Is There "Too Much" Inequality in Health Spending Across Income Groups?*

Laurence Ales
Carnegie Mellon University
ales@cmu.edu

Roozbeh Hosseini
University of Georgia
roozbeh@uga.edu

Larry E. Jones
University of Minnesota
lej@umn.edu

Abstract

In this paper we study the efficient allocation of health resources across individuals. We focus on the relation between health resources and income. In particular we determine the efficient level of the health care social safety net for the indigent. We assume that individuals have different life cycle profiles of productivity. Health care increases survival probability. We adopt the classical approach of welfare economics by considering how a central planner with an egalitarian (ex-ante) perspective would allocate resources. We show that, under the efficient allocation, health care spending increases with labor productivity, but only during the working years. Post retirement, everyone would get the same health care. Quantitatively, we find that the amount of inequality across the income distribution in the data is of the same magnitude as what would be justified solely on the basis of production efficiency. As a rough summary, in U.S. data the top to bottom income quartile spending ratio is about 1.3 for most of the working life, dropping to 1 at retirement. Efficiency implies a steady decline from about 1.3 (at age 25) to 1 at retirement. We find larger inefficiencies in the lower part of the income distribution and ages between 50 to 65.

JEL Classification: I14, H21, E60.
Keywords: Health care expenditure, Inequality, Efficiency.

*We would like to thank Mariacristina De Nardi, Eric French, George-Levi Gayle, Martin Gaynor, Fatih Guvenen, Berthold Herrendorf, Chad Jones, Dirk Krueger, Pricila Maziero, Ellen McGrattan, Serdar Ozkan, Juan Pablo Nicolini, Chris Phelan, Jim Poterba, Jose-Victor Rios-Rull, Seth Richards-Shubik, Todd Schoellman, Sam Schulhofer-Wohl, Erick Sager, Fallaw Sowell, Gustavo Ventura, Amir Yaron, Motohiro Yogo and seminar participants at ITAM, Chicago Fed, Minneapolis Fed, OSU, Rice, Wharton, Midwest Macro Meetings in Vanderbilt, the SED meetings in Gent, QSPS Workshop at Utah State University, Minnesota Macro Workshop, St Louis Fed Policy Conference and the Canadian Macro Meetings for comments and many useful conversations. Special thanks to Karen Kopecky and Selo İmrohoroğlu for thoughtful discussions. We thank Erick Sager for his excellent research assistance. Financial support from NSF grant SES-0962432 is gratefully acknowledged.

1 Introduction

The U.S Department of Health & Human Services reports that in 2009 median health expenditures for people with high income was $1700 while it was $781 for people described as poor.[1] Is this gap in spending too high? How would an egalitarian social planner allocate health care resources across income groups? In this paper we provide answers to these questions.

We proceed in three steps. First, we use data from the Medical Expenditure Panel Survey (MEPS) to document how health care spending varies across individuals in different income quartiles.[2] Second, following Hall and Jones (2007) we develop a life cycle model in which the sole impact of health care spending is in terms of increased survival. We assume that individuals have different life cycle profiles of productivity. We study the extent to which health care expenditures should be different across productivity types. Finally, we calibrate this model and compare the efficient health care spending implied by our model to the MEPS data.

When looking at MEPS data, we focus on individuals that are not experiencing catastrophic health shocks and are employed during MEPS interviews.[3] We document three facts. One, individuals in the lowest income quartile receive about 21 percent of all health care spending when young (25 to 35 years old). Their share of total health spending rises to about 25 percent when they are old (70 to 80 years old). Two, individuals in the highest income quartile receive about 28.5 percent of all health care spending. Their share of total health spending falls to about 24 percent when they are old. Third, the ratio of spending on top to bottom income quartiles is between 1.2 and 1.4 until age 60. It falls to about 1 after age 65.

To provide an answer to the questions posed earlier, we develop a dynamic life cycle model in which health care spending affects survival. We assume that individuals have different life cycle profiles of productivity. We model these profiles as deterministic conditional on an initial productivity level and assume that they follow the standard hump-shaped pattern. Given our interest on aggregate health care spending across income (productivity) groups we abstract from individuals health shocks. Within this environment we characterize the

[1] Source: Center for Financing, Access and Cost Trends, Agency for Healthcare Research and Quality: Medical Expenditure Panel Survey, 2009. Table 1.1. Poor is defined as individuals with income at or lower than poverty line. High income is defined as individuals with more than 4 times the poverty line.

[2] The Medical Expenditure Panel Survey (MEPS) is a survey administered by the U.S. department of Health & Human Services. This data set collects information on total health expenditures at individual level (these expenditures include those directly by individuals, insurance companies and government programs).

[3] As will be explained later, this is done to limit the feedback effect that very poor health has on income. Also, since we use wage income as a proxy for individual productivity, we focus only on employed individuals.

efficient allocation of health care resources: that is, the allocation chosen by an ex ante egalitarian social planner.

We show that the efficient allocation features two robust properties. First, the planner devotes more health care resources towards individuals with higher productivities before retirement. The idea is simple; the planner allocates higher levels of health care (hence granting a higher survival rate) to the highly productive since the opportunity cost of an hour of work from those types is higher to society. Thus, there is a pure (productive) efficiency reason for having different levels of health care spending across the productivity distribution. Second, under the efficient allocation, retired workers receive the same amount of health care independent of their productivity. Since all workers contribute the same output at retirement (namely each worker contributes zero), there is no efficiency gain from having any difference in health care expenditures between them. In this sense, the outcome of this model is similar to the uniform provision of health care post retirement as in Medicare.

The final step is to quantify how much inequality in health care expenditure the efficient allocation generates and compare it to what we observe in MEPS data. To do so we require two steps. We first estimate age specific parameters for the *survival function* (the map between health spending and survival rate at each age) using health spending levels from MEPS and age specific survival rates from the Center for Disease Control and Prevention. Next we go on to estimate the distribution and profile of productivities over the life-cycle using wage income data in the Panel Study of Income Dynamics (PSID) and MEPS. Preference parameters are chosen by matching average hours worked, the capital to output ratio and the *Value of a Statistical Life*.[4]

Quantitatively the model delivers four key findings. First, we find that the efficient level of spending is quite close to the aggregate data in terms of health spending as a share of GDP. However, the age profile of health spending is steeper than what is seen in MEPS data. Spending on younger ages in the MEPS data is higher than the (ex ante) efficient level while the opposite is true for older ages.[5] Second, we find that under the efficient allocation, 22 percent of all health care spending on young individuals (between ages 25 to 35) goes to bottom income quartile. This share rises to 25 percent for individuals at or above retirement age (65 years old in our model). Third, 30 percent of all health care spending on young individuals (between ages 25 to 35) goes to the top income quartile. This share declines to 25 percent among retirees in the model. Fourth, the ratio of spending on top to bottom income quartiles during the working life ranges between 1.34 (for ages 25 to 35) and 1.0 (at

[4]Following Hall and Jones (2007) we use the Department of Transportation Estimates for the Value of Statistical Life as the target for our calibration.

[5]These are consistent with the findings of Hall and Jones (2007) at the aggregate level.

ages 65 and above).

The mechanism which generates inequality in health spending across income groups in the model is very intuitive and simple. Yet, we find this simple mechanism can generate a pattern of inequality in health spending that is remarkably similar to what we observe in MEPS data, both qualitatively and quantitatively. The only point at which the MEPS data demonstrates more inequality than the efficient allocation in the model is during ages 55 to 65 years old. For this age group, the share of spending at the bottom wage income quartile in MEPS is about 21.5 percent while efficiency implies a spending share of 24.8 percent. In MEPS data, the ratio of average spending between the top and bottom income quartiles for this age group is 1.27. The ratio implied by the efficient allocation is almost 1. After retirement however (after age 65), there is a sharp drop in inequality in the data in line with what occurs in the efficient allocation.

We also compare data quantities to what we call the "Laissez Faire" allocation in the model, in which each individual splits his/her own income between consumption and health care over the life cycle. This allocation gives rise to drastically higher levels of inequality than what is seen in the data. In this sense, the current allocation of health care spending in MEPS goes a long way in providing the efficient amount of social insurance (relative to each man for himself). There are still differences between the (ex ante) efficient outcome and what we observe in data, however. These are primarily concentrated among individuals near retirement at the lower end of the income distribution.

Our quantitative results are sensitive to the parameter that determines constant flow utility that individuals enjoy if they are alive. We show that increasing this parameter leads to higher and more equal levels of health care spending in the efficient allocation. We report these results in Section 6.1.

In Section 6.2 we explore implications of the model when productivity types are private information and the planner has to use the allocation of consumption and health care spending to induce truthful revelation of productivity types. One implication of this is that the allocation of health care is unequal even after retirement. We find that under this formulation the inequality in health spending at all ages is much lower in the data than what is implied by constrained efficient allocation (even after the retirement).

Related papers

The paper closest to ours is Hall and Jones (2007). As in their paper the principle motivation for health care spending is to reduce mortality risk.[6] They focus on a representative

[6]Like Hall and Jones (2007), we acknowledge that this is somewhat restrictive. Not all health care spending has an effect on survival (e.g., cosmetic procedures). Also, there are a variety of other kinds of

agent formulation to derive normative implications for overall spending on health care over time and compare this with US aggregate data. In our paper, we introduce heterogeneity across agents and pursue the implications of the model across the income distribution and compare this with household-level data in the MEPS. In this regard our paper can be seen as complementary to Hall and Jones (2007) studying the optimality of health expenditures in the cross-section.

Ozkan (2011) also looks at an environment similar to ours. The focus of the paper however, is in developing a positive description of household health care decisions. Health care expenditure patterns in his model are driven by wealth heterogeneity and moral-hazard-like market inefficiencies. Additional positive papers that study how health care affects lifetime decisions are Yogo (2009) focusing on the elderly portfolio composition, De Nardi et al. (2010) looking at the effect of health expenditures on life-cycle asset accumulation. Prados (2012) estimates the effect of health status on earning ability and explores implications of expanding health insurance coverage on individuals' productivity and welfare. Cole et al. (2012) also study the effect of expanding health insurance coverage in a model in which health affects individuals' productivities. However, they focus on the insurance-incentive trade offs arising from individual's private effort in leading a healthy life.

The question of the efficient allocation of health care has also been addressed by Arrow (1973) and Daniels (1985). The former discuss how the approach of Rawls (1971) extended to health care would be problematic: society would be dealing with a "bottomless pit" in which an ever increasing amount of resources are diverted to the most ill individual. The latter moves towards a characterization of a more desirable allocation of health-care. This is done by relaxing the assumption in Rawls that members of society are well functioning adults and introducing health care as a primary good (using the terminology of Rawls). The theory proposed suggest that a just allocation of health care should feature a minimal levels of health expenditure for the least well off member of society.[7]

Our paper is also related to a large literature that documents inequality in health outcomes and health status and it's relation to income. Refer to Deaton (1999), Deaton and Paxson (1998, 1999), Wagstaff (2002), Skinner and Zhou (2004).

The remainer of the paper is organized as follows: we begin looking at the data on health expenditures in Section 2. In Section 3, we construct a normative model of health expenditures and characterize the (first best) efficient allocation. In Section 4 we calibrate

activities – e.g., non-health spending, environmental factors and life style choices – that also affect survival. We are abstracting from these. For a brief survey of the link between health spending and survival see Hall and Jones (2007) and references therein.

[7]See also Powers and Faden (2006) for a related approach.

the model and in section 5 we compare its quantitative implications with the data. In Section 6 we perform some robustness calculations on our results. Finally, we conclude in Section 7. The Appendix contains additional information on our data sets and includes the proofs of theoretical results.

2 Health Expenditure Facts

In this section we present stylized facts on health expenditures over the life cycle. In particular, we show how health spending varies across different income quartiles over the life cycle. In Section 2.1 we first briefly discuss the data source and sample selection. We present our empirical findings in Section 2.2.

2.1 Data and Sample Selection

Our main data source is the Medical Expenditure Panel Survey (MEPS). This is a rotating panel containing individual level data on hours worked, income and demographic variables, along with extensive information on medical expenditures. The MEPS provides annual data continuously from 1996 to 2009. For each year the MEPS provides data for roughly 30,000 individuals. Refer to Appendix C for additional details on the MEPS data set.

We study individuals between 20 and 90 years of age. Our measure of health expenditure for each individual is the sum of all actual health expenditures incurred either by that individual or provided by other sources on his/her behalf over the course of a year.[8] Our measure of income for each individual is the sum of income from all sources (includes wage, business, unemployment benefits, dividends, interest, pension, Social Security income, etc.)

