerical results suggest that it is valid, especially earlier in life, up to a linear transformation. \(^{38}\)

**A CRRA Utility and a CES Health Production** Consider the following constant relative risk aversion (CRRA) flow utility function with an additive term to which, following Hall and Jones (2007) and Ales, Hosseini, and Jones (2014), we will refer as the value of being alive:

$$u(c) = b + \frac{c^{1-\sigma}}{(1 - \sigma)}.$$  \hspace{1cm} (30)

The parameter $\sigma$ in (30) is the degree of relative risk aversion. It determines the elasticity of intertemporal substitution.

Assume the health production function is given by the following constant elasticity of substitution (CES) form:

$$f(m, h) = A \left[\alpha (z \cdot m)^{\gamma} + (1 - \alpha) h^{\gamma}\right]^{\frac{\beta}{\gamma}},$$  \hspace{1cm} (31)

where $z > 0$ is a measure of technological progress, $\alpha \in (0, 1)$ is the share parameter, and $A > 0$ is the total factor productivity. The other two parameters of interest in (31) are $\gamma \in (-\infty, 1]$ and $\beta \in (0, 1]$, which determine the elasticity of substitution between $m$ and $h$.

\(^{38}\) Given a smooth stream of consumption—which is a rather accurate approximation under the optimal control according to our simulations of the full economy—if $f(\cdot)$ was equal to the life-expectancy, then $V(x, t) / f(m, h, a, t) \approx u(c)$ would be an accurate approximation. However, in the full model, $f(\cdot)$ is not exactly equal to the life-expectancy, but only a rough approximation, up to a linear transformation.
The CRRA utility function is widely used in macroeconomic literature due to the fact that it implies a constant elasticity of marginal utility, a constant degree of relative risk aversion, and a declining degree of absolute risk aversion. However, the constant term \( b \) also plays a crucial role when it comes to the effects of health and health spending, because it determines the level of utility and, consequently, the value of being alive in comparison to the utility at death.\(^{40}\)

A CES health production function, on the other hand, is a novelty of this study. In particular, before this paper, researchers have ignored the significance of the effect of underlying health on the marginal product of health spending as captured by the cross-elasticity of health production with respect to health status and health spending. For instance, Hall and Jones (and many others) restrict their attention to a case where the elasticity of substitution is equal to one, by assuming a health production function in which health spending and “other factors” enter multiplicatively.\(^{41}\)

While the introduction of the elasticity of scale, \( \beta \), in (31) allows us to capture the possibility of diminishing returns as in Hall and Jones, we do not limit ourselves to a Cobb-Douglas functional form. This, as we are going to discuss, makes it possible for the elasticity of health production with respect to health spending to fall rapidly with health status. Consequently, high income individuals have less incentives to allocate resources to health spending, as long as their are “healthy enough.”

In the rest of this paper, we are going to focus on the two functional forms in (30) and

\(^{39}\) Kmenta (1967) was the first paper that added the parameter \( \beta \) to Arrow et al. (1961)’s constant returns to scale CES production function. The addition of this term allows us to nest Hall and Jones (2007), Ales, Hosseini, and Jones (2014), and many others’ health production functions, as special cases.

\(^{40}\) One should also note that, for the standard range of values for \( \sigma \) in the macroeconomic literature, the level of utility in (30) becomes negative when \( b = 0 \). As a result, with \( V^d \) normalized to zero, “mortality becomes a good, rather than a bad” (Hall and Jones 2007). This implies, in the absence of the additive term, individuals would rush to their death!

\(^{41}\) In addition, by assuming that the elasticity of scale is similar for both factor inputs, Hall and Jones are implicitly assuming that the two factors have equal shares in the production (that is \( \alpha = 0.5 \) in (31)). Considering the fact that both in Hall and Jones and our study, health status is a “latent variable,” this is only a matter of normalization.
Our main quantitative challenge is the estimation of the parameters of these functions. To see how these two functional forms help us account for the observed patterns of health spending in the time series and cross section, and how we can use these observations to discipline the structural parameters of interest, let us consider the implications of (30) and (31) for Lemmas 3 and 4.

With the CRRA utility form of Equation (30), the elasticity of utility with respect to consumption is given by

\[ \varepsilon_u = \left[ \frac{1}{b c^{\sigma - 1} - \left( \frac{1}{\sigma - 1} \right) } \right]. \tag{32} \]

On the other hand, for the elasticity of health production with respect to health spending when health production function is given by (31), we have

\[ \varepsilon_f = \frac{\beta}{1 + \left( \frac{1 - \alpha}{\alpha} \right) \left( \frac{h}{z m} \right)^\gamma}. \tag{33} \]

When \( b > 0 \), for a degree of relative risk aversion that is greater than unity (as broadly accepted in the macroeconomics and finance literature), \( \varepsilon_u \) declines rapidly with income (assuming non-health consumption is a normal good). When \( \gamma = 0 \) in (33)—as assumed previously in the literature—the ratio of elasticities in Lemma 3 rises rapidly with income.\(^{42}\)

On the other hand, when health status and health spending are stronger substitutes, that is for elasticities of substitution greater than one (\( \gamma > 0 \) in (30)), \( \varepsilon_f \) no longer remains constant with changes in income. Specifically, if health status is held fixed (as assumed in Lemma 3), \( \varepsilon_f \) increases with rises in income. This, in turn, implies that the ratio \( \varepsilon_f / \varepsilon_u \) increases more rapidly with income, leading to a rapid rise in the share of health spending.\(^{43}\)

In summary, in addition to Hall and Jones’s channel where the rise in the share of health spending is attributable to the rapid decline of the *value of consumption* relative to the *value...*
of being alive, our health production function allows for a new channel for the rise of health spending as a share of income: The rise in the share of health spending in life expectancy—or, more broadly, in marginal utility—compared to the share of health status.\footnote{44}

For the CRRA utility function of Equation (30), the first term in Equality (27) becomes

\[
\frac{\partial y(h, t)}{\partial h} = \left(\frac{1}{b c^\sigma + \frac{c}{1-\sigma}} + \frac{\sigma}{c}\right).
\] (34)

When $\sigma > 1$ and $b > 0$, for large enough values of consumption, the term in the parentheses is positive and declining in consumption.

On the other hand, for the general CES production function in Equation (31), the second term in (27) can be written as

\[
\frac{\gamma}{h \left[1 - \alpha + \alpha \left(m h^{-\gamma}\right)^\gamma\right]}.
\] (35)

A comparison of (34) and (35) reveals that, when $\gamma = 0$, the model of Section 2 (and, as discussed in the previous subsection, our full model) has no hope in accounting for a downward sloping schedule for health spending in the income cross section. On the other hand, when health status and health spending are strong compliments in the production of health—that is, when $\gamma \gg 0$—the model implies an increasing Engel curve in the cross section. Only for values of $\gamma$ which are above zero, the model can deliver a downward-sloping spending curve. If we assume that Inequality (27) holds and $h$ is sufficiently large in Equation (35), a higher substitutability between health spending and health status—as captured by a larger $\gamma$—implies a steeper Engel curve in the cross section.

The following proposition summarizes the preceding arguments on how each of the parameters of the functions in (30) and (31) help us capture an aspect of the health spending patterns in the data.\footnote{45}

\footnote{44. It is worth noting that the “utility channel” allows for the share of health spending to tend to one asymptotically. However, the “health production channel” is limited in its capacity to explain the ever-growing rise in the share of health spending in the last five decades.}

\footnote{45. This proposition is a direct corollary of Lemmas 3 and 4.}
**Proposition 5** With the constant relative risk aversion utility form of Equation (30) and the constant elasticity of substitution health production function of Equation (31):

(i) for any $0 < \gamma < 1$, there exists some $B_{t_0} > 0$ such that, $0 < \partial y (h, t_0) / \partial t < B_{t_0}$ for all $h_0 \in \mathcal{H}$ implies $\partial m^*(h_0, t_0) / \partial h < 0$; and

(ii) when $\partial y (\cdot, t) / \partial t > 0$, then

$$\frac{\int_{\mathcal{H}} m^*(h_0, t_1) \Gamma (dh_0)}{\int_{\mathcal{H}} y (h_0, t_1) \Gamma (dh_0)} < \frac{\int_{\mathcal{H}} m^*(h_0, t_2) \Gamma (dh_0)}{\int_{\mathcal{H}} y (h_0, t_2) \Gamma (dh_0)}, \quad t_1 < t_2,$$

if either (1) $b > 0$ and $\sigma > 1$, (2) $\gamma > 0$, or both.

In the next section, we are going to use Proposition 5 to make inference about the structural parameters of the model—importantly, the value of being alive and the elasticity of substitution between health spending and health status.$^{46}$

4. The Quantitative Analysis

The estimation of health production functions has been historically challenging primarily because of the lack of reliable measures for health status. A straightforward way to see this is in the context of our full model in Section 2. As assumed in our model, an unobserved shock to health status (as captured by the Brownian process governing $\omega_h$) simultaneously affects the individual’s income (through the income equation). Income, in turn, is a main determinant of health spending (as discussed in the previous section). Any examination of the relation between health outcomes (such as mortality, physiological outcomes, or mea-

$^{46}$ At the end of this section, it is worth mentioning that this simple model can be used to study the role of technological progress—that is changes in $z$—on health spending. Under our formulation of health production function, Equation (31), the role of health care technology is to determine the relative price of the two commodities in the economy (as it is the case in many standard macroeconomic models). As a result, a change in $z$ entails a substitution and an income effect. The total impact of the growth in technology, therefore, depends on the magnitude of each of these effects. When $\gamma = 0$, for instance, these two effects cancel out, leaving technological innovations neutral with regard to the level of health spending. With $\gamma > 0$, however, technological improvements lead to an increase in the share of health spending relative to income. This channel is present in our quantitative exercise in the next section.
asures of the burden of diseases) that cannot capture these shocks in a health capital index runs into the possibility of endogeneity and, consequently, biased estimates.\(^{47}\)

A large literature in health economics is dedicated to this topic, suggesting a multitude of instrumental variables to address the problems arising due to the endogeneity. However, almost all of this literature ignores the possibility of cross-effects between the underlying health status and health care spending—specifically, the effect of underlying health on the marginal product of health spending, despite the early evidence on the importance of these cross-effects going as far back as the RAND’s seminal health insurance experiment.\(^{48}\) Using data from the RAND HIE, Brook et al. (1983) show that the effect of health care utilization on health outcomes can be significantly different across different income groups and across groups with different risk factors.\(^{49,50}\)

Instead of instrumenting for health spending (or health capital)—as most of the studies before them do—Hall and Jones (2007) and Ales, Hosseini, and Jones (2014) estimate a restricted form of the health production function in Equation (31) using a time trend as

47. The above argument ignores the effect of idiosyncratic productivity shocks which can exacerbate the endogeneity issue.

48. RAND Health Insurance Experiment (RAND HIE) was a multimillion-dollar randomized controlled trial conducted between 1971 and 1986, founded by the Department of Health, Education, and Welfare, which, to this day, remains the largest health policy study in the United States history. The study randomly assigned families across different health insurance plans with different levels of cost sharing. One of the main findings of the study was that the health care utilization was significantly different across different plans. (Newhouse et al. 1981’s findings remain one of the main references for the price elasticity of health spending, both in macroeconomics and health economics literature, to this day.) Due to its random nature, RAND HIE also provided an excellent instrument to study the effects of health care utilization on health outcomes, including the self-assessed health and detailed physiological outcomes measured by the RAND investigators.

49. For instance, as Phelps (2016) notes, “[f]or persons with relatively high health risks (e.g., from obesity, smoking, high blood pressure), the risk of dying was reduced by about 10 percent in the full-coverage group […].”