Since we are interested in the link between productivity and health spending, we only keep observations whose wage income is at least 10 percent of their total income for working age individuals (those 20 to 65 years old). In addition we exclude those who are unemployed during all their survey interviews (there are five interviews in a period of two years). This procedure allows us to only consider observations with strong link between wage income (as a proxy for productivity) and health spending. We put no such restriction for retired individuals. Finally, we remove individuals who self report to be in *poor health* in all their

[8]This does not include "uncompensated care". The MEPS defines expenditures as "payments made for health care services," which excludes the cost of services for which there was no explicit and identifiable payment linked to a specific patient (except for services provided by public hospitals and clinics). For example, the MEPS does not count provider revenues from government appropriations to hospitals and Medicare and Medicaid disproportionate-share hospital (DSH) payments, since they are not payments for specific patients. See Hadley and Holahan (2003) and Miller et al. (2004) for more details.

interview responses.[9] We also exclude observations with extremely high health expenditures. To do this, we first divide the population in 5 year age-groups from age 20 to 90. In each age-group we calculate the 99th percentile of health expenditures. We then drop observations with health expenditures above the 99th percentile of the respective age-group in each year. The reason for the last restriction on the sample is a potential issue with the endogeneity of income. For example, if a high productivity person becomes sick, his/her ability to work is likely compromised. In this case, income is lower than what would be expected from the productivity and simultaneously medical expenditures are high. This effect biases upward the estimate of spending on low income individuals (by lumping in higher productivity types with them) and biases downward the estimate of spending on high income individuals (by classifying the very sick among them as low income).[10] The individuals that are omitted, i.e. those with consistently poor health and those with extremely high health spending, are most prone to this type of bias. Our final sample includes 235,013 separate observations. For more details on how this procedure affects spending by age and income groups refer to Appendix C.[11]

2.2 Facts on Health Care Spending Across Income Groups

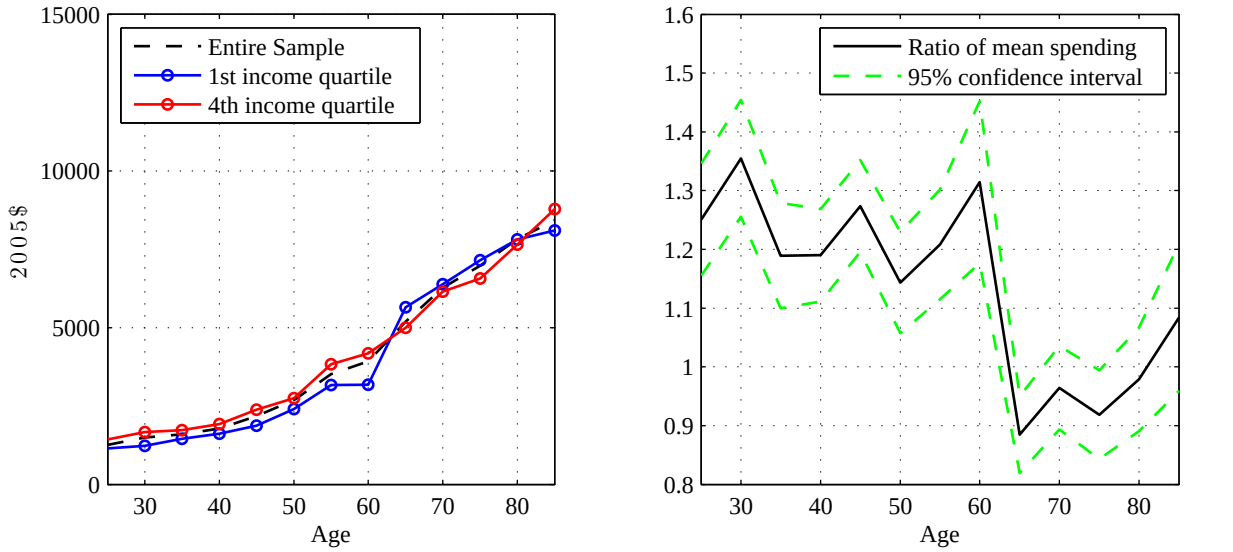
We begin by looking at average spending on health by age. This is the dotted black line in Figure 1(a). In the data, expenditures starts at about $1,100 per person per year for 25 year old individuals and then gradually rises, reaching a peak of about $8,000 per person per year at 80 years of age. Also shown in Figure 1(a) are average expenditures for the top and bottom income quartiles for ages 25 to 85. As can be seen, up to age 65, spending is higher for individuals in the higher income quartile than in the lower income quartile. This can be seen even more clearly in Figure 1(b) where we plot the ratio of average spending between the top and bottom income quartile. Over the course of the working life this ratio is between 1.2 to 1.3. So that the highest income quartile on average spends between 20 to 30 percent more over the entire pre-retirement period.[12] After age 65, we observe a sharp

[9] In a given year an individual is interviewed multiple times. In each interview he is asked to self report his overall health (the exact question asked is: "In general, would you say your health is:"). Self reported health status can take on five possible values: poor, fair, good, very good and excellent.

[10] See Cole et al. (2012) and Prados (2012) for evidence on the effect of health on productivity.

[11] Previous versions of this paper looked at a more restrictive sample focusing on individuals whose self reported health status is within the median range of the self reported health status for the corresponding age group. This approach limits the impact of health on income by limiting by how much health varies across people at the cost of dropping a significant number of observations. Using this sample we found similar results as in the current version of the paper. See Ales et al. (2012) for more details.

[12] Ozkan (2011) and Jung and Tran (2010) document the opposite pattern within the entire MEPS sample. This means that if we do not exclude extreme health expenditures, extreme poor health and unemployed (or zero wage individuals) then we observe higher spending on average at the bottom income quartile. This



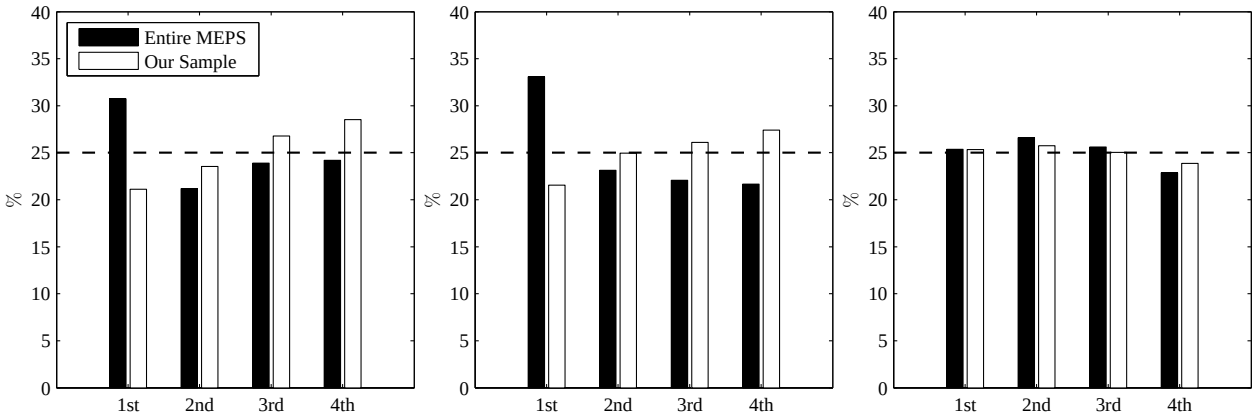
(a) Age profile of health spending.

(b) Ratio of spending of 1^{st} to 4^{th} income quartile.

FIGURE 1: Average Health Spending in MEPS Data over Age.

increase in health spending of the bottom income quartile (this is mostly due to the effect of Medicare spending, as can be seen in Figure 17(b)). This leads to a sharp decline in the ratio of spending as well. After retirement, spending by the bottom income quartile is higher (the ratio drops to 0.9), however over age we observe an increase in inequality, so that by age 85 the highest income quartile spends 10 percent more than the lowest income quartile.

finding is not robust in many dimensions and it is driven by a few extreme health spending observations (which we exclude) and is prone to the income endogeneity issue we outlined earlier. To highlight this further we show in the Appendix D that for the entire MEPS sample, median health spending on the top income quartile is higher than on the bottom income quartile. Also, in the entire sample, the income elasticity of health spending is positive and significant.



(a) Ages 25 to 35. (b) Ages 55 to 65. (c) Ages 70 to 80.

FIGURE 2: Share of Health Spending by Income Quartile and Age Group. Entire vs. Filtered Sample.

Before proceeding, we give a summary on how our sample selection affects the distribution of health spending across income quartiles for various age groups. Figures 2(a), 2(b) and 2(c) show the fraction of all health expenditures that go to each income quartile for three different age groups, for both the entire MEPS sample and our selected sample. The main observation is that the sample selection lowers the share of spending attributed to individuals in the bottom quartile of the income distribution. In addition, for the age group going from 55 to 65 (Figure 2(b)) using our selected sample implies that there is monotonicity of health spending with respect to income (i.e., also for the second and third quartile).

3 A Dynamic Model of Health and Mortality

In this section we present a model in which the sole role played by health care spending is to decrease mortality risk. Our goal is to study the relationship between individual productivity and health expenditures implied by an ex ante efficient allocation with full information.

3.1 Environment

The environment we study is similar to Hall and Jones (2007). We depart from their setup in two basic ways. First, we assume that individuals are heterogeneous in their labor productivities. Second, we allow labor supply to be endogenous. We begin by studying the allocation of a single cohort. In later sections we consider an overlapping generation structure.

The economy is populated by a continuum of finitely lived workers. Workers are heterogeneous with respect to their profile of life-cycle productivity. Productivity types will be

labeled according to $\theta_i \in \Theta = \{\theta_1, ..., \theta_I\}$ of relative size $\pi(\theta_i)$. Assume that $\theta_{i+1} > \theta_i$ for all $i = 1, \ldots, I-1$. Types are drawn at date zero when workers are born and they are permanent, they determine a conditionally non-stochastic life cycle profile of labor productivities. Each workers faces a positive probability of death at each age. We assume that there is a terminal age A so that no worker can survive to age $A+1$. Following Hall and Jones (2007), we assume that an agent's survival rate depends on health spending which we allow to be type and age specific and we denote by $h_a(\theta)$.[13] The probability of surviving to age $a+1$ for an individual of age a with health spending h is given by $P_a(h)$. By assumption $P_A(h) = 0$ for all h. We assume the following:

Assumption 1 *$P_a(h)$, is strictly increasing and strictly concave for all $a < A$. In addition* $\lim_{h\to 0} P_a'(h) = \infty, \quad \forall\ 0 \leq a < A.$

As notation we denote by $N_{a+1}(\theta)$ the fraction of workers of type θ who survive to age $a+1$. Workers of type θ who are alive at age a receive $c_a(\theta)$ units of consumption and work for $l_a(\theta)$ hours. Given an allocation $\{c_a(\theta), l_a(\theta), h_a(\theta), N_a(\theta)\}_{a,\theta}$, lifetime preferences of a worker of type θ are

$$\sum_{a=0}^{A} \beta^a N_a(\theta) u\left(c_a(\theta), l_a(\theta)\right).$$

We assume that the per period utility function is always positive and that the per-period utility function is additively separable between consumption and leisure:

Assumption 2 *Let $u(c,l) = u(c) + v(1-l)$ where $u', -u'', v', -v'' > 0$.*

A worker of type θ who supplies $l_a(\theta)$ hours of work at age a, produces $w_a(\theta) l_a(\theta)$ units of output. We assume that productivity, $w_a(\theta)$, is increasing in θ up to a fixed retirement age a_{ret}.[14] At and following retirement, productivity goes to zero for all types.[15] Resources can be transferred over time, let $R = 1/\beta$ be the rate of return on storage.[16]

Our goal is to study a key efficiency channel that will introduce inequality in health spending: more productive individuals are societally more valuable given the higher output they can generate. We study this channel in a stylized environment. For example we are

[13] We assume $h_a(\theta)$ is the total health spending in terms of consumption goods so that the price of health spending in terms of consumption good is set to one.

[14] In the quantitative section we relax the assumption on monotonicity of productivity for each age.

[15] Here, for simplicity, we assume that the age of retirement, a_{ret}, is fixed across all types. In the quantitative work below, we assume that $a_{ret} = 65$. An interesting extension would be to add endogenous, type specific, retirement ages to the model. See Shourideh and Troshkin (2011) for a version of this with private information, but no health spending.

[16] In the quantitative section the return on storage will be endogenous.

intentionally leaving out many realistic features. To mention a few, in this model only current health care spending affects survival (as opposed to a stock of health model as in Grossman (1972) and Ehrlich and Chuma (1990) or an environment where life style choices affect survival probability); there are no health shocks and finally health care spending only affects survival (it does not impact productivity or the flow utility from consumption and leisure). We acknowledge that these are important features. However, including them in the model increases the complexity of the analysis–both theoretically and empirically–without adding substantial insights. For example, consider a model with agents receiving health shocks in each age. In this case, as long as the distribution of health shocks is identical across types, then the aggregate (per type) amount of health expenditures needed to achieve a certain level of survival will be identical across types. Hence adding this feature to the model will not affect the amount of inequality in the allocation across income groups, the main focus of this paper.

3.2 The Full Information Ex Ante Efficient Allocation

In this section we derive the properties of the allocation of health care resources in our normative benchmark. The approach used in this paper is to maximize the ex-ante equally weighted sum of utility as advocated in Wagstaff (1991).[17] The solution to the utilitarian planner's problem is a Pareto efficient allocation. In particular, this allocation can be viewed as the optimal insurance arrangement against the realization of individual productivities that would be agreed on before individuals are born. The utilitarian equally weighted planning problem is given by

$$\max_{\{c_a(\theta),l_a(\theta),h_a(\theta),N_a(\theta)\}_{\theta,a}} \sum_{\theta\in\Theta} \pi(\theta) \sum_{a=0}^{A} \beta^a N_a(\theta) u\left(c_a(\theta), l_a(\theta)\right), \tag{1}$$

[17] For a general discussion regarding the use of this approach in determining desirable allocations of health care expenditures refer to Powers and Faden (2006). For the purpose of this paper we see two main upsides of this approach. First, we are able to evaluate, from a normative standpoint, the entire distribution of health expenditures. In contrast, the approach described in Daniels (1985) and Powers and Faden (2006) can only effectively evaluate welfare and hence, the amount of health care resources for the worse off individual (or group). Second, the MEPS data set described in Section 2, provides all the necessary data needed to evaluate the distribution of health care expenditures. In contrast, theories as in Powers and Faden (2006) or as in Daniels (1985) would require information on all the *essential dimensions of well being* for the former or *primary goods* for the latter in order to determine which group is the worse off and thus determine the lower bound on health care expenditures. Neither of these is available.

subject to

$$\sum_{\theta\in\Theta}\pi(\theta)\sum_{a=0}^{A}\frac{1}{R^a}N_a(\theta)\left[c_a(\theta)+h_a(\theta)-w_a(\theta)l_a(\theta)\right]\leq 0, \tag{2}$$

$$N_{a+1}(\theta)=P_a(h_a(\theta))N_a(\theta),\quad \forall\ \theta\in\Theta,\quad 0<a\leq A, \tag{3}$$

$$N_0(\theta)=1,\quad \forall\ \theta\in\Theta. \tag{4}$$

In our first proposition we focus on the allocation of health expenditure following retirement. We show that after retirement the ex ante efficient allocation requires that all workers have the same health expenditure. The intuition for this is simple. Following retirement no worker contributes to production. In addition, consumption and leisure are equalized across individuals for each type. This implies that the utility of each worker is the same at every age following retirement. Since the planner values each individual equally, then there is no gain to the planner from having a higher number of a certain type survive to higher ages. On the other hand, the cost of producing health is convex and the return in terms of increasing survival rate is concave. This implies that it is ex ante efficient for everyone to receive the same level of health care and hence, they survive to higher ages at the same rate.

Proposition 1 *Suppose the utility function satisfies Assumptions 1 and 2. Let $h_a^*(\theta)$ be the solution to the planer's problem (1) for age a and type θ. Then $h_a^*(\theta)=h_a^*(\theta')$ for all θ and θ' in Θ, and all $A\geq a\geq a_{ret}-1$.*

Proof. In Appendix A.1. ■

A key step driving the previous result is that following retirement all agents receive the same flow of utility. In Section 6.2 we introduce private information on agent's own productivity. Incentives required to extract this information will cause the agents to be ex-post heterogenous. In particular, highly productive agents will receive a higher level of consumption and health expenditures in post retirement years than those with lower productivites.

We now look at health expenditures prior to retirement. Our next proposition establishes that more productive workers must have higher health expenditures at all pre-retirement ages. To show and provide intuition for this result, it is convenient to introduce a recursive formulation for the planner's problem. The state variables will be given by: the current age of the cohort a; the fraction of type θ workers who survive to age a: $N_a(\theta)$; and the total (net) saving in the economy up to age a: k_a. Define the function $V_a(N_a(\theta_1),N_a(\theta_2),...,N_a(\theta_I);k_a)$

as the solution of the following problem

$$V_a(N_a(\theta_1), N_a(\theta_2), ..., N_a(\theta_I); k_a) = \max \sum_{\theta \in \Theta} \pi(\theta) \sum_{a'=a}^{A} \beta^{a'-a} N_{a'}(\theta) \left[u(c_{a'}(\theta)) + v(1 - l_{a'}(\theta)) \right]$$

$$s.t \qquad \sum_{\theta \in \Theta} \pi(\theta) \sum_{a'=a}^{A} \frac{1}{R^{a'-a}} N_{a'}(\theta) \left[c_{a'}(\theta) + h_{a'}(\theta) - w_{a'}(\theta) l_{a'}(\theta) \right] \leq k_a$$

$$N_{a'+1}(\theta) = P_{a'}(h_{a'}(\theta)) N_{a'}(\theta) \quad \forall\ \theta \in \Theta, \forall\ a \leq A.$$

During pre-retirement ages, higher productivity workers provide more labor effort and produce more output. As we will show in the next proposition, the planner always prefers to have fewer low productivity workers and a higher number of high productivity workers. The reason is that, when there are more high productivity workers, resources are higher and the planner can afford to have everyone work less and enjoy more leisure. The planner can affect the type-composition of the population by providing different amounts of health care (and hence, survival probabilities) to different types. Given that a higher number of more productive workers is desired, it is clear that the planner will provide them with a higher level of health care expenditures leading to higher survival probabilities. The following proposition formalizes this argument.[18]

Proposition 2 *Suppose that the utility function satisfies Assumptions 1 and 2. Let $h_a^*(\theta)$ be the solution to the planner's problem (1) for age a and type θ. Then for all ages $a < a_{ret} - 1$, and all $\theta_{i'} > \theta_i$ it follows that $h_a(\theta_{i'}) > h_a(\theta_i)$.*

Proof. In Appendix A.2. ■

3.3 The Equilibrium Allocation

In this section we briefly describe the steady state version of our environment in an overlapping generations setting. This will be the benchmark environment used in our quantitative analysis (Section 4).

The model we presented above can be interpreted as the steady state of an overlapping generations economy in which a new cohort is born in each period t. Assume that there is no

[18] Previous versions of this paper considered the case where health status appears as a state variable as in Grossman (1972). In this case, the pre-retirement behavior of the efficient allocation is similar to the present: more productive workers receive higher health spending than less productive workers. Upon retirement, the effect is reversed. Less productive workers initially receive a higher level of health spending. This is because at retirement less productive workers are endowed with a lower health stock. In the long run every agent receives the same amount of health expenditures. Details are available upon request.

population growth. Output is generated by a constant returns to scale production function $F(K_t, L_t)$ where K_t is the aggregate capital and L_t is the aggregate amount of efficiency units of labor in period t. Let $c_{t,a}(\theta)$, $l_{t,a}(\theta)$, $h_{t,a}(\theta)$ and $N_{t,a}(\theta)$ be consumption, hours worked, health spending and the fraction of people alive for type θ individuals at age a in cohort born in period t. Each cohort is assigned a cohort planner who takes the return on aggregate capital, R_t, and wage, W_t, as given and maximizes the ex-ante expected welfare of its cohort. The cohort planner solves the following problem:

$$\max \sum_{\theta \in \Theta} \pi(\theta) \sum_{a=0}^{A} N_{t,a}(\theta) \beta^a \left[u\left(c_{t,a}(\theta)\right) + v\left(1 - l_{t,a}(\theta)\right) \right] \tag{5}$$

subject to

$$\sum_{\theta \in \Theta} \pi(\theta) N_{t,a}(\theta) \left[c_{t,a}(\theta) + h_{t,a}(\theta) \right] + k_{t,a+1} \leq R_{t+a} k_{t,a} + W_{t+a} \sum_{\theta \in \Theta} \pi(\theta) N_{t,a}(\theta) w_a(\theta) l_{t,a}(\theta)$$

$k_{t,0} = 0$, together with the law of motion for N in (3) and (4). Cohort planners' allocations must be feasible, i.e.,

$$C_t + H_t + K_{t+1} = F(K_t, L_t) + (1 - \delta) K_t, \quad \forall\, t, \tag{6}$$

$$K_t = \sum_{a=0}^{A} k_{t-a,a}, \quad L_t = \sum_{a=0}^{A} N_{t-a,a}(\theta)\, w_a(\theta)\, l_{t-a,a}(\theta),$$
$$C_t = \sum_{\theta} \pi(\theta) \sum_{a=0}^{A} N_{t-a,a}(\theta)\, c_{t-a,a}(\theta), \quad H_t = \sum_{\theta} \pi(\theta) \sum_{a=0}^{A} N_{t-a,a}(\theta)\, h_{t-a,a}(\theta).$$

Our focus is on steady states. In this case the return on aggregate capital and efficiency units of labor are given by $R = F_K(K, L)$ and $W = F_L(K, L)$. Note that at the steady state, the problem in equation (5) is identical to problem (1) studied in the previous section.

3.4 The Laissez Faire Allocation

The full information, ex ante efficient allocation discussed in the previous section features an extreme level of insurance: each agent receives full insurance against the realization of his own productivity type θ. In this section, we introduce an alternative benchmark allocation. In contrast to the allocation studied above, this benchmark features no redistribution across productivity types. We will call this benchmark the "Laissez Faire" allocation. In this case, individuals choose leisure, consumption and health expenditures constrained by their

intertemporal budget constraints. We assume that each type has access to perfect annuity markets.[19] The maximization problem faced by an individual of type θ is given by:

$$\max_{c_a, l_a, h_a, N_a} \sum_{a=0}^{A} \beta^a N_a u(c_a, l_a) \tag{7}$$

$$s.t. \qquad \sum_{a=0}^{A} \frac{1}{R^a} N_a \left[c_a + h_a - w_a(\theta) l_a\right] \leq 0,$$

together with the law of motion for N in (3) and (4). Using similar steps as for the ex ante efficient allocation, it can be shown that $h_a(\cdot)$ is increasing in θ.

4 Parametrization

Proposition 2 establishes that the ex ante efficient allocation of health expenditures is increasing in productivity. In this section, we study, quantitatively, the magnitude of this dependence.

In the simulations a period is set to one year. Individuals are assumed to enter the economy and start working at age 20 and retire at age 65. Mortality is determined endogenously, however, we impose that individuals live at most 100 years. In order to compute the allocation, we need estimates of: (i) the survival production function, (ii) the life cycle profile of wages for each type, (iii) preference and technology parameters. We describe each step in turn.

4.1 Survival Production Function

Let h_a denote health spending on an individual of age a. Following Hall and Jones (2007), our survival production function is given by:

$$P_a(h_a) = 1 - \frac{1}{x_a} = 1 - \frac{1}{f_a(h_a)}, \quad \forall\ 20 \leq a < 100, \tag{8}$$

where x_a is the inverse of mortality rate at age a. We assume that f_a is given by:

$$x_a = f_a(h_a) = A_a h_a^{\eta_a}, \quad \forall\ 20 \leq a < 100. \tag{9}$$

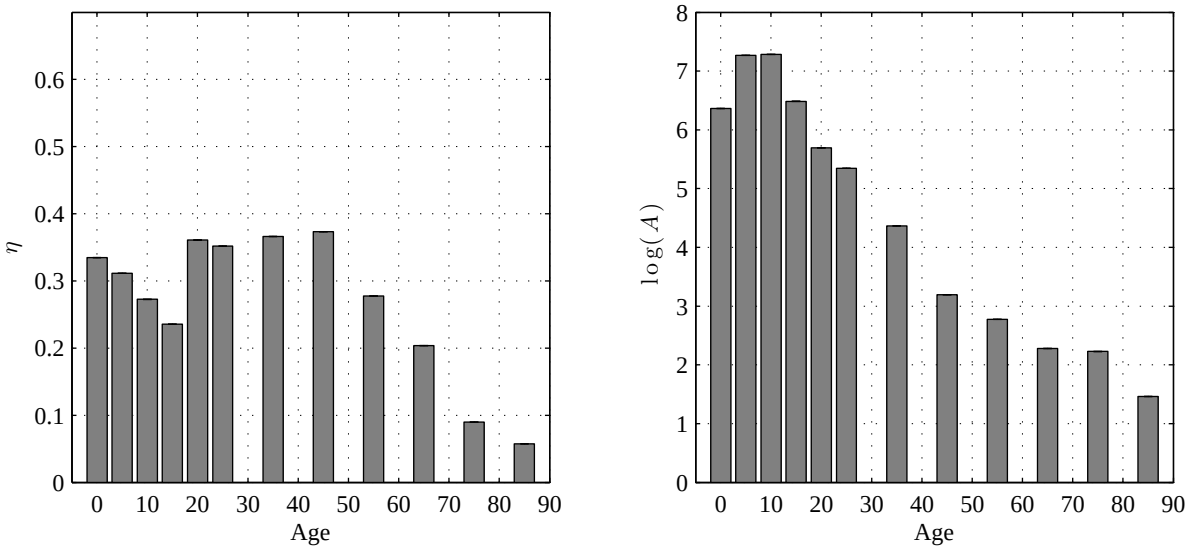
[19] This implies that consumption claims contingent on survival are priced at their actuarially fair value for each type.

As stated earlier $P_{100}(\cdot) = 0$.[20] The estimates of Hall and Jones (2007) for both productivity (A_a) and elasticity (η_a) of health expenditures are based on age specific health spending data which they scaled up so as to match the aggregate values as reported by the BEA.[21] In particular, their data includes nursing home expenditures as a part of health spending. Since MEPS does not include nursing home expenditures, by construction it does not match US aggregates (see Appendix C for further discussion). For this reason, we estimate the parameters A_a and η_a using the MEPS data directly instead of using the Hall and Jones (2007) estimates.[22] To do this, we use variations across observable demographic characteristics in both expenditures and survival rates. In particular we define an observation (group) as a unique combination of year/age/gender/race and census region of residence. Mortality by age for each group is taken from the compressed mortality tables at the CDC. We determine total mortality and subtract mortality due to homicides and suicides. Finally we drop groups that contain less than 20 separate observations in MEPS. This amounts to 4261 separate observations. We estimate (9) using a weighted regression (where weights are given by the number of distinct agents in each separate group) as described in Appendix B. Parameters are estimated for age groups of 5 years up to age 20 and 10 years thereafter. To generate parameter values for every age we then interpolate across age groups. Results are in Figure 3(a) and 3(b). Our estimates of log productivity of the survival production function are similar to those in Hall and Jones (2007). Overall levels are different due to the different levels of aggregate health expenditures used in estimating the production functions. In our case, a smaller level of total spending is required to achieve the same overall mortality rates. An additional difference is the higher estimated elasticities for older individuals. The main reason for this is the absence of nursing home expenditures in the MEPS. Hence, using only the expenditures included in the MEPS to estimate the production functions makes it appear as if health spending is more effective at older ages: a smaller level of total spending is required to achieve the same overall mortality rates.

[20] We have also experimented with a health production function with a baseline mortality so that $P_a(h_a) = 1 - 1/(\overline{x}_a + A_a h_a^{\eta_a})$. This specification provides a positive survival probability without any health spending. Baseline mortality parameters $\overline{x}_a$ were chosen to match survival probabilities in the US at the beginning of the 20th century as in Bell et al. (2005). We found that the inequality results were unaffected by this formulation. Details available upon request.

[21] See, Meara et al. (2004).

[22] The issue of how to treat nursing home expenditures is a difficult one. While some of these expenses are clearly related to health care other parts are not. Moreover, these expenditures are particularly significant for median and low income earners and post retirement individuals. See also Kopecky and Koreshkova (2009).



(a) Elasticity of survival with respect to health. (b) Log productivity of survival production function.

FIGURE 3: Estimation of Survival Production Function.