50. Eslami and Karimi (2018b) use the RAND HIE data to estimate an approximated version of (31) using instrumental variable techniques. To this end, they construct an index of health capital as the common component of socioeconomic correlates of health and use it to estimate the relation between several measures of health outcome, health spending, health capital, and their cross-product, instrumenting for the health care utilization. Their estimates indicate a significant cross-effect between health capital and health spending on most measures of health outcomes. The limited sample size of the RAND HIE data, however, keeps Eslami and Karimi from estimating these cross-effects for different age groups.
an instrument (in a Generalized Method of Moments (GMM) estimation procedure). The logic of this approach is as follows: Restricting the elasticity of substitution between health spending and “other underlying factors” to one implies that technological innovations, the rise in health spending, and the increase in these underlying factors each captures a constant fraction of the improvement in health outcomes over time. If we consider mortality rate (at different ages) as the main indicator of health outcomes and assume that technology in the health sector grows at the same rate as the non-health sector, the only remaining unknown is the growth rate of the underlying factors. In their benchmark analysis, both studies assume that the growth rate of these factors pertains to one third of the total decline of mortality in the United States. This enables them to estimate the elasticity of scale—that is $\beta$ in (31)—by imposing two moment conditions on the detrended rates of mortality in the past five decades: they have zero mean and are uncorrelated with a time trend.

Our discussions in the previous section deem the constraint $\gamma = 0$ on the health production function as “too restrictive.” In this paper, we relax the restriction on the cross-elasticity of health production with respect to health status and health spending. The parameters of the resulting relaxed functional form, however, can no longer be estimated using Hall and Jones’s suggested approach.

Identification Strategy: a Case for Indirect Inference Instead of directly estimating the effects of health spending and health status on health outcomes (mortality, specifically), we take an indirect approach: We use the patterns of health spending in the cross section and over time to make inferences about the parameters of the health production function (beside the value of being alive) using the results of the previous section.

To demonstrate the underlying logic, for the sake of argument, let us assume that we know the relation between health status and income within and across cohorts—as specified by the functional form $y(h, t_0)$ in (3). Moreover, let us focus our attention on the elasticity of substitution and the elasticity of scale parameters by assuming the share parameter $\alpha$ and the growth rate of $z$ in (31) are known.

In the absence of any uncertainty (as in the simplified model of (24)), starting from any initial cohort $t_0$ and income level $y(h_0, t_0)$, two instances of income change suffice to infer
all the (remaining) parameters of interest: a change in cohorts, which corresponds to an increase in income not associated with an increase in health status; and a change in health status in a given cohort.\textsuperscript{51}

The arguments leading to Proposition 5 reveal that these two variations in income, together with the level of spending at the initial state, enable us to deduce $\gamma$, $\beta$, and $b$ in Equations (30) and (31). Specifically, an increase in the share of health spending across two cohorts reveals the ratio of the elasticity of utility with respect to consumption relative to the elasticity of health production with respect to health spending. On the other hand, the slope of health spending with respect to income among the individuals of single cohorts contain valuable information regarding the cross elasticity of health production with respect to health status and health spending. These two pieces of information, when combined with the information contained in the level of spending, suffice to infer all the parameters of utility and health production functions beyond what has already been assumed.\textsuperscript{52}

As discussed in Section 3, there is a close relation between the full model of Section 2 and the simplified version of Proposition 5. As a result, we expect the above logic to extend naturally to the full model. Nevertheless, even in the case of simplified model of Section 3, finding an analytic solution for the model is not possible for a generic set of parameter values.\textsuperscript{53} In the full life-cycle model, computations are considerably more complicated, mainly due to the nonstationarity of the problem. As a result, finding a one-to-one relationship between the parameters of the model and the coefficients of health spending schedules is not feasible.

The approach we take in this paper is the simulated method of moments (SMM): While it

\textsuperscript{51} The total factor productivity, $A$ in Equation (31), has no bearing on the level or the slope of the health spending schedule in the simplified model of Section 3. However, given the share parameter $\alpha$, it has important implications for the distribution of the health outcomes.

\textsuperscript{52} The total factor productivity in the health production function, $A$, is determined through the relation between health spending, health status, and the health outcome of interest.

\textsuperscript{53} Even when we limit ourselves to the case where the degree of relative risk aversion is 2—as is broadly used in the macroeconomics and finance literature—it is not possible to write optimal health spending as an explicit function of income and health status, except when $\gamma = 0$ or $\gamma = 1$. Our choice of the indirect inference as our estimation method is mainly to avoid such simplifications. See Guvenen and Smith (2010) for an excellent discussion.
is not possible to find an analytic solution to the model of Section 2—as characterized by the two partial differential equations, HJB and KF—we still can find a numerical solution for a chosen set of functional forms and structural parameters. This solution, then, can be used to generate a simulated series from the model. The basic idea behind the SMM is to choose the structural parameters such that the moments of interest in the simulated series match those from the data.

The relation between health spending and income in the cross section and its variations over time provide us with the sufficient moments, as suggested by the above arguments. This can be characterized in the form of an estimation equation of the form

\[ m_{a,t}^i = \beta_{0,t}^a + \beta_{1,t}^a \cdot y_{i,t}^a + \beta_{2,t}^a \cdot (y_{i,t}^a)^2 + \beta_{3,t}^a \cdot (y_{i,t}^a)^3 + \epsilon_{i,t}^a, \]  

(36)

known as the auxiliary model. The variables \( m_{a,t}^i \) and \( y_{i,t}^a \) in Equation (36) are health spending and income at time \( t \) for individual \( i \) in the data, respectively, who has age \( a \). For a given time \( t \), the coefficients of Equation (36) (\( \beta_{i,t}^a \)'s) capture the relation between income and health spending in the cross section for individuals of different ages at time \( t \). Estimating this equation for different \( t \)'s, then, characterize the variations of this relation over time. The higher order terms on the right hand side of the auxiliary model capture the fact that the relation between income and health spending is far from being linear, as suggested by our model.\(^{54}\)

Our objective is to choose the structural parameters of the model so that the series generated by the model (under these parameters) look as close as possible to the actual data, as represented by the coefficients of the auxiliary model. This is the basic idea behind the indirect inference approach.

\(^{54}\) If our model is to represent the important mechanisms present in the “real world,” we can expect an estimation of Equation (36) to result in higher order coefficients that are statistically significant; after all, a highly non-linear relationship between income and health spending, as suggested by the model, means that the several initial terms of a Taylor approximation of the “actual” relationship are significant.

One can expect the lagged income to also have an effect on the health spending because of its effect on savings, according to our model. These terms become important specially in the presence of stationary idiosyncratic productivity shocks. Unfortunately, the limitations in MEPS’ panel features prevent us from using these moments.
The indirect inference method, first proposed by Smith (1990, 1993) and further developed by Gourieroux, Monfort, and Renault (1993), provides a criterion—through the use of an auxiliary model—to infer the structural parameters of interest in the model. In effect, the indirect inference approach provides an answer to a key issue in the SMM, through the use of an auxiliary model, and that is which moments to match (Qu 2012). As Guvenen and Smith (2010) write,

“the indirect inference estimator is obtained by choosing the values of the structural parameters so that the estimated model and the US data look as similar as possible when viewed through the lens of the auxiliary model.”

The Income Equation  The above discussion forms the basis of our quantitative analysis with a not-so-trivial shortcoming: The relation between health status and income is not known. Without the knowledge of such a relationship, the identification of the parameters of the health production function is not possible. 55 Importantly, in our analysis, we want to remain faithful to the notion of health status as a latent variable. This prevents us from using an index of observed characteristics as a measure of underlying health status (as is a standard practice in the literature). 56

To overcome this difficulty, we use another source of variations in the data to make inference about the income equation. We take advantage of the relation between the rate of mortality at different ages—as a measure of health outcome—and income, as well as the variations of this relationship over time, to deduce how income and health status are related to each other. 57

These two steps, that is comparing model’s simulations regarding the joint distributions of income and health spending and income and life expectancy to the data, can be combined in

55. Without an income equation, for any given set of parameter values, one can “choose” a level of individual health status such that the model matches the data perfectly.

56. As our discussions at the start of Section 4 suggest, in the absence of reliable instruments, using such measures of health status is prone to an endogeneity problem.

57. Limiting ourselves to the notion of health production as a determinant of longevity—at least in this section—enables us to compare the resulting relation between income and longevity with the actual data, and to make further inference about the parameters of the income equation.
the form of two auxiliary equations: in practice, one can choose the parameters of an income equation (that is the parameters of a functional form of choice for Equation (3)) and those of the utility and health production functions simultaneously such that model’s simulations are “as close as possible” to the actual data. Closeness, in this context, is determined by estimates of Equation (36) and an equation relating life expectancy to income.

In our estimation procedure, however, we are going to perform these two comparisons sequentially: In the first step, for a given set of parameters of the health production function, we choose the parameters of an income equation. This is done such that, should the model mimic the relationship between income and health spending in the data as closely as possible, the resulting joint distribution of income and life expectancy also matches that in the data. This leads to an estimate of the income equation that is conditional on the parameters of the health production function being equal to the SMM estimates. In the second step, the conditional estimate of the income equation and the choice of parameters of the health production function are used to compare the moments of the auxiliary equation (36) between the data and the model. We repeat these steps until the moments of interest in the model are close to those in the data.

In Section 4.2, we are going to explain this procedure in more detail in the context of an estimation algorithm. As discussed above, this procedure uses the joint distributions of income and health spending, and income and mortality rate, at different ages, and their changes over time. In Section 4.1, we will briefly explain the data used for this purpose.

4.1. Data

The data on the joint distribution of income and health spending as a function of age and cohort is taken from the household component of the Medical Expenditures Panel Survey (MEPS). Annually conducted by the Agency for Healthcare Research and Quality (AHRQ) and the National Center for Health Statistics (NCHS) form 1996, the MEPS includes nationally representative surveys of detailed health care utilization and expenditures for the United States’ civilian, non-institutionalized population.

Health care expenditures are based on individuals’ self-reports, but they are verified by
and supplemented with reports from medical providers and employers. Therefore, the MEPS provides a reliable source of information on health care expenditures for surveyed individuals. Since it collects ample individual and family background information, it is also suitable for studying the distribution of health care expenditures by demographic and socioeconomic characteristics.

Total health care expenditures in the MEPS consist of expenditures, regardless of the payer, on most medical services. Health care expenditures are paid out-of-pocket or by private insurance, Medicaid, Medicare, or other local, state, and federal sources. Medical services in MEPS are categorized into nine groups: medical provider visits, hospital outpatient, inpatient and emergency room visits, dental visits, home health care, vision aids, prescribed medicines, and other medical equipment and services.\textsuperscript{58}

MEPS also gathers detailed information on respondents’ family income. Family income is the summation of all family members’ income. An individual’s income includes money made from all sources such as wages and compensations, business incomes, pensions, benefits, rents, interests, dividends, and private cash transfers, \textit{excluding tax refunds and capital gains}.\textsuperscript{59}

A central objective of this paper is to study the effect of the evolution of income on health spending. Therefore, as discuss in more detail in what follows, we use \textit{Center on Budget and Policy Priorities} (CBPP)’s estimates and projections of the growth rate of income at different parts of the income distribution to approximate the evolution of this distribution before and

\textsuperscript{58} Total health care expenditures calculated from the MEPS data are significantly different from the estimates provided by the \textit{National Health Expenditures Accounts} (NHEA), which mainly use aggregate providers’ revenue data. The disparity does not originate from different estimations of expenditures on comparable services but from differences in inclusion of services and in covered populations. For example, expenditures on over-the-counter drugs, longer than 45-day stays in hospitals, and for institutionalized individuals are out of the MEPS’ scope. Once aggregate estimations are adjusted for service and population, and measurement methods are made compatible, they tend to converge. In effect, the average growth rates of per person health care expenditures, driven from MEPS and NHEA, are very similar (See Eslami and Karimi 2018a).

\textsuperscript{59} The exclusion of capital gains from income is consistent with the definition of income in our model. However, we believe that our lack of access to tax refunds and transfers in the MEPS data—specifically for the lower income groups—\textit{does} affect our results, as the \textit{Congressional Budget Office} (CBO)’s \textit{comprehensive income measures} paint a drastically different account of the growth rate of income at the bottom of the income distribution.
after the MEPS’ time span (Stone et al. 2015).