4.2 The Life Cycle Profiles of Income

Our strategy is to estimate the life cycle profile of labor income using the heterogeneous income profile approach (HIP) as in Guvenen (2007, 2009). The motivation for using this approach is that it enables us to capture the large heterogeneity in lifetime labor income that is already established early in the work life.[23] For an individual i at age a our specification of log labor income is:

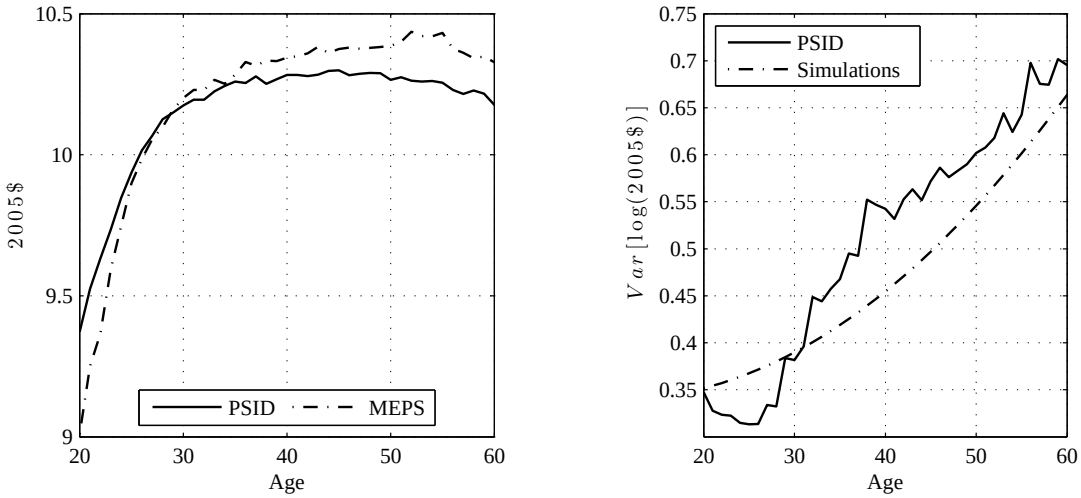
$$y_a^i = g_a + \alpha^i + \beta^i a, \quad \forall\ a = \{20, \ldots, 65\}. \tag{10}$$

Where g_a is the common (across individuals) lifecycle component of labor income. The pair (α^i, β^i) represents the source of uncertainty in our environment. As in Guvenen (2007, 2009) we assume that the parameters are jointly normally distributed with variances σ_α^2, σ_β^2 and covariance $\sigma_{\alpha\beta}$. With respect to Guvenen (2007, 2009) we have two main points of departure. First g_a will be point-wise determined at every age rather than being approximated by a cubic polynomial on age. Second we abstract from any additional uncertainty.[24] Removing additional uncertainty does not affect our result. There are two reasons for this. First, when looking at a limited number of productivity types additional uncertainty averages out. Second, in our benchmark, the planner will provide full insurance against these observable

[23] See for example, Keane and Wolpin (1997).

[24] Guvenen (2007, 2009) allows for an additional autoregressive component affecting labor income.

labor income shocks.



(a) Mean labor income over age.

(b) Variance of labor income over age.

FIGURE 4: Properties of Labor Income.

To estimate the labor income process we will use two data sources. The first data set will be the MEPS described earlier; the second data source is the panel study of income dynamics (PSID).[25] In both data sets we apply the same sampling procedure. For each household we consider both the head and its spouse. Following Guvenen (2009) we restrict the sample to individuals who work at least 520 hours per year and at most 5200. Finally, as described in Section 2, we filter out individuals that have low wages but high income. To make both data sets compatible (PSID data is available from 1967 to 2002 whereas MEPS is from 1996 to 2009) we remove cohort effects as described in Guvenen (2009) and Heathcote et al. (2010).

To determine our productivity types we proceed in several steps. The age dependent function g_a will be estimated by looking at average labor income in MEPS over the course of the life cycle. This profile is shown in Figure 4(a); in this figure it can be seen that the age specific wage profile is similar in both MEPS and PSID. As a point of comparison with Hall and Jones (2007), note that our wage profile will be hump shaped.

We next estimate the joint distribution of (α^i, β^i). To estimate σ_α^2, σ_β^2 and $\sigma_{\alpha\beta}$, we look at the variance of labor income over age in PSID. The profile of life cycle variance of labor income is displayed in Figure 4(b).[26] The age profile of the cross-sectional variance in our

[25] For additional information regarding the PSID refer to Heathcote et al. (2010).

[26] We focus on PSID rather than MEPS for one key reason. Income in MEPS is top-coded at the top 1% (income is top-coded if any of the 14 sources of income considered is above the 1% threshold in it's category). This implies a difference of an order of magnitude between top labor income earners in PSID and MEPS.

specification will be given by:

$$\text{Var}(y_a) = \sigma_a^2 + 2\sigma_{\alpha\beta}a + \sigma_b^2 a^2.$$

From the above, the values σ_α^2, σ_β^2 and $\sigma_{\alpha\beta}$ are estimated regressing the variance of income over age and age squared. We find the following values: $\sigma_\alpha^2 = 0.345$, $\sigma_\beta^2 = 0.000134$ and $\sigma_{\alpha\beta} = 0.0021$. The value of σ_α^2 we estimate is higher than what is reported in Guvenen (2009). This is expected since we do not consider the additional heterogeneity in income due to the autoregressive component. Our estimate of σ_β^2 is similar but slightly lower than what is reported in Guvenen (2009). In addition, the covariance is of the opposite sign than found in Guvenen (2009). This is due to our focus on both the head of the household and the spouse. If, as in Guvenen (2009), we were to restrict to only heads of household we would estimate a similar value for both σ_β^2 and $\sigma_{\alpha\beta}$.

With the estimated values of σ_α^2, σ_β^2 and $\sigma_{\alpha\beta}$ we proceed and draw 10,000 pairs of (α^i, β^i). As a final step we approximate the empirical distribution of (α^i, β^i) by generating a grid of 100 points. We consider a grid of ten values over the distribution of α and β. The grid approximation is done using the procedure described in Tauchen (1986), where the grid on α and β is "rotated" to take into account the covariance between these two values. The grid on α contains 10 values corresponding to the adjusted deciles of the distribution. The grid on β is constructed by 8 deciles of the distribution on β (from the 2nd to the 9th), the lowest and highest values of the grid on β are estimated by looking at the 2nd and 98th percentile of the distribution of β_i. This is done to ensure a correct approximation of the variance of labor income at older ages. The profile of the variance of log labor income in our approximation is displayed in Figure 4(b) where it is compared to the same profile from the PSID.

The final step is to map the 100 different profiles of labor income to the productivity types used in the model. In the model, a type is characterized by a given labor productivity. To estimate profiles of labor productivity we divide the yearly labor income profiles by the average yearly hours worked. We take average hours from MEPS.This step will generate 100 age profiles for labor productivity w_a each associated to one of the 100 types $\theta \in \{\theta_1, \ldots, \theta_{100}\}$. Finally we assume that $w_a = 0$ for $a > 65$.

4.3 Technology and Preference Parameters

The production function is assumed to be Cobb-Douglas: $F(K,L) = AK^{\alpha}L^{1-\alpha}$, where we set $\alpha = 0.36$. The utility function in each period is assumed to be of the form:

$$U(c,l) = b + \frac{c^{1-\gamma}}{1-\gamma} + \psi\frac{(1-l)^{1-\epsilon}}{1-\epsilon}. \tag{11}$$

Here, the parameter b determines the flow value of being alive. The parameters γ and ϵ determine the curvature of the utility function with respect to consumption and leisure, respectively. Finally, ψ denotes the weight of leisure in the utility function relative to consumption.

We begin by calibrating ϵ. Using data from CEX and PSID, Heathcote et al. (2009) estimate a micro-Frisch elasticity of labor equal to 0.38. In our utility specification the Frisch elasticity (η_F) for an individual working l hours is given by

$$\eta_F = \frac{1-l}{l}\frac{1}{\epsilon}. \tag{12}$$

Using the approach of Browning et al. (1999) we calibrate our curvature parameter to match average yearly hours in MEPS.[27] In MEPS, individuals aged 20 to 65 work an average of 39.5 hours per week. We assume that individuals, on average, work for 49 weeks per year. This implies average yearly hours equal to 1935.5. A standard value used for the number of feasible yearly hours is 5200 (normalized to 1 in our model). This implies that, on average, $l = 0.38$. Using (12), this implies that $\epsilon = 4.29$.

A key parameter of interest in this specification is b. This parameter plays a key role in determining the value of life in the model – the higher b is, the higher is the value of a life in the model. As in Hall and Jones (2007), we choose this parameter to match the empirical value of statistical life, *VSL* hereafter. For 34-39 year olds in the year 2000 ($3 million).[28]

In our benchmark calibration we calibrate jointly β, b and ψ by targeting a capital-output ratio equal to 3, a value of statistical life equal to $3 million (at age 37) and average hours worked equal $l = 0.38$. In the first stage of the estimation process we minimize the equally weighted distance between model generated values and targets in the data. In a second stage we set A in the production function so that the output per person in the model matches

[27] Refer to Browning et al. (1999) footnote a in section 3.4.1.

[28] Greenstone et al. (2012) also empirically find support for a VSL of $3 million. For additional discussion on the value of life refer to Hall and Jones (2007), Thaler and Rosen (1976) and Rosen (1988). Refer to Viscusi and Aldy (2003), Ashenfelter (2006) for a quantitative analysis. Technically, b is chosen so that *VSL*=$3,000,000=$[B_1(\theta) + B_2(\theta)]$ where $B_1(\theta)$ is the value, in consumption terms of the discounted value of utility flows from age 37 on of a type θ agent and $B_2(\theta)$ is the value, to the planner of the excess production by a type θ. Details available from the authors on request.

Parameter	Source/Target	Value
γ	Hall and Jones (2007)	2
ϵ	labor supply elasticity of 0.38	4.2936
β	capital-output ratio of 3	0.9689
ψ	average hours of 0.38	6.6332×10^{-6}
b	VSL at age 37 = \$3mln	1.3808×10^{-4}
α	capital share of income from NIPA	0.36
δ	investment-output ratio of 0.235	0.0683
A	set $F(K,L)$ equal to $\frac{GDP}{\text{person 20yrs and older}}$	1.0142

TABLE 1: Baseline Parameters.

GDP per person older than 20 years of age in year 2005. [29] The depreciation rate δ is chosen to match the investment share of output. We find a value of β, b and ψ accordingly) on our results. The baseline parameters used for our simulation are summarized in Table 1.

5 Comparing Model and Data Quantities

In this section we compare our simulation results with the MEPS data. Before looking at the implications of the model for individual health expenditures we look at the performance of the model in terms of aggregate health expenditures.

	Ex Ante Efficient	Data	
		Entire MEPS	Our Sample
In 2005 \$[a]	\$3789	\$3779.5	\$2887.5
As Share of GDP	6.39%	6.38%	4.87%

[a] per person 20 years and older.

TABLE 2: Aggregate Health Expenditures: Ex Ante Efficent vs. Data.

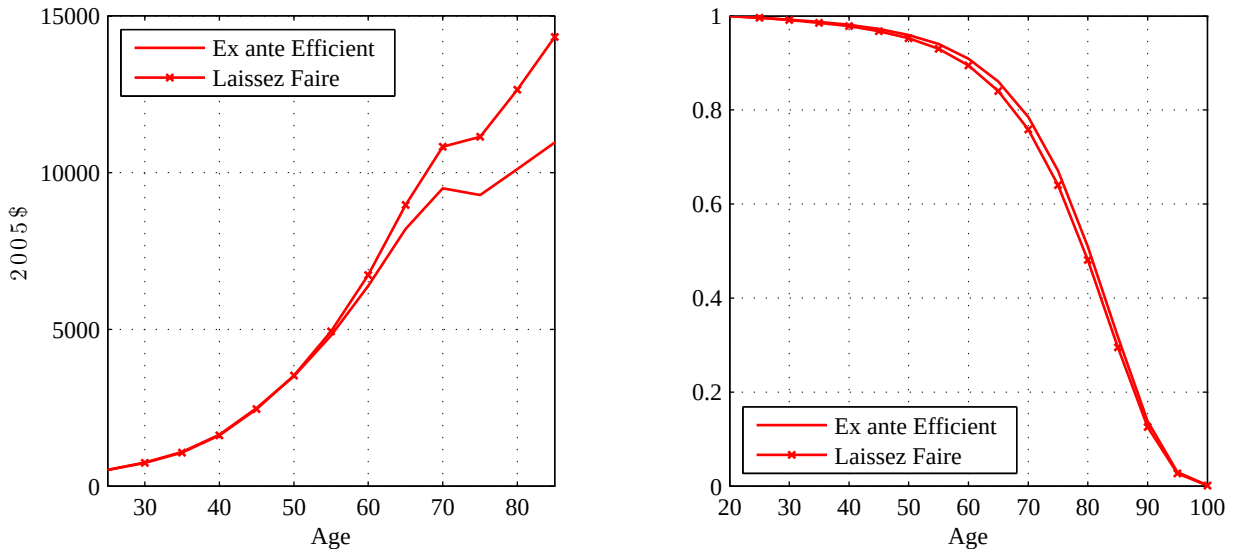
In our model individuals are aged 20 and older. Therefore, we compare all per capita aggregate variables in the model with aggregate variables per person older than 20 years old in the data. As we see in Table 2, health care expenditures per person in the model are very close to those on individuals over 20 years old in the MEPS. However, the model implies higher per person health care spending than what is seen in our selected sample.

[29] It is important that total resources in our model model be consistent with data since the share of expenditure on health care depends on the level of income. In addition, when calibrating the value of life parameter, b, we use the dollar value of a statistical life. It is then important that the arguments in the utility function (consumption and leisure) have the appropriate units and scale. In the case of a model with exogenous output, we could follow the approach of Hall and Jones (2007) and normalize everything by output. However, in our model output is endogenous.