Data on the relation between income and life expectancy is taken from Chetty et al. (2016)’s seminal work on the association between the life expectancy and income in the United States. Chetty et al.’s analysis uses a database of federal income tax and Social Security records that includes all individuals with a valid Social Security Number between 1999 and 2014. Chetty et al. construct the period life expectancy conditional on income percentile by (i) estimating mortality rates for the ages of 40 to 76 years; (ii) extrapolating mortality rates beyond the age of 76 years and calculating the life expectancy; and (iii) adjusting for differences in the proportion of racial and ethnic groups across percentiles. Using the Social Security Administration (SSA)’s death records, these steps lead to estimates for period life expectancy at 40 for men and women at different levels of income over the period 2001–2014. (See Figures 2 and 3 in Chetty et al. 2016.)

To compute the variance of health shocks, we use information from Ales, Hosseini, and Jones (2014): the variance of log income at different ages. Ales, Hosseini, and Jones use the data from the Panel Study of Income Dynamics (PSID) to construct this measure. (See Figure 4 in Ales, Hosseini, and Jones 2014.)

4.1.1. Data Preliminaries

We use the disposable family income in the MEPS—by the Current Population Survey (CPS)’s definition of family—as individuals’ income. At every stage of our estimations, we use family-level weights provided by the MEPS, so that the survey samples provide as close a representation to the United States’ non-institutionalized population as possible. All dollar values are adjusted for inflation, using the personal consumption expenditure (PCE) index. For expenditures on components of medical services, we used the corresponding personal health care (PHC) components such as PHC for hospital care, for physician and clinical services, and for dental services. All real values are in 2009 dollars. In addition, all individuals

60. This is consistent with Chetty et al. (2016)’s measure of income.
with zero income in a given year are dropped from the sample.  

To use the variations of income over time in the estimation of the structural parameters of the model, we divide the MEPS’ sample period into two sub-periods: 1996–2005 and 2006–2015—consistent with our assumption that each period in our model is equivalent to an interval of ten years in the data, as discussed later on. The resulting two sub-periods include 300,610 and 339,039 individual-year observations, respectively. We also group individuals in each sub-period into four age groups of ten-year intervals, from 40 to 70, and an age group of individuals older than 80 years old. Table 1 provides a summary of the MEPS data in each age group and across the two sub-samples.

Using the MEPS data, structured as described in Table 1, we construct the empirical distributions of income for different age groups in each of the sub-periods. As noted before, one of the central objectives of this paper is to use the variations in income over time to make statistical inferences about the production of health. To address the fact that, at any given age and year, rational agents also take these variations—both in the past and future—into account, we need to extend the age-specific distributions of income to time periods before

61. We drop zero-income individuals for two reasons. First, our model does not allow for consumption to fall below a certain level—because the utility must remain positive at all dates. Second, the SSA records do not provide reliable data for such individuals.

62. Dropping the top and bottom 2.5% of the income distribution in each time period in a robustness exercise does not alter our results significantly.

63. Hall and Jones (2007) consider five-year periods.

64. Individuals below the age of 40 are dropped to remain consistent with Chetty et al. (2016)’s estimates.

65. There are two main reasons for choosing ten-year time periods as the length of one period in the model: First, due to year-to-year changes in the randomly selected MEPS’ samples, there are year-to-year (sometimes irregular) fluctuations in health care expenditure estimates, especially when sub-groups are identified by more than one characteristic. However, using time intervals of length five years show that the patterns of changes in income and health spending are remarkably similar to those between the two periods, 1996–2005 and 2006–2015.

The second—and more important—reason is that this rough temporal grid leads to far fewer structural parameters of the health production function. This smaller set of unknowns, in turn, eases the computation burden of the estimation procedure significantly. The authors have implemented a version of the numerical method that uses intervals of length five years. However, at the time of writing this draft, the results of this implementation are not reliable.
<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Obs.</td>
<td>Percentage of Males</td>
<td>Avg. Age</td>
</tr>
<tr>
<td>40–49</td>
<td>42,417</td>
<td>49</td>
<td>44.36 (0.02)</td>
</tr>
<tr>
<td>50–59</td>
<td>32,078</td>
<td>48</td>
<td>54.14 (0.02)</td>
</tr>
<tr>
<td>60–69</td>
<td>20,384</td>
<td>47</td>
<td>64.22 (0.03)</td>
</tr>
<tr>
<td>70–79</td>
<td>15,509</td>
<td>43</td>
<td>74.19 (0.03)</td>
</tr>
<tr>
<td>≥ 80</td>
<td>8,158</td>
<td>36</td>
<td>83.65 (0.04)</td>
</tr>
<tr>
<td>≥ 40</td>
<td>118,546</td>
<td>46</td>
<td>57.13 (0.08)</td>
</tr>
</tbody>
</table>

**Source:** Medical Expenditure Panel Survey (1996–2015). **Note:** Observations with zero income have been dropped from the sample. For the percentage of males and average age in each age group, personal weights, provided by the MEPS are used. The total number of observations do not take survey weights into account.
and after the MEPS’ relatively short time span.\footnote{An important decision in the quantitative studies of life-cycle phenomena is how to deal with cohort effects. In particular, as noted, we assume that each time interval of unit length in our model corresponds to a period of ten years in the MEPS data. This means that, in each of the sub-periods, we have individuals who have entered the economy before the start of the sample’s time span and who will leave the economy after the sample’s end date.}

To this end, we use the CBPP’s estimates—using the CBO’s data—to extrapolate the age-specific distributions of income in the first time period. This gives us the age-specific distributions of income down to the year 1980. For the years before 1980, we assume a unified growth rate (equal to the growth rate of GDP per capita) for income at all income percentiles. The CBPP’s projections are used in a similar procedure to extrapolate the income distributions up to 2035.\footnote{This is the year at which individuals who are 40 at 2010 retire.} We use these in the estimation of the parameters of the income equation in periods that fall outside the MEPS’s time span.

We use the MEPS data to estimate the joint distribution of health spending and income for each age group, in each of the time periods. This distribution, then, is used to compute the average level of health spending in each income ventile, for each age group and time period.

\begin{footnotesize}
\footnote{An important decision in the quantitative studies of life-cycle phenomena is how to deal with cohort effects. In particular, as noted, we assume that each time interval of unit length in our model corresponds to a period of ten years in the MEPS data. This means that, in each of the sub-periods, we have individuals who have entered the economy before the start of the sample’s time span and who will leave the economy after the sample’s end date.}
\end{footnotesize}

\begin{footnotesize}
\footnote{A rather standard approach in the quantitative literature is to ignore these effects. For example, by pooling the MEPS’s data, Ales, Hosseini, and Jones (2014) assume that individuals of age $a$ in the model are going to face the same income as the individuals of age $a + 1$ in the sample, when they become $a' = a + 1$. This approach obviates the need to make additional assumptions regarding the evolution of variables outside the sample. However, a more important advantage of it is to eliminate an aggregate state from the problem—that is time—which has a considerable impact on the computational burden.

In practice, this is the same approach used in many longevity studies, including Chetty et al. (2016), to calculate the life expectancy from the period life tables. (And, as we are going to talk about in Section 4.3, we start our numerical search for the model parameters by assuming this is the case.) Due to the nature of our claims, however, we find it hard to justify that individuals in our economy do not take into account the evolution of their income—and also technological innovations in the health sector—when making decisions.

With regard to the future evolution of income, using a projection seems justifiable (e.g., Arnold and Plootinsky 2018). For the evolution of a cohort’s income distribution over its lifetime, what matters, from the perspective of our model, is the distribution of asset holdings. The MEPS, however, does not provide information on this variable. Our use of the CBPP’s estimates for the evolution of income before the MEPS’ sample, in effect, serves as a tool to construct this distribution for individuals of different ages who, at the beginning of the MEPS sample, are older than $a$.

Another possibility is using other data sources to impute asset holdings at different income percentiles. At the time of writing this draft, the authors are exploring this possibility using PSID.}
\end{footnotesize}
To calculate health status as a function of age and income for any given set of parameter values, we need to have the survival rate at each age among different income groups in each of the sub periods. To this end, we reconstruct the mortality rate at different ages, at each point in Chetty et al.’s sample period, from the reported life expectancies at 40. We do this by inverting the procedure used to construct the average life expectancy from the period life tables, as follows:68 We start by extrapolating Chetty et al.’s estimates to the years in the MEPS that are missing from Chetty et al.’s sample period. Next, we compute the average life expectancy at 40 across the income groups in each time period of interest, 1996–2005 and 2006–2015, using the MEPS’ distribution of individuals by their gender in each period.

The Gompertz equation defines the rate of mortality as a function of age as

\[
\ln (mortality \ at \ age \ a) = g_1 + g_2 \ln (a). \tag{GL}
\]

Assuming that the Gompertz law provides an accurate description of the mortality as a function of age at each level of income and each time period, we estimate \( g_1 \) and \( g_2 \) in Equation (GL) for each time period and each income ventile. We do this such that the resulting life expectancy at 40 matches those reported by Chetty et al., under the assumption that \( \bar{a} = 100 \). Figure 1 illustrates an example of the resulting mortality and survival rates for the bottom and top income ventiles, in the first time period.

4.2. Estimation Strategy

We start the discussion of our estimation algorithm by choosing the specific functional forms whose parameters, beside those of the CRRA and CES utility and health production functions of Section 3, are going to be directly or indirectly targeted in the SMM method.

68. This is the procedure used by Chetty et al. to compute life expectancy from mortality rates at different ages, as discussed previously.

69. Chetty et al. show that the Gompertz equation provides a remarkably good description of the rate of mortality as a function of age, at different levels of income, up to a certain age. (See Figure 1 in Chetty et al. 2016.) In their estimates of the life expectancy, they use this equation to extrapolate the mortality rate beyond the age of 76.

It should be noted that the fit of the Gompertz model declines drastically after the age of 90.
**Figure 1.** Gompertz Approximations in the 5th and 95th Income Percentiles

*Source:* Authors’ calculations based on Chetty et al. (2016).
These include the depreciation function, the income equation, and the subsidy function.\textsuperscript{70}

The depreciation of health status, at each given age \(a\), is assumed to be an affine function of the natural logarithm of health status at that age. Formally, we assume that \(g(\cdot)\) in Equation (1) takes the form

\[
g(h,a) = -\left[a_\delta(a) + b_\delta(a) \cdot \ln(h)\right].
\] (37)

We assume that at each age \(a\), a linear function relates the natural logarithm of income to the logarithm of health status, as

\[
\ln(y(h,a,t)) = \bar{y}(a,t) + \varphi(a,t) \cdot \ln(h).
\] (38)

The term \(\bar{y}(a,t)\) in this equation captures the common component of income among the individuals of cohort \(t-(a-a)\) at age \(a\), whereas \(\varphi(a,t)\) characterizes the income variations arising due to the heterogeneity in health.\textsuperscript{71}

Following Guvenen and Smith (2010), after retirement, income is a function of the level of income at the age of retirement and the average income in the economy \(\bar{Y}\):

\[
\phi(y^R,t) = ay\left[y^R \bar{Y}(t)\right] + by\left[y^R \bar{Y}(t)\right] \cdot \left[y^R \bar{Y}(t)\right].
\] (40)

\textsuperscript{70}. This is mainly due to the fact that we find having these relationships at hand helpful to the flow of our discussions. Nevertheless, our methodology can be generalized to other assumed functions.

\textsuperscript{71}. In the presence of productivity shocks, we modify Equation (38) as

\[
\ln(y(h,\nu,a,t)) = \bar{y}(a,t) + \varphi(a,t) \cdot \ln(h) + \nu.
\] (39)

This is similar to the functional form considered by many in the literature for earnings and labor income, modified to include the impact of health status. Examples are Guvenen (2007, 2009), who instead of allowing \(\bar{y}(\cdot)\) to change freely with age, assume that the life-cycle profile of income is given by a quadratic function of age. In the health economics literature, Scholz and Seshadri (2011) consider the same functional form but abstract from the effects of health status on income.