The efficient allocation in the model also implies a health care spending share that is very close to the MEPS data for individuals older than 20. However, these values are much smaller than health care spending as share of GDP in NIPA which is about 11% in 2005. As we pointed out earlier there is a discrepancy between aggregate health spending computed in MEPS and aggregate health expenditure in NIPA. Since we use MEPS to estimate our survival function we think comparing model output to MEPS is the appropriate comparison. Refer to Appendix C for additional details.

5.1 Model Results

We begin by looking at the relationship between age and average health expenditures for all types. These are shown in Figure 5(a).

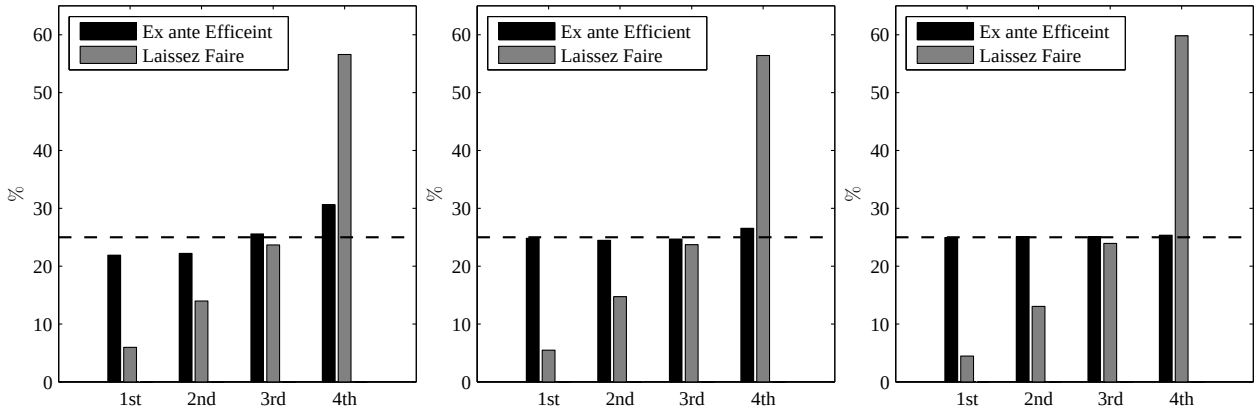


(a) Average Health Spending.

(b) Unconditional survival probability to each age.

FIGURE 5: Comparison Between Allocations. Ex Ante Efficient and Laissez Faire.

As can be seen, average expenditures, for both the Laissez Faire and the ex ante efficient allocations, increase slowly as a function of age, beginning at about $550 per person per year at age 25 and increasing up to about $10,000 per person per year at age 80 for the ex ante efficient allocation. Further, the differences between the average levels of spending for the two allocations are small until retirement age. Health spending in the Laissez Faire allocation grows significantly faster after retirement. Although the average levels of the two allocations are quite similar for most ages, there are significant differences in how these expenditures are distributed across the productivity distribution. This is shown in Figures 6(a), 6(b), and



(a) Ages 25 to 35. (b) Ages 55 to 65. (c) Ages 70 to 80.

FIGURE 6: Share of Health Spending by Income Quartile and Age Group.

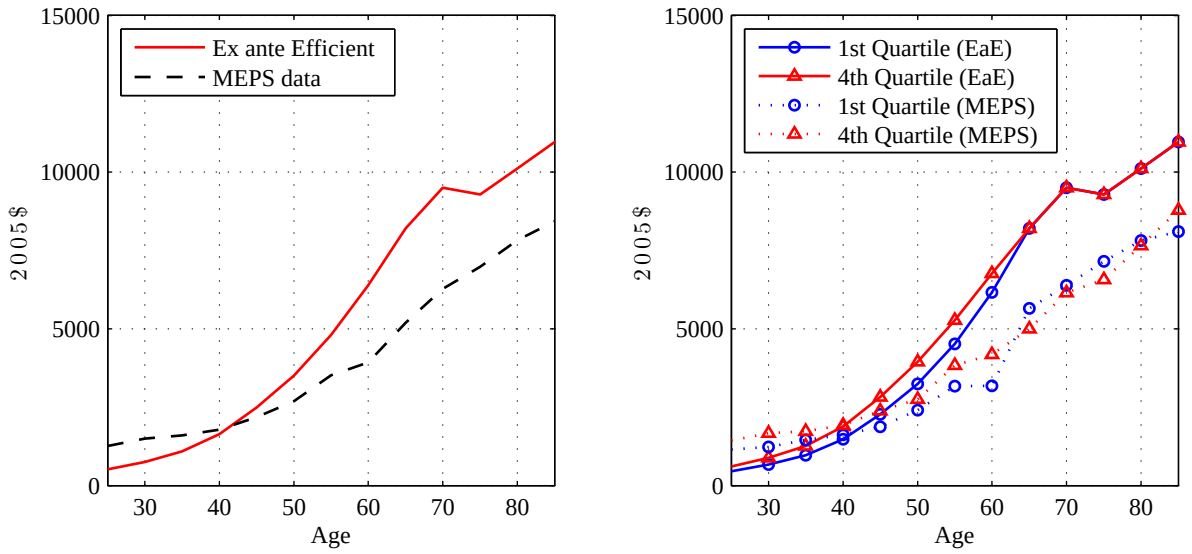
6(c). Here we break down total spending for three age groups across the four productivity quartiles for both the ex ante efficient and Laissez Faire allocations. We observe that the lowest productivity quartile accounts for about 22% of the health spending for individuals between ages 25 to 35 in the ex ante efficient allocation while about 30% of spending is on the highest types. The share going to the lowest types increases monotonically with age while that for the highest types falls. After age 65, the expenditure shares are of course 25% for all of the 4 groups (as expected given the result in Proposition 1). Overall we find that the ex ante efficiency implies a fairly generous social safety net in terms of health expenditures for the poorest income quartile.

Also shown is the distribution of spending across the groups in the Laissez Faire allocation. As can be seen, these are much more unequal. Throughout the lifetime the lowest income quartile receives less than 10% of total health spending while the top income quartile receives more than 50% of total health spending. In both of our benchmark models, expenditures on health are made to increase survival probabilities. These (i.e., average across types of unconditional survival probabilities) are shown, as a function of age, for the two allocations in Figure 5(b). As can be seen, these are quite similar in the two allocations (on average). Note, however, that because there is significant difference in health spending between high income and low income groups under the Laissez Faire allocation, there is also a significant difference in their mortality/survival. In fact this is the reason there is a sharp increase in average spending under the Laissez Faire allocation after retirement (Figure 5(a)). Since lower productivity individuals die at faster rate, high productivity individuals are a higher fraction of the population. This drives up average spending as individuals become older. Note that we don't see this effect under the ex ante efficient allocation since, after

retirement, all individuals receive the same health spending and have the same conditional survival probabilities.

5.2 Model and Data

We now compare the results from the previous section with MEPS data. We begin by comparing overall spending by age in Figure 7(a). As we see, average spending per person of age 25 to 40 in the data is above what is called for under the ex ante efficient allocation. However, the overall spending level in MEPS data grows more slowly than what is called for in the ex ante efficient allocation. Therefore, health spending in the data at old ages is significantly lower than in the efficient allocation.

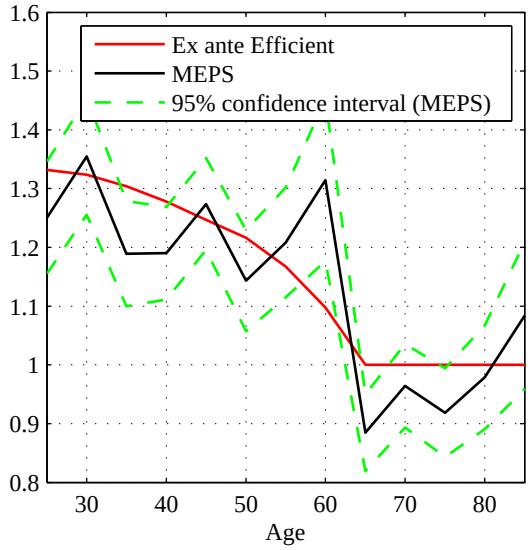


(a) Entire Sample.

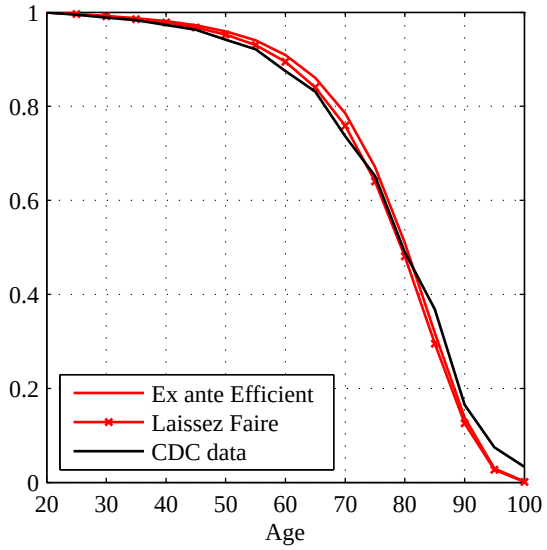
(b) Top and Bottom Quartiles of income.

FIGURE 7: Average Expenditures over Age. Ex Ante Efficient (EaE) and MEPS Data.

Next, we turn to the differences in health spending by different income groups between the model and data in Figure 7(b). We find that inequality in spending by different income groups in MEPS data is very close to what would be called for in the ex ante efficient allocation for all ages. This is best summarized in Figure 8(a) which shows the ratio of average spending of the top to the bottom income quartiles in the ex ante efficient allocation and in the data. This ratio is approximately between 1.2 to 1.3 for all ages up to 50 in the data. We see a similar pattern for the ex ante efficient allocation. However, in the ex ante efficient allocation this ratio steadily decreases towards 1 while in the data it remains above 1.2 in all pre retirement ages. Hence, we see evidence of excess inequality in health



(a) Ratio of average health spending of top to bottom income quartile.



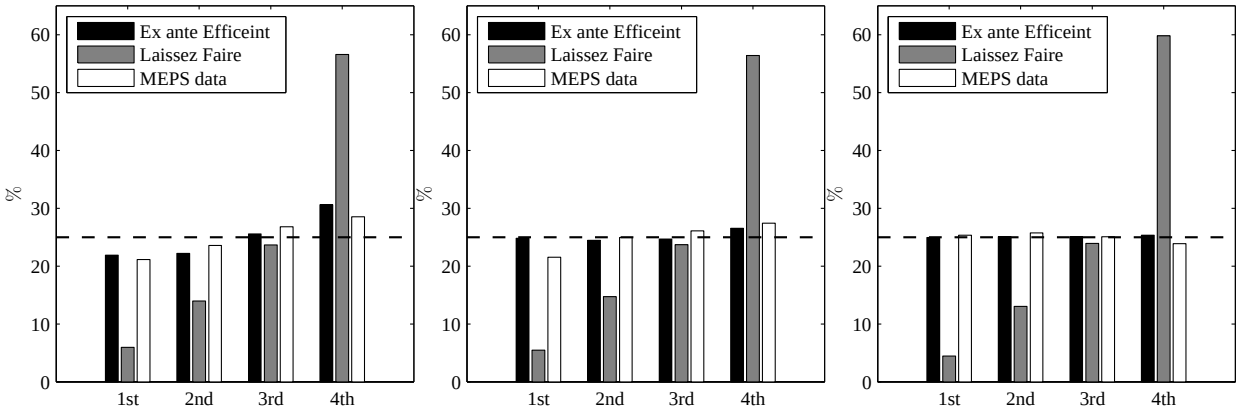
(b) Unconditional survival probabilities to each age.

FIGURE 8: Comparison Between Allocations. Ex Ante Efficient, Laissez Faire and MEPS Data.

expenditures in pre-retirement ages. After retirement, inequality in health expenditures drops significantly as prescribed by the ex ante efficient allocation. For these ages, the ratio of average health spending of the top to the bottom income quartiles fluctuates around 1. We do observe a progressive increase in inequality as individuals age.

Compared to the Laissez Faire allocation, the data exhibits much less inequality in spending. Thus, in this sense, the current allocation in the U.S. is much closer to ex ante efficient than to Laissez Faire. The next three figures: Figure 9(a), 9(b) and 9(c), give more detail about the breakdown in spending across the income distribution for the three allocations for three different age groups. As can be seen, the distribution in the data (white bars) is fairly close to that in the ex ante efficient allocation (black bars) in all the age groups. As a summary we conclude that health care spending in the MEPS data is considerably less unequal than what would hold in the Laissez Faire allocation. In this sense, it appears that the overall level of social insurance, i.e. health care spending transfers from high income to low income, in the U.S. currently is significant. Figure 8(b) provides an additional useful overall summary. As can be seen under the ex ante efficient allocation, unconditional survival probabilities are larger than that seen in the data. This is expected because for most ages spending in the model is higher than in the data.[30] We also see that the Laissez Faire allocation implies lower average survival rates than in the data (and in the ex ante efficient

[30]Since the maximum age in the model 100 years, spending starts to fall significantly after age 90. This is the reason that survival in the data is higher than the model for ages higher than 90.



(a) Ages 25 to 35. (b) Ages 55 to 65. (c) Ages 70 to 80.

FIGURE 9: Share of Health Spending by Income Quartile and Age Group.

allocation). This is true even though (average) overall spending is even more in the Laissez Faire allocation than in the ex ante efficient one. The reason for this is that under the Laissez Faire allocation there is more heterogeneity in mortality across the income distribution. Under this allocation, lower income individuals die at much faster rates and this brings the overall mean survival rate in the model down.[31]

6 Robustness

In this section we depart from the benchmark environment and perform robustness calculations on the results of the previous section. In particular we look at two extensions: (i) sensitivity to the value of life parameter and (ii) the introduction of private information.

6.1 Sensitivity to Value of Life Parameter

The flow utility to an agent from being alive is governed by the parameter b. This parameter can be calibrated by looking at the micro based estimates of the Value of Statistical Life (VSL). The literature studying VSL provides a wide range of values. In their meta-study Viscusi and Aldy (2003) provide a reference value of \$6.7 million (with median at \$7 million) with a standard error of \$5.6 million. In our benchmark we target an average VSL at age 37 equal to \$3 million. In Figure 10, we see that for different productivity types, the VSL ranges from \$2.5 to about \$5 million at the top of the income distribution.

[31] High spending by the rich does not increase their survival rates enough to offset this in the population due to the low elasticity of survival to health spending.

Case	b	β	ψ	A
VSL = \$3 million*	1.38×10^{-4}	0.97	6.63×10^{-6}	1.014
VSL = \$7 million	3.19×10^{-4}	0.964	7.41×10^{-6}	1.016

TABLE 3: Calibrated Parameters for Different VSL Targets (* benchmark calibration).

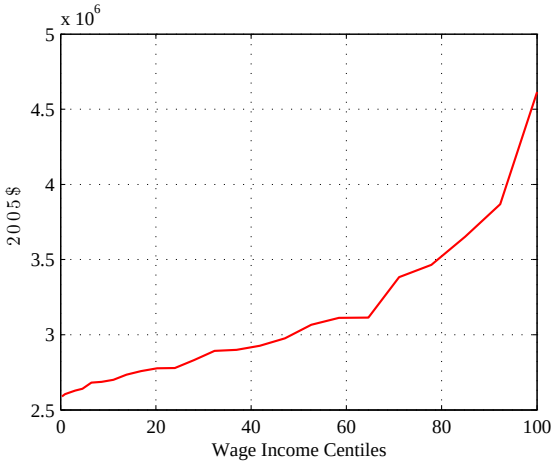
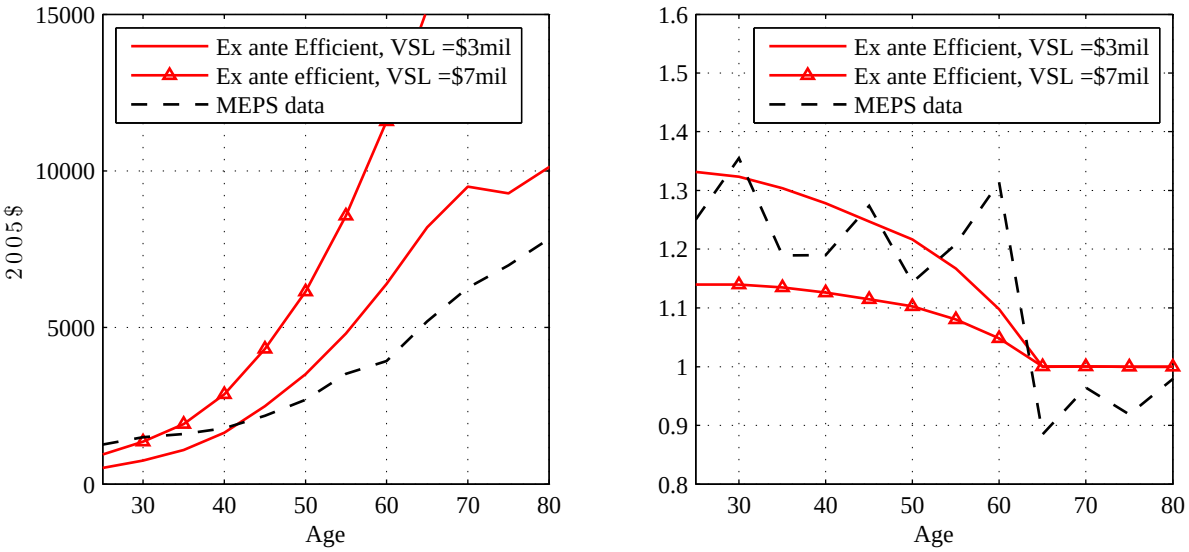


FIGURE 10: VSL over Wage Income at Age 37.

Given what the findings displayed in Figure 10 and the work in Viscusi and Aldy (2003), a potential critique is that the VSL in the benchmark model is too low. To take this into account we now explore an alternative choice for our VSL target. In particular, we target a value for VSL equal to \$7.0 million. In this case, we re-calibrate the values of b, β and ψ. The calibrated parameters for each alternative value of VSL are presented in Table 3. As is expected a higher VSL target implies a higher calibrated value of life parameter b. In Figure 11(a), we report average health spending by age for the two different values of VSL. As expected, a higher base flow value of utility increases the value, to the planner, of keeping each individual alive and this causes spending on health to increase and consumption to fall (not shown). Therefore, the model output with higher VSL target implies higher overall spending.[32] In this sense, when looking at aggregate health expenditures levels, the benchmark choice of a VSL equal to \$3 million seems to be the more appropriate one when looking at aggregate data (as in, Hall and Jones (2007)).

We now look at inequality. A higher value of being alive implies less inequality in health spending across income groups. We can see this in Figure 11(b), which shows the ratio of health care spending between top and bottom income quartiles. Therefore, the model output with a higher VSL target implies less inequality in health care spending across income groups. This is very intuitive. Note that in this model the planner faces a trade

[32]Clearly the opposite would be true with a lower VSL target.



(a) Mean Health Spending by Age.

(b) Ratio of average health spending of top to bottom income quartile.

FIGURE 11: Ex ante Efficient Allocation for Alternative Values of VSL.

off. On one hand, it is benficial to keep everyone alive as long as possible to increase ex ante expected utility. On the other hand, total output (and therefore consumption) is higher if more productive individuals live longer (on average). Although, it is always optimal to have health spending be somewhat unequal, the amount of inequality that is optimal is determined by the trade-off between these two motives. When b is larger, the utility value of keeping everyone alive (irrespective of their productivity) is higher. Therefore, the first effect pushes for less inequality in spending.

Quantitatively, the model prescribes less inequality in health expenditures from what we see in the data but not drastically so (for the age group 40-50 we are still within the bounds of 95% confidence interval). In particular, also in this case, clear inefficiencies emerge in pre-retirement ages. Finally, the amount of inequality in the data is closer to the ex ante efficient allocation than to the Laissez Faire one for both values of VSL that we examined.

6.2 Adding Private Information

So far, we have focused on one reason for making health care unequal – society gets more output as a whole when more productive agents survive longer. A possible criticism of the benchmark environment is that it misses other efficiency reasons for inequality in health spending. In this section we explore a second motive for it to be optimal to introduce inequality in health expenditures: incentives. Suppose that the productivity type of the

agent is his own private information (as in Mirrlees (1971)). In such an environment, the planner can use differences in consumption, leisure and in our case health care expenditures to provide agents incentives to reveal their true productivity types. These differences will be spread across worker lifetimes so that we can expect that following retirement, positive inequality in health expenditures will emerge.

We assume that there are two productivity types $\theta_H > \theta_L$ and that these are the private information of the individual.[33] Further, as in the benchmark environment, we assume that this productivity type is permanent. Preferences are as in (11). In this environment, the planner has to provide incentives for productive individuals to truthfully report their types. Since, as in the previous section, the realization of the type, θ, in the first period determines the entire path of labor productivity, the planner has to induce workers to reveal their types only once - in the first period. For this reason only one incentive constraint is required to ensure truthful revelation of types. The ex ante constrained efficient allocation will be given by the solution of the planner's problem in (1) with the addition of the following incentive compatibility constraint:

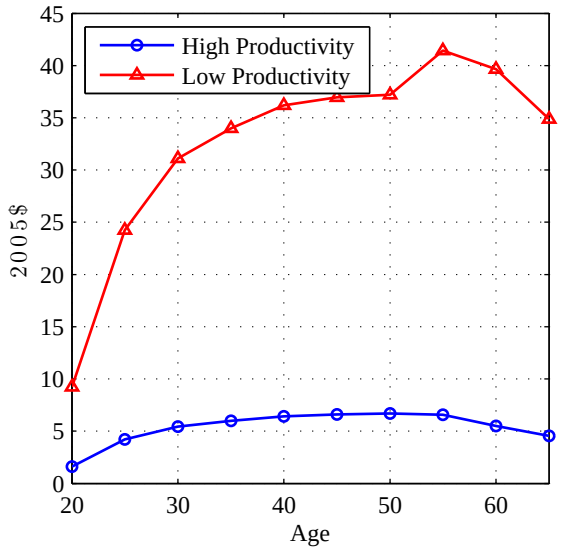
$$\sum_{a=0}^{A} \beta^a N_a(\theta_H) \left[b + \frac{c_a(\theta_H)^{1-\gamma}}{1-\gamma} + \phi \frac{(1-l_a(\theta_H))^{1-\epsilon}}{1-\epsilon} \right] \geq \tag{13}$$
$$\sum_{a=0}^{A} \beta^a N_a(\theta_L) \left[b + \frac{c_a(\theta_L)^{1-\gamma}}{1-\gamma} + \phi \frac{(1-l_a(\theta_L) w_a(\theta_L)/w_a(\theta_H))^{1-\epsilon}}{1-\epsilon} \right].$$

This additional constraint will provide the productive individual with higher lifetime utility, which is provided via higher consumption, leisure, and a higher survival probability through higher health care spending. It can be shown that the constrained efficient allocation will feature $c_a(\theta_H) > c_a(\theta_L)$ for all ages, so that the environment will generate positive inequality in consumption.[34] Characterizing the behavior of health expenditure proves more challenging. Given this, we proceed numerically. For the example that follows we use the parameters of the benchmark case (refer to Table 1). To accommodate the assumption of $I = 2$, we assume $\pi(\theta_H) = \pi(\theta_L) = 0.5$ and set $w_a(\theta_H)$ equal to the average of the top half of wages from the MEPS and $w_a(\theta_L)$ equal to the average of the bottom half of wages from MEPS. Figure 12(a) shows these productivity profiles. When comparing the model with the data we use average health expenditures for the top and bottom halves of the wage income distribution in the MEPS (as opposed to the top and bottom wage income quartiles in the previous sections).

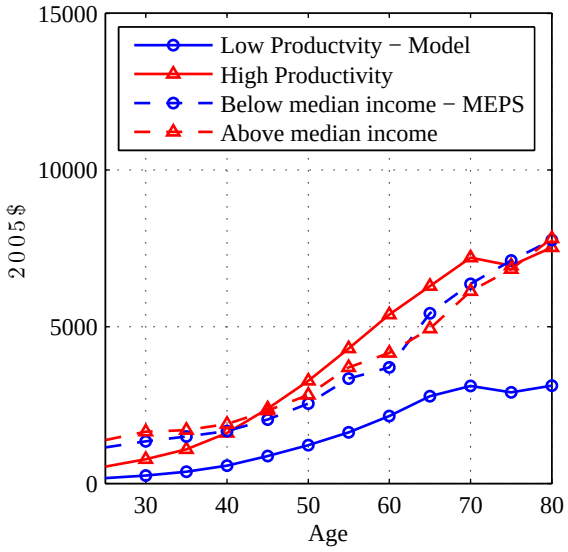
Figure 12(b) shows average expenditures for high and low productivity types and the

[33] Assuming two types will maximize the effect of private information on health inequality.

[34] For additional considerations regarding consumption and leisure in an environment with private information also refer to Ales and Maziero (2010).



(a) Life cycle wage profiles averages above and below the median. MEPS data.



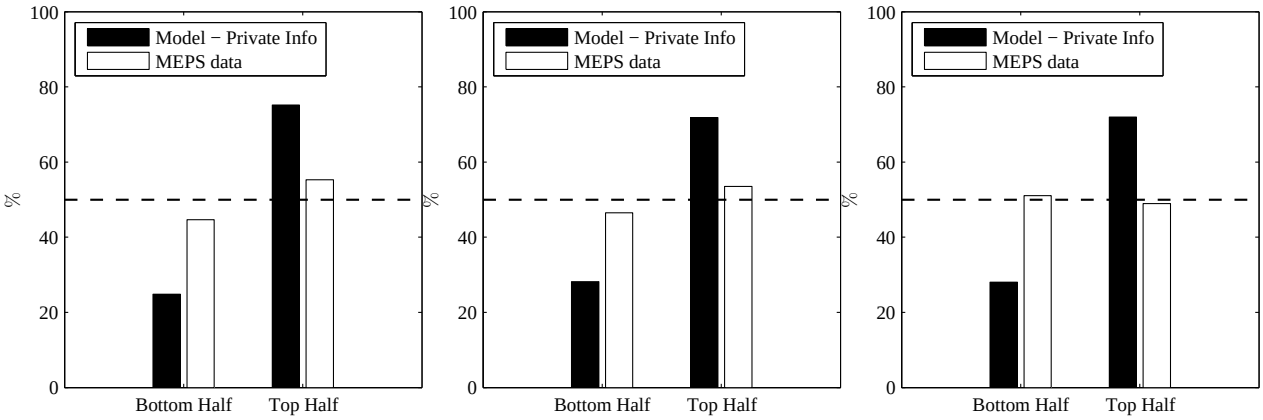
(b) Mean health expenditures averages above and below the median. Constrained efficient allocation (CE) and MEPS data.

FIGURE 12: Benchmark Environment with Private Information.

corresponding quantities from the MEPS. The dashed lines are the MEPS data and the solid lines are taken from the model. In terms of average expenditures, the level of health expenditure in the data is higher than efficient. The is primarily due to an overall drop in output that occurs because of the informational friction. Hence, it is, in part, an artifact of that choice not to recalibrate the model to match output levels in this version. The interesting aspect that emerges in this example is that spending levels are no longer equal after retirement. As can be seen in Figures 13(a), 13(b) and 13(c) the amount of inequality in the constrained efficient allocation is roughly independent of age and is higher than what is seen in the MEPS data. There are two, related, reasons for why this is true. First the planner uses survival rates as an additional instrument, above and beyond consumption and leisure, to 'separate' the types. In addition, the planner has a further motive for high survival rates for high types post retirement – flow utility is higher for the high types.