Equation (39) is similar to the functional form considered by Fonseca et al. (2009), with the consideration that "health status" in Fonseca et al. is assumed to take discrete values. They, however, assume \(\bar{y}(\cdot)\) is a quadratic function of age.
As discussed in Section 3, the flow utility is assumed to take the CRRA form with a constant term—the value of being alive:

$$u(c) = b + \frac{c^{1-\sigma}}{1 - \sigma}. \quad (41)$$

For the health production function, we modify the CES functional form of Section 3 to allow for individuals’ age to have an effect on the probability of survival:

$$f(h, m, a, t) = A(a) \left[ \alpha [z(t) \cdot m]^{\gamma(a)} + (1 - \alpha) h^{\gamma(a)} \right]^{\frac{\alpha(\gamma)}{\gamma(\sigma)}}. \quad (42)$$

This form allows our model to nest the health production functions of Hall and Jones (2007) and Ales, Hosseini, and Jones (2014) as special cases. We will denote the growth rate of health technology $z(\cdot)$ by $g_z$.

Finally, we consider the following functional forms for the rate of subsidy, before and after the retirement:

$$s(y, a) = \begin{cases} [a_s \cdot \exp(b_s \cdot y)]^{-1} & \text{if } a \in [\bar{a}, a^R], \\ (a_s^R + b_s^R \cdot y)^{-1} & \text{if } a \in [a^R, \bar{a}]. \end{cases} \quad (43)$$

72. As mentioned before, Equation (43) summarizes many different government health programs in the United States which include, but are not limited to, Medicaid, Medicare, State Children’s Health Insurance Program (SCHIP), the Department of Defense TRICARE and TRICARE for Life programs (DOD TRICARE), the Veterans Health Administration (VHA) program, the Indian Health Service (IHS) program.

Before retirement, Medicaid is the dominant provider among these government programs. Because of Medicaid’s means-tested nature, its share in the total health spending declines rapidly by income, justifying the exponential form of subsidies in Equation (43). Our estimations show that this functional form in fact does an excellent job in representing United States health care subsidization programs before retirement.

After retirement, Medicare replaces Medicaid as the major public provider of health care services. While Medicare is not means-tested, two factors seem to be responsible for the share of total health expenditures paid by government entities to be declining in income. First, despite the dominant role of Medicare, Medicaid remains as a complimentary provider of services that are not covered by Medicare for lower income individuals. Second, Medicare’s provisions are not similar for all medical services. Therefore, the differences in the type of health services that are consumed by each income group result in the rate of subsidy to be non-homogeneous in income. (Our model does not capture these differences in the type of services that are demanded by each income group. See Ozkan 2014 for an example where this consideration is explicitly modeled.) Equation (43)
To estimate the parameters of these functional forms, we use an iterative procedure that follows the logic discussed at the beginning of Section 4: to search for a set of parameters under which the simulated data generated by the model looks as close as possible to the actual data, when viewed through the lens of an auxiliary model. To this end, we keep updating the values of the unknown parameters of the model until no further improvements can be achieved upon a “closeness” criterion. Some of the parameters of the model, however, are estimated or calibrated outside this iterative loop. We will explain these variables first.

4.2.1. Preset Parameters

As noted earlier, we assume that an interval of unit length in our model corresponds to a period of ten years in the data. Therefore, the two time periods under consideration in the MEPS data, 1996–2005 and 2006–2015, correspond to a time interval of length two in our model. For the sake of consistency (and convenience), we are going to denote the approximate midpoints of the two time periods by \( t_1 \) and \( t_2 \), in what follows: \( t_1 = 2000 \) and \( t_2 = 2010 \).

We set \( \bar{a} = 40 \). This value is consistent with the initial age in Chetty et al. (2016)’s sample. We assume individuals live up to \( \bar{a} = 100 \) years (Hall and Jones 2007).73 For the age of retirement, we choose \( a^R = 65 \). While, in our sample, there is a lot of variation in the age of retirement, this value ensures that the individual is eligible to receive Medicare compensations if \( a > a^R \).

The degree of risk aversion, \( \sigma \) in Equation (41), is set to 2.0, as it is the gold standard in the macroeconomics and finance literature. This is the value that has been widely used in the literature after Mehra and Prescott (1985)’s seminal work, and the parameter used by Ales, Hosseini, and Jones (2014). In their benchmark quantitative analysis, Hall and Jones use the same value. The value of \( \rho \) is chosen so that the annual discount rate is 0.98. \( r \) is set to match the average long-term rate of return on the United States treasury bills (that is does a good job in consolidating these factors up to a certain threshold (specifically, up to 400% of the federal poverty line). After this threshold, however, the model fit declines.

73. This number seems to be consistent with Chetty et al. (2016)’s upper bounds using Gompertz extrapolation for the 99% income percentile of their sample.
A consequence of our insistence on treating the health status as a latent variable is that our estimation strategy cannot identify the initial level of health status from the share parameter, \( \alpha \) in Equation (42), in the first time period. In addition, the growth rate of average health status cannot be identified from the growth rate of health technology, parameter \( z \) in Equation (42). One can see this by noticing that none of these parameters are invariant to the normalizations of health status: should the measurement unit of \( h \) change in Equation (42), these variables change as well. Therefore, in our simulations, we normalize \( \alpha \) to 0.1 and \( z (t_1) \) to 0.25. Following Hall and Jones, we assume \( z (\cdot) \) grows at the same annual rate as the long-run growth rate of GDP per capita in the United States economy (1960–2016); 2.03%.\(^75,76\)

The parameters \( a_y \) and \( b_y \) in Equation (40) are borrowed from Guvenen and Smith (2010), so that

\[
\phi \left( y^R, t \right) = \bar{Y} \left( t \right) \times \begin{cases} 
0.9 \times \bar{y}, & \text{if } \bar{y} \leq 0.3, \\
2.27 + 0.32 \times (\bar{y} - 0.3), & \text{if } 0.3 < \bar{y} \leq 0.2, \\
0.81 + 0.15 \times (\bar{y} - 2.0), & \text{if } 2.0 < \bar{y} \leq 4.1, \\
1.13, & \text{if } 4.1 \leq \bar{y},
\end{cases}
\] (44)

where \( \bar{y} := y^R / \bar{Y} \left( t \right) \). The MEPS data is used to estimate \( \bar{Y} \left( t \right) \) for \( t = t_1 \) and \( t = t_2 \). Outside the MEPS sample, \( \bar{Y} (\cdot) \) is assumed to grow at the same rate as the GDP per capita.

\(^74\) Note that, unlike most of the macroeconomic literature, the gross rate of return is not equal to the inverse of discount rate in our calibrations, as dictated by the Euler equation. This assumption, besides the fact that individuals are not infinitely lived in our economy, is mainly justified by the endogenous chance of mortality.

\(^75\) Our results do not show any change as a result of a change in \( z (t_1) \) or any significant change as a result of a change in \( \alpha \), confirming our claim that they act as normalization parameters.

\(^76\) This means \( g_z \), the growth rate of \( z (\cdot) \) in the model is chosen such that

\[
g_z \frac{10}{10} = \ln \left( \frac{z (t_1 + 0.1)}{z (t_1)} \right) = 0.0203.
\]

This is because of the assumption that an interval of length \( \Delta t = 1 \) in our model corresponds to ten years in the data.
Table 2: Preset Parameters

<table>
<thead>
<tr>
<th>Parameter</th>
<th>Value</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>$\sigma$</td>
<td>2.0</td>
<td>Mehra and Prescott (1985), Ales, Hosseini, and Jones (2014), Hall and Jones (2007)</td>
</tr>
<tr>
<td>$\rho$</td>
<td>0.2</td>
<td></td>
</tr>
<tr>
<td>$r$</td>
<td>0.32</td>
<td>United States’ Department of Treasury</td>
</tr>
<tr>
<td>$z(t_1)$</td>
<td>0.25</td>
<td></td>
</tr>
<tr>
<td>$\alpha$</td>
<td>0.1</td>
<td>Hall and Jones (2007), National Income and Product Accounts</td>
</tr>
<tr>
<td>$g_s$</td>
<td>0.20</td>
<td>Hall and Jones (2007), National Income and Product Accounts</td>
</tr>
<tr>
<td>$\bar{Y}(t_1)$</td>
<td>77,475</td>
<td>Medical Expenditure Panel Survey</td>
</tr>
<tr>
<td>$\bar{Y}(t_2)$</td>
<td>77,876</td>
<td></td>
</tr>
<tr>
<td>$g_Y$</td>
<td>0.20</td>
<td>National Income and Product Accounts</td>
</tr>
<tr>
<td>$a_s$</td>
<td>1.660</td>
<td></td>
</tr>
<tr>
<td>$b_s$</td>
<td>0.069</td>
<td>Medical Expenditure Panel Survey</td>
</tr>
<tr>
<td>$a^R_s$</td>
<td>1.384</td>
<td></td>
</tr>
<tr>
<td>$b^R_s$</td>
<td>0.013</td>
<td></td>
</tr>
</tbody>
</table>

Source: Authors’ calibrations. Note: $\alpha$ and $z(t_1)$ are normalization parameters. Values of $r$, $\rho$, $g_s$, and $g_Y$ are calibrated noting that one year in the data corresponds to a period of length 0.1 in the model. Parameters of the policy function are estimated by fitting the functional forms in Equation (43) to the share of total health expenditures that are paid by government entities, before and after the age 65, over the entire MEPS sample period, 1996–2015.

Finally, the parameters of the policy function, $a_s$, $b_s$, $a^R_s$, and $b^R_s$, are estimated so that the resulting subsidy schedule matches the average share of total expenses that are paid by government entities in the MEPS data as a function of income, before and after retirement. The tax rate $\tau(\cdot)$ is calibrated so that the government’s budget in Equation (22) is balanced in each period.

Table 2 summarizes the parameters that are estimated or calibrated outside our main SMM loop.
4.2.2. Simulated Method of Moments

The remaining structural parameters of the model consist of $b, \{A(a)\}_a, \{\gamma(a)\}_a, \{\beta(a)\}_a, \{a_b(a)\}_a, \{b_b(a)\}_a, \sigma_h, \{g(a, t)\}_{a,t},$ and $\{\varphi(a, t)\}_{a,t}.$ In the continuous time model of Section 2, $t$ and $a$ can take on all the values on the real line and in the interval $[\bar{a}, \bar{a}],$ respectively. Since it is not feasible to estimate the parameters that are functions of age and/or time over their entire domain, we have to restrict our estimates to certain cross sections. As our discussions of Section 4.1 suggest, for the time sections, we are going to estimate the parameters for $t \in \{t_1, t_2\}$ in the MEPS sample and $t \in \{1950, 1960, 1970, 1980, 1990, 2020, 2030\}$ outside the MEPS time span. For the variables that are functions of age, we limit our estimations to the average ages in each of the five age groups (as given in Table 1).

One can choose the parameters of the health production function, the utility function, and the income equation so that the moments generated through the model match those of two sets of auxiliary models: One characterized by Equation (36); and one relating the life expectancy to income, at different ages. But, this means that our estimator will be a vector of intractable dimensions, making the search for the global optima unfeasible.

To overcome this difficulty, as suggested before, we take another route and divide the set of structural parameters of our model into two subsets: The first set consists of parameters that are “estimated directly” to “target” the moments of the auxiliary model, Equation (36).

77. To summarize, for the parameters that depend on $t$ and $a$, the estimations are limited to

$$ (t, a) \in \left\{ (t, 44.36), (t, 54.14), (t, 64.22), (t, 74.19), (t, 83.65) ; \; t \in \{1950, \ldots, 2000\} \right\} \quad (45) $$

and

$$ (t, a) \in \left\{ (t, 44.61), (t, 54.30), (t, 64.11), (t, 74.12), (t, 83.33) ; \; t \in \{2010, 2020, 2030\} \right\}. \quad (46) $$

For the parameters which are assumed to remain the same over the two periods, we limit our estimators to

$$ a \in \{44.48, 54.23, 64.15, 74.15, 83.47\}. \quad (47) $$

78. In our simulations, when $a$ and $t$ fall between two sections $a_1$ and $a_2,$ and $t_a$ and $t_b$, we use a bi-linear interpolation of the parameter values at $(a_1, t_a), (a_1, t_b), (a_2, t_a),$ and $(a_2, t_b)$. 