7 Conclusion

A key policy debate in the U.S. focuses on reforming the health care component of the social safety net. This paper contributes to this policy discussion by characterizing, in a



(a) Ages 25 to 35. (b) Ages 55 to 65. (c) Ages 70 to 80.

FIGURE 13: Share of health spending by age group.

normative model of health care spending, the ex-ante efficient allocation. We are interested in two dimensions of the efficient allocation: the amount of inequality in health spending driven by productivity differentials and the amount of health care resources spent on the least productive segment of the population. When comparing this allocation to what we see in the data we reach two main conclusions: first, we find that the ex ante efficiency implies a fairly generous social safety net: the efficient share of healthcare spending for the bottom quartile of the income distribution is about 22% of total healthcare spending for ages 25 to 35 and it increases gradually up to 25% at retirement. Second, the amount of inequality across the income distribution that is seen in the data is consistent to what would be justified solely on the basis of production efficiency for most of the working life. We observe an inefficiently large level of inequality for pre-retirement individuals.

As usual, these conclusions depend on the particular model that we have chosen to adopt to address these questions, and thus, it is of considerable interest to see how sensitive our results are to this specification. For example, we have not included health as a 'state' – spending on health today only affects survival today, not at future dates (see Yogo (2009) for a specification of the health production function with this feature). Other examples of useful extensions include, but are not limited to: explicitly including illness in the model, allowing productivity to move stochastically over the life cycle and a more thorough examination of the case with private information. We leave these questions for future work.

References

Albouy, V., L. Davezies, and T. Debrand (2009): "Dynamic Estimation of Health Expenditure: A New Approach For Simulating Individual Expenditure," *IRDES institut for research and information in health economics. Working Paper 20*.

Ales, L., R. Hosseini, and L. Jones (2012): "Is There "Too Much" Inequality in Health Spending Across Income Groups?" *NBER Working Paper*, w17937.

Ales, L. and P. Maziero (2010): "Accounting for Private Information," *Working Paper*.

Arrow, K. (1973): "Some Ordinalist-utilitarian Notes on Rawls's Theory of Justice," *Journal of Philosophy*, 70, 245–263.

Ashenfelter, O. (2006): "Measuring the Value of a Statistical Life: Problems and Prospects," *The Economic Journal*, 116, C10–C23.

Bell, F., A. Wade, and S. Goss (2005): "Life Tables for the United States Social Security Area 1900-2100," *SSA Pub*, 120.

Browning, M., L. Hansen, and J. Heckman (1999): "Micro Data and General Equilibrium Models," *Handbook of macroeconomics*, 1, 543–633.

Cole, H. L., S. Kim, and D. Krueger (2012): "Analyzing the Effects of Insuring Health Risks: On the Trade-o between Short Run Insurance Benefits vs. Long Run Incentive Costs," *Working Paper*.

Daniels, N. (1985): *Just Health Care*, Cambridge University Press.

De Nardi, M., E. French, and J. Jones (2010): "Why Do the Elderly Save? the Role of Medical Expenses," *Journal of Political Economy*, 118, 39–75.

Deaton, A. (1999): "Inequalities in Income and Inequalities in Health," *NBER Working Paper*, w7141.

Deaton, A. and C. Paxson (1998): "Aging and Inequality in Income and Health," *American Economic Review*, 248–253.

——— (1999): "Mortality, Education, Income, and Inequality Among American Cohorts," *NBER Chapters*, c10324.

Duetsch, M. (2008): "Out-of-Pocket Health Care Spending Patterns of Older Americans as Measured by the Consumer Expenditure Survey," *Consumer Expenditure Survey Anthology*, 46–51.

Ehrlich, I. and H. Chuma (1990): "A Model of the Demand for Longevity and the Value of Life Extension," *The Journal of Political Economy*, 98, 761–782.

Greenstone, M., S. P. Ryan, and M. Yankovich (2012): "The Value of a Statistical Life: Evidence from Military Retention Incentives and Occupation-Specific Mortality Hazards," *Working Paper*.

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

Guvenen, F. (2007): "Learning Your Earning: Are Labor Income Shocks Really Very Persistent?" *The American Economic Review*, 687–712.

——— (2009): "An Empirical Investigation of Labor Income Processes," *Review of Economic Dynamics*, 12, 58–79.

Hadley, J. and J. Holahan (2003): "How Much Medical Care Do The Uninsured Use, And Who Pays For It?" *Health affairs (Project Hope)*, W3.

Hall, R. and C. Jones (2007): "The Value of Life and the Rise in Health Spending*," *The Quarterly Journal of Economics*, 122, 39–72.

Hartman, M., R. Kornfeld, and A. Catlin (2010): "Health Care Expenditures in the National Health Expenditures Accounts and in Gross Domestic Product: A Reconciliation," *BEA Working Paper*.

Heathcote, J., F. Perri, and G. Violante (2010): "Unequal We Stand: An Empirical Analysis of Economic Inequality in the United States, 1967-2006," *Review of Economic Dynamics*, 13, 15–51.

Heathcote, J., K. Storesletten, and G. Violante (2009): "Consumption and Labor Supply With Partial Insurance: An Analytical Framework," *NBER Working Paper*, w15257.

Jung, J. and C. Tran (2010): "Medical Consumption over the Life Cycle: Facts from a U.S. Medical Expenditure Panel Survey," *Working Paper*.

Keane, M. and K. Wolpin (1997): "The Career Decisions of Young Men," *Journal of political Economy*, 105, 473–522.

Kopecky, K. and T. Koreshkova (2009): "The Impact of Medical and Nursing Home Expenses and Social Insurance Policies on Savings and Inequality," *Working Paper*.

Meara, E., C. White, and D. Cutler (2004): "Trends in Medical Spending by Age, 1963–2000," *Health Affairs*, 23, 176.

Miller, E., J. S. Banthin, and J. F. Moeller (2004): "Covering the Uninsured: Estimates of the Impact on Total Health Expenditures for 2002," Working Paper 04007, Agency for Healthcare Research and Quality.

Mirrlees, J. (1971): "An Exploration in the Theory of Optimum Income Taxation," *Review of Economic Studies*, 38, 175–208.

Ozkan, S. (2011): "Income Differences and Health Care Expenditures over the Life Cycle," *Mimeo, University of Pennsylvania*.

Powers, M. and R. Faden (2006): *Social Justice: the Moral Foundations of Public Health and Health Policy*, Oxford University Press, USA.

Prados, M. J. (2012): "Health And Earnings Inequality Over The Life Cycle: The Redistributive Potential Of Health Policies," *Working Paper*.

Rawls, J. (1971): *A Theory of Justice*, The Belknap Press.

Rosen, S. (1988): "The Value of Changes in Life Expectancy," *Journal of Risk and Uncertainty*, 1, 285–304.

Selden, T., K. Levit, J. Cohen, S. Zuvekas, J. Moeller, D. McKusick, and R. Arnett (2001): "Reconciling Medical Expenditure Estimates From the MEPS and the NHA, 1996," *Health Care Financing Review*, 23, 161–178.

Shourideh, A. and M. Troshkin (2011): "Providing Efficient Incentives to Work: Retirement Ages and the Pension System," *University of Minnesota Working paper*.

Skinner, J. and W. Zhou (2004): "The Measurement and Evolution of Health Inequality: Evidence from the US Medicare Population," *NBER Working Paper*, w10842.

Tauchen, G. (1986): "Finite State Markov-chain Approximations to Univariate and Vector Autoregressions," *Economics letters*, 20, 177–181.

Thaler, R. and S. Rosen (1976): "The Value of Saving a Life: Evidence from the Labor Market," *NBER Chapters*, c3964, 265–302.

Viscusi, W. and J. Aldy (2003): "The Value of a Statistical Life: a Critical Review of Market Estimates Throughout the World," *Journal of risk and uncertainty*, 27, 5–76.

Wagstaff, A. (1991): "QALYs and the Equity-efficiency Trade-off," *Journal of Health Economics*, 10, 21–41.

——— (2002): "Inequality Aversion, Health Inequalities and Health Achievement," *Journal of health economics*, 21, 627–641.

Yogo, M. (2009): "Portfolio Choice in Retirement: Health Risk and the Demand for Annuities, Housing, and Risky Assets," *NBER Working Paper*, w15307.

Appendix

A Proofs of Section 3.2

Before proving the main propositions we show the following

Lemma 1 *The efficient allocation features:*

1. $h_A^*(\theta) = 0$ *for all* $\theta \in \Theta$.
2. $c_a^*(\theta) = c^*$ *for all* $\theta \in \Theta$, *and all* $0 \leq a \leq A$.
3. $l_a^*(\theta) = 0$ *for all* $\theta \in \Theta$, *and all* $a \geq a_{ret}$.
4. $l_a^*(\theta') > l_a^*(\theta)$ *for all* $\theta, \theta' \in \Theta$ *with* $\theta' > \theta$ *and all* $0 \leq a < a_{ret}$.
5. $h_a^*(\theta) > 0$ *for all* $\theta \in \Theta$, *and all* $0 \leq a \leq A - 1$

Proof. Statements *1.* and *2.* follow directly from Assumption 2 and the assumption that no agent can survive to age $A+1$. Statement *3.* follows from the assumption that $w_a(\theta) = 0$ for all $a_{ret} \leq a \leq A$ and for all $\theta \in \Theta$. From the first order conditions for consumption and leisure we have $w_a(\theta)u'(c^*) = v'(1 - l_a^*(\theta))$ for all θ and all $a < a_{ret}$. Which implies *4.*. Finally *5.* follows from Assumption 1. ■

From Lemma 1 we have that:

Corollary 1 *For all* $0 \leq a \leq A$ *and for all* $\theta \in \Theta$, $N_a(\theta) > 0$.

A.1 Proof of Proposition 1

Proof. We first show that the claim must hold for $a = A-1$. Let $h^*_{A-1}(\theta)$ and $h^*_{A-1}(\theta')$ be the solution to the planner problem for age $A-1$ and types θ and θ'. The allocation $h^*_{A-1}(\theta)$ and $h^*_{A-1}(\theta')$ must be the solution to the following maximization problem

$$\max_{\hat{h}_{A-1}(\theta),\hat{h}_{A-1}(\theta')} \pi(\theta')N^*_{A-1}(\theta')P_{A-1}(\hat{h}_{A-1}(\theta')) + \pi(\theta)N^*_{A-1}(\theta)P_{A-1}(\hat{h}_{A-1}(\theta))$$

subject to

$$s.t. \quad \pi(\theta)N^*_{A-1}(\theta)\left(\hat{h}_{A-1}(\theta) - h^*_{A-1}(\theta)\right) + \pi(\theta')N^*_{A-1}(\theta')\left(\hat{h}_{A-1}(\theta') - h^*_{A-1}(\theta')\right) = 0$$

If not it can be replaced by the solution of this problem. The resulting allocation will use the same amount of resources and increases the ex ante welfare (recall that consumption is the same for all types and workers do not work after retirement). It follows immediately from strict concavity of $P_{A-1}(\cdot)$ that $h^*_{A-1}(\theta') = h^*_{A-1}(\theta)$ for θ and θ'.

Next suppose the claim is true for all ages $a+1, a+2, \ldots, A$-1. We show it must be true at age a. Note that the claim is true for all ages above a, therefore

$$\frac{N^*_{a'}(\theta)}{N^*_{a+1}(\theta)} = \frac{N^*_{a'}(\theta')}{N^*_{a+1}(\theta')} \quad \text{for all } a' > a+1$$

$h^*_a(\theta)$ and $h^*_a(\theta')$ must be the solution to the following maximization problem

$$\max_{\hat{h}_a(\theta),\hat{h}_a(\theta')} \pi(\theta')N^*_a(\theta')P_a(\hat{h}_a(\theta')) \sum_{a'=a+1}^{A} \beta^a \left(\frac{N^*_{a'}(\theta')}{N^*_{a+1}(\theta')}\right) + \pi(\theta)N^*_a(\theta)P_a(\hat{h}_a(\theta)) \sum_{a'=a+1}^{A} \beta^a \left(\frac{N^*_{a'}(\theta)}{N^*_{a+1}(\theta)}\right)$$

subject to

$$\begin{aligned}
&\pi(\theta)\frac{1}{R^a}N^*_a(\theta)\hat{h}_a(\theta) + \pi(\theta)N^*_a(\theta)P_a(\hat{h}_a(\theta)) \sum_{a'=a+1}^{A} \frac{1}{R^{a+1}}\left(\frac{N^*_{a'}(\theta)}{N^*_{a+1}(\theta)}h^*_{a+1}(\theta)\right) \\
&+\pi(\theta')\frac{1}{R^a}N^*_a(\theta')\hat{h}_a(\theta') + \pi(\theta')N^*_a(\theta')P_a(\hat{h}_a(\theta')) \sum_{a'=a+1}^{A} \frac{1}{R^{a+1}}\left(\frac{N^*_{a'}(\theta')}{N^*_{a+1}(\theta')}h^*_{a+1}(\theta')\right) \quad = \\
&\pi(\theta)\frac{1}{R^a}N^*_a(\theta)h^*_a(\theta) + \pi(\theta)N^*_a(\theta)P_a(h^*_a(\theta)) \sum_{a'=a+1}^{A} \frac{1}{R^{a+1}}\left(\frac{N^*_{a'}(\theta)}{N^*_{a+1}(\theta)}h^*_{a+1}(\theta)\right) \\
&+\pi(\theta')\frac{1}{R^a}N^*_a(\theta')h^*_a(\theta') + \pi(\theta')N^*_a(\theta')P_a(h^*_a(\theta')) \sum_{a'=a+1}^{A} \frac{1}{R^{a+1}}\left(\frac{N^*_{a'}(\theta')}{N^*_{a+1}(\theta')}h^*_{a+1}(\theta')\right)
\end{aligned}$$

The first order conditions imply $\frac{P'_a(h^*_a(\theta))}{1+P'_a(h^*_a(\theta))} = \frac{P'_a(h^*_a(\theta'))}{1+P'_a(h^*_a(\theta'))}$, and therefore $h^*_a(\theta') = h^*_a(\theta)$. ∎

A.2 Proof of Proposition 2

Before beginning the proof we note that the function V_a defined in Section 3.2 can be found using standard arguments as the solution of the following Bellman equation:

$$
\begin{aligned}
&V_a(N_a(\theta_1), N_a(\theta_2), ..., N_a(\theta_I); k_a) = \\
&\max \sum_{\theta\in\Theta} \pi(\theta) N_a(\theta) \left[u(c_a(\theta)) + v(1 - l_a(\theta))\right] + \beta V_{a+1}(N_{a+1}(\theta_1), N_{a+1}(\theta_2), ..., N_{a+1}(\theta_I); k_{a+1}) \\
&s.t. \quad \sum_{\theta\in\Theta} \pi(\theta) N_a(\theta) \left[c_a(\theta) + h_a(\theta)\right] + k_{a+1} \leq R k_a + \sum_{\theta\in\Theta} \pi(\theta) N_a(\theta) w_a(\theta) l_a(\theta) \\
&N_{a+1}(\theta) = P_a(h_a(\theta)) N_a(\theta) \quad \forall\ \theta \in \Theta, \forall\ a \leq A; \qquad k_0 = 0.
\end{aligned}
$$

We now go back to the proof of Proposition 2.