46
We denote this set by $\Lambda$:

$$
\Lambda := \left\{ b, \{ A(a), \gamma(a), \beta(a) \}_a \right\}.
$$

(48)

The second set of parameters are estimated indirectly, conditioned on $\Lambda$. This set includes

$$
\Theta := \left\{ \sigma_h, \{ a_\delta(a), b_\delta(a) \}_a, \{ \bar{y}(a,t), \varphi(a,t) \}_{a,t} \right\}.
$$

(49)

The logic behind the estimation of $\Theta$ conditioned on a set $\Lambda$ is as follows: If we knew the values of the parameters in $\Lambda$ for the “true” underlying data-generating model, finding the values in $\Theta$ would boil down to several ordinary least squares (OLS) regressions. The reason is that, given $\Lambda$, we could use the joint distribution of health spending, income, and mortality rates to deduce the joint distribution of health status and income, using the health production function. This distribution (at different ages), in turn, could be used to estimate $a_\delta(\cdot)$ and $b_\delta(\cdot)$, at different ages which, itself, determines the evolution of the distribution of health status during the life-cycle of individuals. This distribution, together with the evolution of the distribution of income for each cohort, enable us to estimate $\bar{y}(\cdot)$ and $\varphi(\cdot)$ at different ages and different dates. The variance $\sigma_h$, then, could be chosen so that the variance of income as a function of age matches that in the data as closely as possible.

When $\Lambda$ is not known, for a given guess $\tilde{\Lambda}$, should the moments of the simulated data match those of the actual data as close as possible, we expect the joint distribution of income and health spending to be similar between the generated and actual data. Therefore, for the model to be able to also predict the joint distribution of income and life expectancy as it prevails in the actual data, $a_\delta(\cdot), b_\delta(\cdot), \bar{y}(\cdot)$ and $\varphi(\cdot)$ must take certain values. The same is true for the parameter $\sigma_h$.

This enables us to estimate $\Theta$ conditioned on a given guess for $\Lambda$. If $\tilde{\Lambda}$ does in fact result in a data generating machine that closely resembles the actual data generation process (as seen through the lens of the auxiliary model), we can rest assured that we have the “right” estimate for $\Theta$ as well.

What makes this “sequential” estimation procedure possible is the fact that, under the
assumed functional forms, there is a direct correspondence between some moments in the data and some of the parameters of the model (namely $\Theta$). Hence, given an estimate of the health status, it is easy to estimate the parameters in $\Theta$ using the conventional techniques. This logic reduces the number of unknowns that are directly chosen in the indirect inference method to 16. These are the parameters in $\Lambda$ which are chosen to target the 50 reduced form data moments that are derived from the auxiliary equations, Equation (36).

Estimating this over-identified set of moments requires the use of an efficient weighting matrix. However, as we will discuss in more details in the steps that follow, instead of trying to reduce the distance between the data and model moments using a weighting matrix, we minimize a Gaussian objective function, as suggested by Guvenen and Smith (2010).

**Step 1: Generating the Shocks** The starting period of the economy in our simulations is when the cohort of individuals who are 90 at $t_1$ enter the economy. We denote this starting point by $\bar{t} := t_1 - 50$. The final period of the simulations is denoted $\bar{t} = t_2 + 60$; the date at which the cohort $t_2$ reaches the age of $\bar{a}$.

In the first step, we generate a set of random shocks corresponding to each individual in our sample: For each individual in our sample, $i \in \mathcal{I}$, we generate a Wiener process of length $\bar{t} - \bar{t}$. We repeat this simulation $N$ number of times. (We pick $N = 10$.) This leads to $N$ sets of random shocks, each of size $|\mathcal{I}|$. We denote this set by $\mathcal{N}$,

$$\mathcal{N} := \{(i, n) ; i \in \mathcal{I}, n \in \{1, 2, \ldots, N\}\},$$

and the corresponding Wiener process by $\omega_h (\cdot ; i, n)$. For each $(i, n) \in \mathcal{N}$, $\omega_h (\cdot ; i, n)$ is the path of health shocks that affects individual $i$ during her lifetime. These simulated health shocks are going to remain fixed throughout our simulations.

**Step 2: An Initial Guess for $\Lambda$** We make an initial guess for the set of parameters that are estimated directly by targeting the moments of interest through the indirect inference approach. Let us denote this initial guess by $\Lambda_0$.

79. The number of individuals with non-zero income who are above 90 in the MEPS data is virtually zero.
**Step 3: Computing the Health status** Health spending for each age group and each time period $t_1$ and $t_2$ is given by the MEPS data at different income levels, $m(a, y, t)$. Given $\Lambda_0$ and the log mortality rate at different ages in $t_1$ and $t_2$ in each income ventile, $\log(\chi(a, y, t))$, we can compute the average health status in each income ventile and for different ages using Kmenta (1967)'s approximation of the health production function:  

$$
\log(\chi(a, y, t)) = \log(A(a)) + \alpha \beta(a) \log(z(t) \cdot m(a, y, t)) 
+ (1 - \alpha) \beta(a) \log(h(a, y, t)) 
+ \frac{1}{2} \alpha (1 - \alpha) \beta(a) \gamma(a) [\log(z(t) \cdot m(a, y, t)) - \log(h(a, y, t))]^2. 
$$  

(50)

**Step 4: Estimating the Depreciation Function** Assuming that, after the age of 40, there are no systematic movements between different percentiles of the log income during the course of individuals’ life, we use the average health status of individuals in a given income ventile who are $a$-years old in $t_1$ and the same variable for individuals who are $a + 1$ in $t_2$ to find an approximation for the depreciation function, $a_\delta(\cdot)$ and $b_\delta(\cdot)$ in Equa-

80. The use of Kmenta (1967)'s translog approximation of the health production function is to emphasize that all variables are in natural logarithms, and has no real bearing on our simulations.

81. This assumption is different from saying there are no movements between the different percentiles of log income. In fact, as Guvenen and Smith (2010) argue, there are differences in the growth rate of log income during the life-cycle. But, as long as these differences are not in a way that, on average, individuals of a given income ventile end up in a higher ventile in the next decade of their life, our assumption is valid. It is worth mentioning that this is the same assumption that Chetty et al. (2016) make, when using the mortality rate of an individual of age $a + 1$ in a given income percentile, as the future mortality rate of an individual in the same income percentile, but at age $a$. Chetty et al. (2016) argue that this is a reasonable assumption.
tion (37).\textsuperscript{82,83}

**Step 5: Computing the Evolution of Health Status** Using our estimates for the depreciation function and the distribution of health status for individuals of different cohorts in \(t_1\) and \(t_2\), we can use the law of motion of health status to construct the initial distribution of health status for each of the cohorts between \(t\) and \(t_2\), together with its evolution.\textsuperscript{84}

**Step 6: Estimating the Income Equation** The average log health status in each income ventile, at different dates and for different age groups, can be used in conjunction with the evolution of income distribution to estimate the parameters of the income equation, Equation (38).

**Step 7: Estimating the Variance of Health Shocks** Given \(\varphi (\cdot )\) at different dates and ages, we estimate the variance of health shocks \(\sigma_h\) such that the variance of log income in the model, on average, has the same age profile as the one estimated by Ales, Hosseini, and...

82. More precisely, the law of motion of health status implies that
\[
E_{h(t_1)} [\log (h(t_2))] = \int_0^1 [-a_\delta (a + t) - b_\delta (a + t) \ln (h)] dt.
\] (51)

Given \(E [\log (h(t_1))] = \log (h(a, y_v, t_1))\) at different income ventiles \((y_v)'s\), one can find \(a_\delta (\cdot )\) and \(b_\delta (\cdot )\) such that \(E [\log (h(t_2))] = \log (h(a + 1, y_v, t_2))\) in Equation (51). Under the assumption that \(a_\delta (\tilde{a})\) and \(b_\delta (\tilde{a})\) remain constant for \(\tilde{a} \in [a, a + 1]\), finding these parameters is rather easy. However, the assumption that for \(\tilde{a} \in (a, a + 1), a_\delta (\tilde{a})\) and \(b_\delta (\tilde{a})\) are interpolations of their values at \(a\) and \(a + 1\) makes the calculations more cumbersome.

83. A more accurate approach to the estimation of the depreciation function is to use Kolmogorov’s backward equation to write the empirical distribution of health status in \(t_1\) and age \(a - 1\) as a function of its distribution at \(t_2\) and age \(a\), the depreciation function, and \(\sigma_h\). Starting from an initial guess for \(\sigma_h\), one can iterate on Steps 4 through 7 to pin down \(a_\delta (a), b_\delta (b)\), and \(\sigma_h\).

This adds another estimation loop to an already numerically expensive problem that we want to avoid. Particularly, when \(\sigma_h\) is small (and the curvature of the distribution of health status is not large), we do not expect this step to add much to our estimations.

One should note that neither of these approaches takes into account the attrition due to mortality which we expect to be higher at lower levels of health status.

84. Cohort \(t_2\) is the last cohort that enters our economy.
Jones (2014) during the MEPS’ time period.\footnote{Note that, for a diffusion process, the variance of the sample path at any future date is equal to the product of the elapsed time and the variance of the underlying Wiener process. With a standard Wiener process, given the income equation and the law of motion of log health status (Equation (1)), the variance of income at age $a$ for individuals of cohort $t_0$ is given by}

This step concludes the estimation of $\Theta$ conditioned on $\Lambda_0$, if the model is to generate the joint distribution of income and mortality at different ages as observed in the data in $t_1$ and $t_2$.

**Step 8: Computing the Path of Health Status for Each Individual** We can, now, use the income equation to deduce the health status of each individual $\tilde{i} \in \mathcal{I}$. For any $\tilde{n} \in \{1, 2, \ldots, N\}$, the income equation and the law of motion of health status can be used to construct the path of health status for individual $\tilde{i}$, under $\omega_h (\cdot; \tilde{i}, \tilde{n})$:

$$h (\cdot; \tilde{i}, \tilde{n}) : [a, \bar{a}] \rightarrow \mathbb{R}_+.$$  \hspace{1cm} (53)

This is done for all the simulated Wiener processes in $\mathcal{N}$.

**Step 9: Finding the Optimal Markov Control** Given the functional forms and the support of the distribution of health status, we can now solve the individual’s problem, Problem (15), by finding the solution to the HJB equation. This, in turn, gives us the optimal feedback rule under $\Lambda_0$: $u_{\Lambda_0}$.

The computational approach that we take to solve the HJB equation is the Markov chain approximation method. We discuss this method briefly in Section 4.3 and leave the detailed discussion for the supplementary appendix.

**Step 10: Simulating the Path of Individual’s Health Expenditures** For each individual observation $\tilde{i} \in \mathcal{I}$ in the MEPS data, let us denote by $t_\tilde{i}$ and $a_\tilde{i}$ the corresponding
year (of the observation) and age (of the individual), respectively. For any \((i, n) \in \mathcal{N}\), we use \(u_{\Lambda_0}\), together with \(h(0; i, n)\) and individual’s cohort of entry \(t_i - (a_i - a)\), to construct the sample path of optimal health spending under \(\omega_h(\cdot; i, \tilde{t})\):

\[
m(\cdot; i, n) : [a, \bar{a}] \to \mathbb{R}_+.
\]

(54)

**Step 11: Consolidating the Simulated Data** For any \(n \in \{1, 2, \ldots, N\}\), we can use \(h(\cdot; i, n)\) in (53) (with the income equation) and \(m(\cdot; i, n)\) in (54) to construct the simulated pair of health spending and income at age \(a_i\), for individual \(i\), under the shock process \(n\):

\[
(y(a_i; i, n), m(a_i; i, n)).
\]

We do this for all \(i \in \mathcal{I}\) to consolidate the simulated data, given \(n\), in a set \(\text{sim}_{\Lambda_0}(n)\). Then,

\[
\text{sim}_{\Lambda_0} := \{\text{sim}_{\Lambda_0}(n); n \in \{1, 2, \ldots, n\}\}.
\]

In the next step, we use this simulated set to estimate the coefficients of the auxiliary model.

**Step 12: The Auxiliary Model** Now, we can estimate the coefficients of the auxiliary model using each of the simulated data sets: \(\text{sim}_{\Lambda_0}(n)\), for \(n \in \{1, 2, \ldots, N\}\). We do this separately for each age group and time period of Table 1, using the OLS method and will denote the resulting parameters by

\[
\hat{\beta}(\text{sim}_{\Lambda_0}(n)) := \left(\hat{\beta}_{0,t}^{a}(\text{sim}_{\Lambda_0}(n)), \hat{\beta}_{1,t}^{a}(\text{sim}_{\Lambda_0}(n)), \hat{\beta}_{2,t}^{a}(\text{sim}_{\Lambda_0}(n)), \hat{\beta}_{3,t}^{a}(\text{sim}_{\Lambda_0}(n))\right)_{a,t}
\]

and \(\hat{\sigma}(\text{sim}_{\Lambda_0}(n))\). We denote the average values of these estimates for different \(n\)’s by \(\bar{\beta}(\text{sim}_{\Lambda_0})\) and \(\bar{\sigma}(\text{sim}_{\Lambda_0})\):

\[
\bar{\beta}(\text{sim}_{\Lambda_0}) = \frac{1}{N} \sum_{n=1}^{N} \hat{\beta}(\text{sim}_{\Lambda_0}(n)) \quad \text{and} \quad \bar{\sigma}(\text{sim}_{\Lambda_0}) = \frac{1}{N} \sum_{n=1}^{N} \hat{\sigma}(\text{sim}_{\Lambda_0}(n)).
\]
The same equation is estimated using the “actual data” from the MEPS in the two time periods and for different age groups to yield the set of reduced form parameters $\hat{\beta} (data)$ and $\hat{\sigma} (data)$.

**Step 13: The Gaussian Objective Function** Our indirect inference estimator is the one suggested by Guvenen and Smith (2010). Following their notation, given the set of parameters $\hat{\beta}$ and $\hat{\sigma}$, define $\epsilon (\cdot)$ as

$$
\epsilon^a_{i,t} \left( \hat{\beta}, data \right) := m^a_{i,t} - \hat{\beta}^a_{0,t} - \hat{\beta}^a_{1,t} \cdot y^a_{i,t} - \hat{\beta}^a_{2,t} \cdot (y^a_{i,t})^2 - \hat{\beta}^a_{3,t} \cdot (y^a_{i,t})^3,
$$

where $m^a_{i,t}$ and $y^a_{i,t}$’s are from the MEPS data. (That is what $data$ in this definition stands for.) These residuals are used to compute the following Gaussian objective function:

$$
\mathcal{L} \left( \hat{\beta}, \hat{\sigma}, data \right) = \left( \frac{1}{2\pi \hat{\sigma}^2} \right)^{\frac{|\mathcal{I}|}{2}} \exp \left( -\frac{1}{2\hat{\sigma}^2} \sum_{i \in \mathcal{I}} \left[ \epsilon^a_{i,t} \left( \hat{\beta}, data \right) \right]^2 \right).
$$

Given these definitions, our indirect inference objective function is given by

$$
\mathcal{G}_{\Lambda_0} := \mathcal{L} \left( \hat{\beta} (data) , \hat{\sigma} (data) , data \right) - \mathcal{L} \left( \hat{\beta} (sim_{\Lambda_0}) , \hat{\sigma} (sim_{\Lambda_0}) , data \right).
$$

**Step 14: The Closeness Criterion** Our goal is to find the value of $\hat{\Gamma}$ (and corresponding $\hat{\Theta}_{\Gamma}$) to minimize $\mathcal{G}_{\Gamma}$. That is, our indirect inference estimator is given by

$$
\hat{\Gamma} := \arg \min_{\Lambda} \{ \mathcal{G}_{\Lambda} \}.
$$

To find the estimates of the parameters in $\Lambda$ and $\Theta$, we need to repeat this procedure, starting from Step 2, until our “closeness criterion” is met. This criterion is provided by the

---

86. As Guvenen and Smith argue, using (57) as the objective function “obviates the need to estimate an efficient weighting matrix.” Estimating this matrix, in our problem, is rather hard and time-consuming, making Guvenen and Smith’s approach very appealing.

Nevertheless, testing an objective function that simply minimizes the *Euclidean distance* between the parameters of the auxiliary model estimated separately using the actual and simulated data leads to the same results.
optimization algorithm of choice. We use a simulated annealing approach, for the reasons that are going to be discussed briefly in the next section.

4.3. Remarks on the Computational Approach

In practice, finding the indirect inference estimates of the model parameters using the SMM procedure of Section 4.2.2 boils down to choosing an optimization algorithm. Starting from an initial guess for $\Lambda$, such an algorithm recommends a direction of movement in each iteration of the above procedure, together with a closeness criterion.

With a control vector of length 16 and an objective function that is not very well-behaved, standard optimization algorithms (like Newton-Raphson, adapted for a multi-dimensional control space) do not, by themselves, guarantee a global optimum.

To ensure a global optimum, we use the simulated annealing method. In this algorithm, in each iteration of the SMM, conditioned on the value of the objective function $G_\Lambda$, there is a chance of an “uphill movement. This probability, however, depends on the system’s “temperature,” which asymptotically tends to zero.\(^{87}\) Starting from any feasible initial guess $\Lambda_0$, with a large number of repetitions of the SMM procedure from $\Lambda_0$, we are more confident that our results are, in fact, close to the global optimum of Problem (58).

From a computational standpoint, almost all of the numerical burden of the simulation procedure is on the HJB equation, Equation (16): This is a PDE in four individual and one aggregate states, making it extremely costly to solve. Even though the assumption that $x$ is governed by a diffusion process simplifies the first order conditions on the right hand side of the equation to some extent, it should not be forgotten that the “actual” underlying state is still a jump-diffusion with controlled jumps.\(^{88}\) This results in highly non-linear optimality conditions and, consequently, adding to the numerical intensity of the problem.

To alleviate these difficulties, we start our global search by assuming no cohort effects: that is, in each time interval $t_i$, we assume that the economy is in a stationary equilibrium.

\(^{87}\) In our numerical simulations, the temperature of the system is assumed to follow a simple reciprocal form. In each iteration, the probability of an uphill movement is proportional to $\exp(-G_\Lambda/temp)$.

\(^{88}\) The intensity of the mortality rate is still controlled by the health spending.
In addition, we assume $\sigma_h = 0$ in the benchmark model. This, in turn, eliminates the need to keep track of $h^R(\cdot)$ as a state variable after retirement. With two fewer states, we can find the solution to the HJB equation rather quickly.

After a very thorough global search for the best candidates for $\Lambda$, under these simplifying assumptions, we initiate several local searches starting from the global candidates. The local searches are performed under the complete set of assumptions until no further improvement seems feasible.\textsuperscript{89}

An extensively used method for solving the HJB equation is the finite difference (FD) method. In this approach, the PDE in (16) is approximated by a discrete equation.\textsuperscript{90} In this paper, however, we take a novel approach, known as the Markov chain approximation method. In this method, which is developed by Kushner and Dupuis (2014), instead of approximating the PDE itself, we approximate the underlying state vector $x$ by a Markov chain. Then, the individual’s problem is written for this approximating chain. This problem is a functional equation (known as the Bellman equation) that can be solved iteratively. Importantly, unlike the FD method, under some regulatory conditions on the approximating chain—called the local consistency conditions—the solution to the Bellman equation is guaranteed to converge to the solution to the HJB equation.\textsuperscript{91} We discuss this method in more

\textsuperscript{89} For the full economy, a complete cycle of the SMM procedure takes approximately twenty minutes on a workstation (with twelve cores working in parallel under OpenMP directives). Simplifying the economy, as explained above, reduces this time to less than two minutes. Starting from the simplified economy allows us to pin down the optimum—from a relevant initial guess—in about three months on the same station. Using the Minnesota Supercomputer Institute (MSI)’s Linux cluster, we do this in less than two weeks using ten nodes working under MPI directives.

\textsuperscript{90} See Achdou et al. (2014) and Tourin (2010) for excellent discussions.

\textsuperscript{91} Approximating a diffusion process by a discrete Markov chain is not, by any means, equivalent to starting from a general Markovian process. As Dixit (1993) notes, a discrete representation of a diffusion process is a random walk that satisfies $\Delta h = \sigma \sqrt{\Delta t}$, where $\Delta h$ is the size of the spacial jumps and $\Delta t$ is the size of the temporal grid.

The fact that the discretization takes the form of a random walk, however, has strict implications for the controls: They can only change the probability of upward or downward movements, but cannot affect the size of the jumps. Consequently, the first order conditions would, in general, be considerably simpler than those of a discrete-time economy in which the states can, in theory, move freely.

However, in the presence of controlled jumps, these approximations lose their attraction, at least to some extent.
Table 3. Estimation Results

<table>
<thead>
<tr>
<th>Age Group</th>
<th>40–49</th>
<th>50–59</th>
<th>60–69</th>
<th>70–79</th>
<th>80–89</th>
</tr>
</thead>
<tbody>
<tr>
<td>Elasticity of Scale, $\beta (a)$</td>
<td>0.40***</td>
<td>0.22**</td>
<td>0.11*</td>
<td>0.10</td>
<td>0.10***</td>
</tr>
<tr>
<td></td>
<td>(0.09)</td>
<td>(0.10)</td>
<td>(0.06)</td>
<td>(0.08)</td>
<td>(0.03)</td>
</tr>
<tr>
<td>Cross Elasticity, $\gamma (a)$</td>
<td>0.375***</td>
<td>0.21***</td>
<td>0.16</td>
<td>0.20***</td>
<td>0.20***</td>
</tr>
<tr>
<td></td>
<td>(0.08)</td>
<td>(0.08)</td>
<td>(0.13)</td>
<td>(0.03)</td>
<td>(0.07)</td>
</tr>
</tbody>
</table>

Source: Authors’ SMM estimates. Note: The elasticity of scale and the cross elasticity of health production with respect to health spending and health status are estimated, together with the value of being alive and the factor productivities, using the SMM. The targeted moments are those of the auxiliary equation, Equation (36). See Section 4.2 for a detailed explanation. The standard errors are computed using the bootstrap method with ten re-samplings of the data. See Footnote 92 for more details.

detail in the supplementary appendix.

5. The Results

Table 3 summarizes our estimation results for two of the parameters of interest: the elasticity of scale and the cross elasticity of health outcomes with respect to health spending and health status. Our estimate for the value of being alive $b$ is 110 (with a 95% confidence interval of (97.68, 122.32)).

Except two instances, all the estimates are statistically significant. These two are the elasticity of scale for 70–79 year-old individuals and the cross elasticity for the age group 60–69. Moreover, except for the 40–49 age group, our estimates of the elasticity of scale are close to Hall and Jones (2007)’s estimates. This parameter follows the same declining trend as their results after the age of 40 suggest.

92. At the time of writing this draft, the standard errors are computed using the bootstrap method with only ten re-samplings. However, in each of the re-samplings, the SMM procedure is restricted to a local search around the “true” candidate to avoid the extremely costly global search. See Footnote 89 for more details.
Importantly, our estimates of the elasticity of substitution are above one, ranging from 1.60 to 1.25, depending on age.\textsuperscript{93} This implies that health status and health spending are relatively strong substitutes. In turn, this means that the effectiveness of health spending, in extending life, is relatively low at higher levels of health status.\textsuperscript{94}

Therefore, as our discussions of Section 3 and, in particular, Proposition 5 suggest, if health status and income are strongly correlated among the individuals of a given cohort, high-income agents tend to have lower health spending (even in absolute terms) than low-income individuals.

This is best shown in Figure 2, which depicts the model’s fit in the two time periods of interest relative to the data. The model performs relatively well among the individuals of the depicted cohorts. In particular, our estimates can capture the declining share of health spending in the cross section and its increasing share in the time series: a necessity in the cross section and a luxury over time.

An interesting observation is that, while income has not increased by much between the two time periods in our data among the individuals of ages 40 to 49, the model suggests that they still increase their spending by a relatively large margin (as it does in the data). This can be attributed to two factors: an increase in $z$ which, in effect, leads to a decline in the relative price of health spending; and individuals’ expectations of higher future income levels (as the average income depicts a significant increase at older ages in the sample). As a result of this expected rise, individuals suppress their savings early in life—to some extend—which frees up resources to be allocated to health care.

After the age 70, the model fit starts to diminish. In particular, our model predicts very large levels of spending at ages over 80. The deterioration of model fit after the age of 80

\textsuperscript{93} We believe that the decline in the substitutability of health spending and health status is rather intuitive. Specifically, at lower ages, the medical services that an individual receives tend to be closer in nature to a replacement for the underlying health: a heart valve surgery, an insulin injection, or an artificial limb, are all substitutes for a functioning organ. While these health services remain a major determinant of health spending as individual ages, they tend to become effective only if the underlying health has not deteriorated greatly.

\textsuperscript{94} Results of Table 3 suggest that, after the age of 60, the absolute value of the marginal product of health spending declines with health status. This is beside the decline in the marginal product relative to the average product (which is the main determinant of health spending).
**Figure 2.** Model Fit: Health Spending Relative to Income

*Source:* Authors’ simulations & MEPS.
can be attributed to several factors: First and foremost, the Gompertz equation used to infer the mortality rates at different ages looses its predictive power at old ages. Second, the variations in mortality rates among different income groups vanishes after a certain age, making our inferences on the health status at older ages even more inaccurate. Finally, our sample sizes decline considerably for the last two age groups (as Table 1 suggests).

Figure 3 illustrates the predictive power of the model in the time series. Each point on the figure illustrates the average log health spending in a given year during the MEPS sample, versus the average log income in the same year. While the model can match the rising share of health spending over time (through the slope of the fitted line, which is significantly greater than one), it misses the level of spending. This, as we mentioned above, is mainly because of the prediction of very high levels of spending at older ages (and among the bottom 2% of the population).

Our results suggest that initial health status has, in general, improved between the two time periods, as depicted in Figure 4. In spite of an increase in general underlying health, this rise has not been large enough to dominate the growth in income (or, at least, the
expectations thereof) or technology. Otherwise, as our discussions in Section 3 suggest, we should have observed a decline in the share of health spending between the two periods. Finally, even though the general health status has improved across the two cohorts in $t_1$ and $t_2$, this increase has not been the same for all income groups, as the distribution in Figure 4 has become significantly more skewed to the left. Such an increase in the variance of initial health status is responsible for the rising gap of longevity across the income groups, as documented by Chetty et al. (2016).

5.1. Marginal Cost of Saving a Life

Table 4 depicts the marginal cost of saving a statistical life for different age groups and at the two time periods. The cost of saving a statistical life, at a given age (and time), is the total amount of resources that are required to reduce the average number of deaths, at that age, by one.

If we assume $\zeta$ denotes the marginal effect of health spending on the rate of mortality (at
a given age), the resources required to prevent one death among the whole population—that is to save one statistical life—is exactly \( 1/\varsigma \). In the context of our model, this is given by

\[
\frac{f^2(m, h, a, t)}{\partial f(m, h, a, t) / \partial m},
\]

for a given age and at a given time.

To compute this value for a given age group, we calculate (59) using our estimates of the health production function and constructed health status, under the “actual” level of health spending, at different ages within an age group. These are compounded to compute the cost of saving a statistical life, as in Table 4.

Table 4 demonstrates these results at four different income levels: bottom and top 5%, median, and top 20%. The last two rows of the table provides a rough estimate for the values of statistical life (VSL) that are widely used in the literature, as function of age and time.95

Some remarks are in order. First, the general trend of marginal cost of saving a life, over the life-cycle, is what one expects: as individuals get older, it becomes significantly more expensive to save one statistical life. Aldy and Viscusi (2008)’s results show the same trend: while estimated using a different approach, the trend of VSL after the age of 40 is similar to ours. In addition, the marginal cost of saving a life has increased dramatically across the two periods (except for the bottom 5% of income distribution). The same trend is apparent in Hall and Jones (2007) and Aldy and Viscusi (2008)’s estimates.

Third, except within the first age group in which our results are significantly larger than the conventional estimates of the VSL, our estimates of the cost of saving a life for the median agent in our sample is comparable to the VSLs that prevails in the literature. If we assume that the VSL is an index for the social value of life, this implies that health expenditures have been—and still are—in their “efficient” range. In particular, if we accept the trend of the marginal costs in Table 4, then, it appears that the average American is

95. The values for the year 2000 are from Aldy and Viscusi (2008)’s VSL estimates at different ages. (All of Aldy and Viscusi’s estimates seem to fall around the midpoints of the estimates in the literature, as reported by Viscusi 2003.) For the year 2010, the growth rate of total VSL, estimated by Felder and Werblow (2009) for Switzerland, was applied to the values in the United States to provide a rough picture.
## Table 4. Cost of Saving a Statistical Life (thousands USD)

<table>
<thead>
<tr>
<th>Age Group</th>
<th>40–49</th>
<th>50–59</th>
<th>60–69</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>1996–2005:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottom 5%</td>
<td>6,570</td>
<td>2,487</td>
<td>919</td>
</tr>
<tr>
<td>Median</td>
<td>22,737</td>
<td>9,347</td>
<td>2,660</td>
</tr>
<tr>
<td>80th Percentile</td>
<td>247,836</td>
<td>52,484</td>
<td>17,786</td>
</tr>
<tr>
<td>Top 5%</td>
<td>754,643</td>
<td>122,555</td>
<td>33,733</td>
</tr>
<tr>
<td><strong>2006–2015:</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bottom 5%</td>
<td>5,566</td>
<td>2,415</td>
<td>809</td>
</tr>
<tr>
<td>Median</td>
<td>67,778</td>
<td>18,608</td>
<td>6,555</td>
</tr>
<tr>
<td>80th Percentile</td>
<td>1,843,551</td>
<td>290,139</td>
<td>82,398</td>
</tr>
<tr>
<td>Top 5%</td>
<td>7,205,732</td>
<td>856,369</td>
<td>190,652</td>
</tr>
</tbody>
</table>

### Value of Statistical Life:

<table>
<thead>
<tr>
<th></th>
<th>2000</th>
<th>2010</th>
</tr>
</thead>
<tbody>
<tr>
<td>2000</td>
<td>11,800</td>
<td>9,900</td>
</tr>
<tr>
<td>2010</td>
<td>20,400</td>
<td>16,400</td>
</tr>
</tbody>
</table>

**Source:** Authors’ calculations. **Note:** The cost of saving a statistical life at age $a$ is given by Equation (59). The cost, in each age group, is computed as the compounded cost of saving a statistical life at different ages in the age interval, using the indirect inference estimates. For each time period, the actual health spending and the constructed health status in each of the income groups are used. The value of a statistical life is provided for comparison. The values are from Aldy and Viscusi (2008) for 2000, and adjusted by the growth rate estimates of Felder and Werblow (2009) for 2010. VSLs are adjusted by the GDP deflator.
spending more than what is efficient at early ages and less than the efficient levels when she gets older.

Four, the cost of saving a statistical life is dramatically different across different income groups. This observation, which goes to the root of this study, results from the claim that, at any given level of technology, the effectiveness of health spending declines dramatically with underlying health; an observation that has been extensively ignored so far in the literature. With a strong correlation between income and health status, this means that a considerably larger sum of resources are required to save an individual at the top of the income distribution compared to one at the bottom. This observation is going to play a central role in the policy analysis of the next section.96

Finally, the value of statistical life has not, in general, risen in the bottom of the income distribution. This is despite the fact that both health spending and health status have grown among this group (though not by much). The main reason is that the rise in spending and health status have fallen behind the technological innovations over this period. This has led to a rise in the marginal product of health spending and, consequently, not a major shift in the marginal cost of saving a life within this group.

5.2. Revisiting Identification

As we discussed in detail in Section 3, the growth of income over time translates to a rise in the share of health spending through its effect on the elasticity of utility relative to the elasticity of health with respect to spending. On the other hand, the downward-sloping schedule of health spending in the cross section arises as a result of a declining productivity of health spending. This decline, in turn, is due to a strong correlation between income and health status. This logic comprises the bases of our identification strategy in Section 4.

To demonstrate this, in Figures 5 and 6, we revisit our initial claims regarding the patterns of health spending in the cross section and over time, together with model’s performance

96. This statement, by no means, should be taken to imply that the resources allocated to health at the top of the income distribution are extensive and inefficient: without a comparable study that investigates the VSL at different levels of income, such conclusions can be greatly misleading.
in accounting for them.

In Figure 5, we have used an extrapolation of income and health for an average American to extend our results back to 1960. The figure depicts the NHEA data, as well as the model simulations, for the estimated parameters and under the assumption that the value of being alive is underestimated by 90%. As our previous discussions suggest, the value of being alive plays a crucial role in determining the trend of health spending over time. This point, which is one of Hall and Jones (2007)'s important contributions, is best understood through Equation (32).

Figure 6 compares the model’s performance in the cross section, for the 40–49 year old individuals in the first time period, to the MEPS data. As the figure suggests, an underestimation of the elasticity parameter, $\gamma$ by 90% (at all ages) leads to an upward sloping spending schedule in the cross section: an observation that is in stark contrast with the data.

97. As noted earlier, the health spending data in MEPS underestimates the total health spending by a constant margin, as compared to the NHEA, because of the exclusion of some health services. For this reason, the model simulations have been shifted upward for presentation purposes.
At the end, it is worth mentioning that none of our main parameters have an isolated effect on the cross sectional or time series trends in our simulations. Nevertheless, one can conceivably argue that the effect of each of these parameters is more pronounced on one aspect of the results.

6. Implications for Policy

To the extent that a downward sloping health spending curve in the cross section is caused by a declining productivity of health care, our results have clear implications for health care policy. Importantly, the interaction between health status and productivity of health care expenditures has been largely ignored in the public finance literature on the consequences of health care reforms.

To separate the role of productivity differentials in the cross section, in Figure 7, we compare the share of health spending (relative to income) in the model to that in the data, for
Figure 7. Health Spending in the Cross Section: the Role of Policy (40–49 year olds in 1996–2005)

Source: Authors’ simulations & MEPS.

four income quartiles of the sample after eliminating the subsidies on health spending during the entire life-cycle \( s(y, a) = 0 \). (Only the initial age group, 40–49, is illustrated for the sake of presentation.) As the figure suggests, to the extent that our model is a relevant description of the underlying mechanisms, policy plays an important role in the relatively high levels of health spending in the first income quartile. However, this role diminishes quickly with income. All of slight changes in the spending at the top median are virtually because of the elimination of Medicare subsidies after the retirement which forces individuals to increase their saving when they are young.

Figure 7 ascertains our claim that most of the decline in the level of health spending in the cross section is because of a decline in the marginal productivity of health spending in extending life. If this effect of health status on the marginal product of health spending is ignored, any policy evaluation exercise results in an overestimation of the favorable impli-

98. If our model does not capture the important aspects of the actual world, Figure 7 is simply the corollary of a proposition regarding a statement whose inaccuracy forms the bases of the mentioned proposition!
cations of the policy for “wealthy and healthy” individuals. If, for instance, it is assumed that health status has no effect on the marginal product of spending (as is the case under the assumption that $\gamma = 0$ in Equation (31)), starting from the status quo, an increase in the health care subsidies at all levels of income would be viewed as having the same life-extending effect for all individuals, regardless of their income. However, as our results in Table 4 so clearly illustrate, the cost of saving a life at the top of the income distribution is dramatically higher than that for individuals from low-income groups. The “optimal policy,” then, is a matter of the planner’s weighting scheme, as well as the productivity of these individuals.99

In the next section, we will compare the effects of two policy shifts that are meant to represent two of the popular policy proposals in the United States: an extension of the United States health care policy that currently prevails after the retirement to all ages; and an expansion of the current pre-retirement health care policy for the low-income individuals to cover more services. With a slight abuse of terminology and for the lack of better terms, we will refer to these policy proposals as “Medicare for all” and “Medicaid expansion,” respectively.100 Not to draw normative judgments about the Pareto weights that a planner might assign to each individual in the economy, we will discuss the effects of each of these proposals on different parts of the income distribution.101

99. Ales, Hosseini, and Jones (2014) is a seminal paper that takes the differences in productivity into account, when considering the problem of an Egalitarian planner.

100. We should emphasize that neither Medicare nor Medicaid comprise “all” of United States post- or pre-retirement health care policies. However, as mentioned before, they are the most important public providers after and before the age of retirement, respectively. Refer to Footnote 72 for a more detailed explanation.

101. This does not mean that a policy can only be justified from a redistributive standpoint in our economy. In the presence of incomplete markets and, in particular, uninsured health shocks, there might exist a policy intervention that is Pareto improving from an ex ante sense: Even if income was fully insured against idiosyncratic shocks, when the cross-elasticity of health outcomes with respect to health spending and health status is non-zero, an individual might be willing to enter into a contract to transfer resources from one state of the world to the other. Full characterization of such a policy, however, is outside the scope of this paper.
6.1. Medicare for All vs Medicaid Expansion

Starting from the status quo policy, \( \tau (\cdot) \) and \( s (\cdot) \), consider two policy shifts: (i) Medicare for all, which is an extension of the post-retirement health policy to all ages, denoted by \( s_1 (\cdot) \); (ii) and Medicaid expansion, \( s_2 (\cdot) \), under which the rate of subsidy for low-income families is increased (in a way that it delivers the same level of welfare to the lowest income individual in the sample as under Medicare for all, considering the required income tax), while keeping the rate of subsidy for the top-earners unchanged. Formally,

\[
s_1 (y) = \left( a_s^R + b_s^R \cdot y \right)^{-1}
\]

and

\[
s_2 (y, a) = \begin{cases} 
[a_s' \exp (b_s' y)]^{-1} & \text{if } a \in [\bar{a}, a^R), \\
(a_s^R + b_s^R \cdot y)^{-1} & \text{if } a \in [a^R, \bar{a}],
\end{cases}
\]

where \( a_s^R \) and \( b_s^R \) represent the status quo—given in Table 2. Both policies are assumed to be financed through an increased tax on income so that the government’s budget, Equation (22), is balanced. We denote the resulting tax rates by \( \tau_1 (\cdot) \) and \( \tau_2 (\cdot) \), so that

\[
\tau_i (t) = \tau (t) + \Delta \tau_i (t).
\]

For the estimated parameters of the model, the coefficients of the policy functions in Table 2, and \( a_s' = 1.35 \) and \( b_s' = 0.07 \), these policy shifts entail an increase in the income tax rate of 5.8 and 0.8 percentage points, respectively, in the first decade (1996–2005): \( \Delta \tau_1 (t_1) = 5.8\% \) and \( \Delta \tau_2 (t_1) = 1.0\% \). The resulting subsidy schedules are depicted in Figure 8. The solid green line in this figure illustrates the rate of pre-retirement subsidy on health care—as a function of income—under the current health care policy in the United States. The red dashed line, on the other hand, is the rate of subsidy on health services that is only applicable after the age of retirement. The blue dotted line depicts the proposed increase in the subsidy rates before retirement.\(^{102}\)

\(^{102}\) The Medicaid expansion is calibrated so that the rate of subsidy received by the lowest-income family in the sample is equal to that under Medicare for all, but the high-income families receive the same rate before
Figure 8. Medicaid Expansion vs Medicare for All: Health Spending Subsidy by Income

![Graph showing Medicaid Expansion vs Medicare for All health spending subsidy by income.]

**Source:** Authors’ simulations.

Figure 9 shows the effect of these policy changes on the health spending of each of the income groups during the first decade of life. Both policies lead to an increase in the share of spending in income because of their price effects. As one would expect, Medicare for all has a larger impact on spending than Medicaid expansion.

However, the increased health spending does not necessarily result in improved welfare for all income groups as Figure 10 illustrates. This figure shows the percentage change in the lifetime stream of consumption (relative to the status quo) that results in the same change in expected lifetime utility for the individuals that enter the economy in \( t_1 \). As the figure suggests, both policies entail a large and positive welfare impact on low-income families. The reason for this large improvement in welfare is the relatively large impact of increased spending on the probability of survival for these groups.\(^{103}\) Consequently, despite the after-

103. Our simulations show that both policies are associated with a 50-day increase in the life expectancy of an individual at the very bottom of the income distribution.
Figure 9. Medicaid Expansion vs Medicare for All: Effect on Health Spending by Income (40–49 year-olds in 1996–2005)

Source: Authors’ simulations & MEPS.
tax decline in income, low-income individuals are significantly better off.

Even though both policies have a similar positive impact on the lower tail of the income distribution, these welfare gains diminish quickly. Under Medicaid expansion, welfare effects become negative after the 12th percentile of income, while under Medicare for all this happens at the 17th percentile.

The dissipation of the positive welfare effects under Medicare for all can be understood in light of our results for the cost of saving a life, Table 2: As the underlying health improves with income, the effectiveness of health spending in expanding life diminishes. This is most felt at younger ages when high-income individuals enjoy a considerably better health. Importantly, these are the exact same age groups that the proposed policy targets. As a result, high-income individuals gain very little in terms of reduced chance of mortality following the policy’s implementation.

However, the increased income tax means that, not only they can not spend as much as before on consumption, but also they have to give up part of their savings which constitute
the source of most of their health spending at old ages when health status has depreciated by a great deal. The total effect is a decline in life expectancy, rather than an increase, for the high-income groups.\(^{104}\)

For the very wealthy individuals, it is hard to replace this loss in expected life-years by consumption. As a result, we observe a surprising decline in welfare that, for the top-earners, surpasses the income loss due to taxes. A quote from Hall and Jones (2007) provides a nice intuition for this:

“As we get older and richer, which is more valuable: a third car, yet another television, more clothing—or an extra year of life?”

The intuition for the effects of Medicaid expansion is, more or less, the same: Even the middle income individuals who do receive some of the fruits of health care reform (in terms of increased subsidies), do not gain much due to the lower productivity of health care for them. The welfare losses, however, at the top of the income distribution are dwarfed by those under Medicare for all because of considerably lower tax rates that are required by Medicaid expansion.\(^{105}\)

The punchline in our arguments in this section is that the key contributor to the mechanism that delivers the flat cross sectional Engel curve in our model, that is the substitutability between health status and health spending, also plays a crucial role in determining the optimal direction that health care policy should take. In the absence of cross-effects between health status and health spending—when the marginal cost of saving the life of a high income individual is similar to that of a low income person—health spending is more valuable to a high income, high productivity individual. This is true from the perspective of both an

\(^{104}\) Based on our calculations, the life expectancy of an individual at the 99th percentile of income decreases by 5 days, after implementing Medicare for all.

\(^{105}\) It is important to remember that there are two sources of overestimation of (favorable) welfare effects in our calculations which stem from the inelastic supply of labor in our model: First, an increase in the labor income tax would have an additional negative income effect due to its impact on the supply of labor. Second, the increases in the income tax rates that are required to support the proposed policies are likely underestimated because of the fact that they do not take the resulting price effects into account. These two effects are likely more important in the first policy shift, Medicare for all, since it is more resource intensive.
As a result, any policy evaluation that fails to account for such differences can lead to misleading conclusions.

Finally, Figure 11 illustrates the inverse of the marginal utility at age \( a \) under the status quo policies. This variable can be interpreted as a rough estimate for the Pareto weights that are required for a planner to deliver the observed level of consumption under the current system—if everything was observable. The authors believe, without having to draw normative conclusions, thinking about the political aspects of implementing each of these policies will be misleading if one does not take such weighting schemes into account.

106. To see how adding cross-elasticity considerations to Ales, Hosseini, and Jones (2014) can alter the direction of optimal health spending from the perspective of an Egalitarian planner, consider the simplified economy of Section 3. Suppose the initial distribution of health status is given by a two-point distribution of the form \( \Gamma (h_0) = \frac{1}{2} \) if \( h_0 \in \{ b, \bar{h} \} \) and \( \Gamma (h_0) = 0 \) otherwise, where \( b < \bar{h} \). Then, the problem of an Egalitarian planner, when \( r = 0 \), is given by

\[
\begin{align*}
\max_{c(h_0), m(h_0)} & \sum_{h_0} \frac{1}{2} f (h_0, m(h_0)) \cdot u (c(h_0)) \\
\text{s.t.} & \sum_{h_0} \frac{1}{2} f (h_0, m(h_0)) [y (h_0) - c(h_0) - m(h_0)] \geq 0.
\end{align*}
\]

The first order conditions to this problem imply that, in the optimum, \( c(h_0) = c^* \) for all \( h_0 \) (as in Ales, Hosseini, and Jones) and the planner equalizes

\[
\frac{f (h_0, m(h_0))}{f_m (h_0, m(h_0))} + c^* + m(h_0) - y(h_0)
\]

across all individuals.

Assuming a CES form for the health production function, when \( \gamma = 0 \), these conditions imply that the optimal health spending is proportional to the individuals' income, as claimed by Ales, Hosseini, and Jones. However, for large enough \( \gamma \) and a strong enough correlation between income and health status, this result no longer holds.

107. In a simulation exercise, we examine the effect of the mis-estimation of parameter \( \gamma \) on the results of this section. Our findings confirm the importance of considering the cross-elasticity of health outcomes with respect to health status and health care spending. For instance, a 20% underestimation of the parameter \( \gamma \) at all ages leads to about 20% underestimation of the welfare effects of Medicare for all for the lowest-income individuals.
7. Concluding Remarks

We develop a life-cycle model with heterogeneity in income and health status, where individuals allocate their income between consumption and health spending. While consumption determines the flow of utility through a standard utility function that incorporates the value of being alive, health spending and health status enter a health production function to determine individuals’ longevity.

In our framework, for a given level of health status, the growth of income over time leads to a decline in the value of consumption relative to the value of being alive. This relative change creates a luxury-good channel that causes the share of health spending in income to increase over time. On the other hand, a strong correlation between income and health status leads high-income individuals to devote fewer resources to health spending. The reason is that, if health status and health care are substitutable, the marginal effect of one dollar health spending on lifetime utility is smaller for wealthier and healthier agents.

These two channels enable our model to account for the conflicting patterns of health
spending in the cross section and time series. We take advantage of the distinct implications of each channel for the cross section and time series to estimate the structural parameters of the model. We use this insight to estimate the parameters of a health production function using income variations in the cross section and over time in the United States. Our estimation results confirm that the elasticity of substitution between health spending and health status is significantly above one at all ages under consideration—an observation that has been largely neglected in the literature.

Substitutability of health status and health expenditures has important implications for the effects of health care policy. We show this by comparing the welfare consequences of two popular policy proposals: an expansion in the pre-retirement health spending subsidies for low-income families—Medicaid expansion—and an extension of the post-retirement United States health care policy to all ages—Medicare for all. Both of these policies entail positive and comparable welfare effects for the lower income individuals. However, our finding that the value of health care is low for high-health status individuals implies that wealthy individuals have little to gain from increased subsidies on health expenditures under Medicare for all. Therefore, the much larger tax increments required to finance Medicare for all lead to greater welfare losses at the top of the income distribution, compared to Medicaid expansion.

Finally, two components missing from our framework are an endogenous process for the accumulation of health status and a mechanism accounting for the initial differences in health status. A vast literature in health economics relates the health differences among individuals later in life to the early-life environment suggesting that, to account for initial heterogeneity in health, one has to incorporate intergenerational links and altruistic motives into the model. A seminal paper that models the life-cycle investment in health capital is Ozkan (2014). In his framework, preventive health capital plays the role of health status, and individuals decide to invest in preventive health to avoid unfavorable health outcomes in the future.

Central to these endogenous channels of health capital variations is an investment function. We believe our approach in using various patterns of health expenditures and health outcomes in the data to quantify a life-cycle model can be generalized and applied in these
frameworks to discipline such investment functions.108

References


108. For instance, in Ozkan’s model, if individuals are only heterogeneous regarding their initial income and the marginal product of investment in health capital is large enough, we should expect a rapid rise in the preventive care utilization by lower-income individuals as their income grows over time. Similarly, if investment in children’s future health has diminishing returns, we should expect a regression to the mean for the health outcomes of different income groups (an implication that is rejected, partly, by rising gap in the life expectancy in the United States).


Tourin, Agnes. 2010. *An Introduction to Finite Difference Methods for PDEs in Finance*.