Proof. We break the proof in two steps. In Step 1, we establish that for $a = a_{ret} - 1$ and for all $\varepsilon > 0$ and $(N_a(\theta_1), N_a(\theta_2), ..., N_a(\theta_I))$,

$$
\begin{aligned}
&V_a\left(N_a(\theta_1), ..., , N_a(\theta_i) - \varepsilon, ..., N_a(\theta_{i'}) + \frac{\pi(\theta_i)}{\pi(\theta_{i'})}\varepsilon, \ldots, N_a(\theta_I); k_a\right) > \\
&V_a\left(N_a(\theta_1), ..., , N_a(\theta_i), ..., N_a(\theta_{i'}), \ldots, N_a(\theta_I); k_a\right) \qquad (14)
\end{aligned}
$$

for all $\theta_{i'} > \theta_i$. In Step 2 we show by induction that (14) holds for all $a < a_{ret}$. In this step we will also complete the proof of the proposition.

Step1
Without loss of generality and in order to make notation easier to follow we present the proof for θ_1 and θ_I. Let $(c_a^*(\theta), l_a^*(\theta), h_a^*(\theta), N_a^*(\theta))$ and k_a^* be the solution to the planner's problem. Consider the following perturbation in the distribution of types alive at age $a_{ret} - 1$.

$$
\begin{aligned}
\tilde{N}_{a_{ret}-1}(\theta_1) &= N^*_{a_{ret}-1}(\theta_1) - \epsilon \\
\tilde{N}_{a_{ret}-1}(\theta_I) &= N^*_{a_{ret}-1}(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon \\
\tilde{N}_{a_{ret}-1}(\theta) &= N^*_{a_{ret}-1}(\theta) \quad \text{for all other } \theta
\end{aligned}
$$

Consider an alternative allocation $(\tilde{c}_{a_{ret}-1}(\theta), \tilde{l}_{a_{ret}-1}(\theta), \tilde{h}_{a_{ret}-1}(\theta))$ such that

$$
\begin{aligned}
\tilde{c}_{a_{ret}-1}(\theta) &= c^*_{a_{ret}-1}(\theta) \\
\tilde{h}_{a_{ret}-1}(\theta) &= h^*_{a_{ret}-1}(\theta) \\
\tilde{l}_{a_{ret}-1}(\theta) &= l^*_{a_{ret}-1}(\theta) \quad \theta = \theta_1, \ldots, \theta_{I-1} \\
\tilde{l}_{a_{ret}-1}(\theta_I) &= \frac{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I) w_{a_{ret}-1}(\theta_I) l^*_{a_{ret}-1}(\theta_I) + \epsilon\pi(\theta_1) w_{a_{ret}-1}(\theta_1) l^*_{a_{ret}-1}(\theta_1)}{\left(N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1)\right) w_{a_{ret}-1}(\theta_I)}
\end{aligned}
$$

A couple of remarks about this alternative allocation. One, note that $\tilde{l}_{a_{ret}-1}(\theta_I)$ has been chosen so that the total output in period $a_{ret} - 1$ is unchanged. Second, we have shown in

proposition 1 that $h^*_{a_{ret}-1}(\theta)$ is the same for all types. Therefore, the fact that there are more θ_I types under the alternative allocations does not affect feasibility.

Note that the utility from consumption under both allocations is the same:

$$\sum_\theta \pi(\theta)\tilde{N}_{a_{ret}-1}(\theta)u(\tilde{c}_{a_{ret}-1}(\theta)) = \sum_\theta \pi(\theta)N^*_{a_{ret}-1}(\theta)u(c^*_{a_{ret}-1}(\theta)).$$

To establish the claim we need to show

$$\sum_\theta \pi(\theta)\tilde{N}_{a_{ret}-1}(\theta)v(1-\tilde{l}_{a_{ret}-1}(\theta)) > \sum_\theta \pi(\theta)N^*_{a_{ret}-1}(\theta)v(1-l^*_{a_{ret}-1}(\theta)).$$

Note that

$$\begin{aligned}
&\tilde{l}_{a_{ret}-1}(\theta_I) = \frac{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I)w_{a_{ret}-1}(\theta_I)l^*_{a_{ret}-1}(\theta_I) + \epsilon\pi(\theta_1)w_{a_{ret}-1}(\theta_1)l^*_{a_{ret}-1}(\theta_1)}{\left(N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1)\right)w_{a_{ret}-1}(\theta_I)} \\
= &\frac{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I)}{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1)}l^*_{a_{ret}-1}(\theta_I) + \frac{\epsilon\pi(\theta_1)}{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1)}\frac{w_{a_{ret}-1}(\theta_1)}{w_{a_{ret}-1}(\theta_I)}l^*_{a_{ret}-1}(\theta_1) \\
< &\frac{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I)}{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1)}l^*_{a_{ret}-1}(\theta_I) + \frac{\epsilon\pi(\theta_1)}{N^*_{a_{ret}-1}(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1)}l^*_{a_{ret}-1}(\theta_1) \equiv \hat{l}_{a_{ret}-1}(\theta_I).
\end{aligned}$$

where we have defined $\hat{l}_{a_{ret}-1}(\theta_I)$ as the weighted average of $l^*_{a_{ret}-1}(\theta_1)$ and $l^*_{a_{ret}-1}(\theta_I)$. Then,

$$\begin{aligned}
&\sum_\theta \pi(\theta)\tilde{N}_{a_{ret}-1}(\theta)v(1-\tilde{l}_{a_{ret}-1}(\theta)) - \sum_\theta \pi(\theta)N^*_{a_{ret}-1}(\theta)v(1-l^*_{a_{ret}-1}(\theta)) = \\
= &(N^*(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1))v(1-\tilde{l}_{a_{ret}-1}(\theta_I)) - \\
&N(\theta_I)\pi(\theta_I)v(1-l^*_{a_{ret}-1}(\theta_I)) - \epsilon\pi(\theta_1)v(1-l^*_{a_{ret}-1}(\theta_I)) \\
> &(N(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1))v(1-\hat{l}_{a_{ret}-1}(\theta_I)) - \\
&N(\theta_I)\pi(\theta_I)v(1-l^*_{a_{ret}-1}(\theta_I)) - \epsilon\pi(\theta_1)v(1-l^*_{a_{ret}-1}(\theta_I)) > 0.
\end{aligned}$$

The last inequality follows from strict concavity of $v(\cdot)$.

Step 2
Suppose for all $\varepsilon > 0$, $(N_{a+1}(\theta_1), N_{a+1}(\theta_2), ..., N_{a+1}(\theta_I))$,

$$\begin{aligned}
&V_{a+1}\left(N_{a+1}(\theta_1) - \varepsilon, ..., N_{a+1}(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\varepsilon; k_{a+1}\right) > \\
&V_{a+1}(N_{a+1}(\theta_1), ..., N_{a+1}(\theta_I); k_{a+1}).
\end{aligned}$$

First we show that $h^*_a(\theta_1) < h^*_a(\theta_I)$. Let $(N^*_{a+1}(\theta_1), N^*_{a+1}(\theta_2), \ldots, N^*_{a+1}(\theta_I))$ and $(h^*_a(\theta_1), \ldots, h^*_a(\theta_I))$ be the solution given $(N_a(\theta_1), N_a(\theta_2), \ldots, N_a(\theta_I))$. Suppose $h^*_a(\theta_1) > h^*_a(\theta_I)$. Consider the

following alternative allocation:

$$\begin{aligned}
\tilde{c}_a(\theta) &= c_a^*(\theta) \\
\tilde{l}_a(\theta) &= l_a^*(\theta) \\
\tilde{N}_{a+1}(\theta_1) &= N_{a+1}^*(\theta_1) - \epsilon \\
\tilde{N}_{a+1}(\theta_I) &= N_{a+1}^*(\theta_I) + \frac{\pi(\theta_i)}{\pi(\theta_I)}\epsilon \\
\tilde{N}_{a+1}(\theta) &= N_{a+1}^*(\theta) \quad \text{for all other } \theta
\end{aligned}$$

$$\tilde{h}_a(\theta) = h_a^*(\theta) \quad \theta = \theta_2, \ldots, \theta_{I-1}.$$

We choose $\tilde{h}_a(\theta_1)$ and $\tilde{h}_a(\theta_I)$ such that the total number of people who survive to the next period is the same

$$\begin{aligned}
P_a(\tilde{h}_a(\theta_1)) &= \frac{\tilde{N}_{a+1}(\theta_1)}{N_a(\theta_1)} = \frac{N_{a+1}^*(\theta_1)}{N_a(\theta_1)} - \frac{\epsilon}{N_a(\theta_1)}, \\
P_a(\tilde{h}_a(\theta_I)) &= \frac{\tilde{N}_{a+1}(\theta_I)}{N_a(\theta_I)} = \frac{N_{a+1}^*(\theta_I)}{N_a(\theta_I)} + \frac{\pi(\theta_1)}{\pi(\theta_I)}\frac{\epsilon}{N_a(\theta_I)}.
\end{aligned}$$

We will show next that this allocation is feasible. In particular, we will show that the cost of delivering these health statuses for type θ_1 and θ_I is lower than that of the original allocation.

Note that because $P_a(\cdot)$ is concave, we have

$$\begin{aligned}
P_a(\tilde{h}_a(\theta_1)) - P_a(h_a^*(\theta_1)) &= -\frac{\epsilon}{N_a(\theta_1)} > P_a'(\tilde{h}_a(\theta_1))(\tilde{h}_a(\theta_1) - h_a^*(\theta_1)) \\
P_a(\tilde{h}_a(\theta_I)) - P_a(h_a^*(\theta_I)) &= \frac{\pi(\theta_1)}{\pi(\theta_I)}\frac{\epsilon}{N_a(\theta_I)} > P_a'(\tilde{h}_a(\theta_I))(\tilde{h}_a(\theta_I) - h_a^*(\theta_I))
\end{aligned}$$

and therefore

$$\begin{aligned}
\tilde{h}_a(\theta_1) - h_a^*(\theta_1) &< -\frac{\epsilon}{N_a(\theta_1)P_a'(\tilde{h}_a(\theta_1))} \\
\tilde{h}_a(\theta_I) - h_a^*(\theta_I) &< \frac{\pi(\theta_1)}{\pi(\theta_I)}\frac{\epsilon}{N_a(\theta_I)P_a'(\tilde{h}_a(\theta_I))}.
\end{aligned}$$

Next, we will show that $\sum_\theta \pi(\theta)N_a(\theta)\tilde{h}_a(\theta) - \sum_\theta \pi(\theta)N_a(\theta)h_a^*(\theta) < 0$. Note that

$$\begin{aligned}
&\sum_\theta \pi(\theta)N_a(\theta)\tilde{h}_a(\theta) - \sum_\theta \pi(\theta)N_a(\theta)h_a^*(\theta) \\
= \;& N_a(\theta_1)\pi(\theta_1)\left(\tilde{h}_a(\theta_1) - h_a^*(\theta_1)\right) + N_a(\theta_I)\pi(\theta_I)\left(\tilde{h}_a(\theta_I) - h_a^*(\theta_I)\right) \\
< \;& N_a(\theta_1)\pi(\theta_1)\left(-\frac{\epsilon}{N_a(\theta_1)P_a'(\tilde{h}_a(\theta_1))}\right) + N_a(\theta_I)\pi(\theta_I)\left(\frac{\pi(\theta_1)}{\pi(\theta_I)}\frac{\epsilon}{N_a(\theta_I)P_a'(\tilde{h}_a(\theta_I))}\right) \\
= \;& \epsilon\pi(\theta_1)\left(\frac{1}{P_a'(\tilde{h}_a(\theta_I))} - \frac{1}{P_a'(\tilde{h}_a(\theta_1))}\right) \leq 0
\end{aligned}$$

Since we assumed $h_a^*(\theta_1) > h_a^*(\theta_I)$, we can always find ϵ small enough such that $\tilde{h}_a(\theta_1) > \tilde{h}_a(\theta_I)$. Then, the last inequality follows from strict concavity of $P_a(\cdot)$.

So far we have shown that if $h_a^*(\theta_1) > h_a^*(\theta_I)$, there is an alternative allocation which is feasible, uses less resources and increases the value of the planner's objective, since

$$V_{a+1}(N_{a+1}^*(\theta_1), \ldots, N_{a+1}^*(\theta_I); k_{a+1}^*) < V_{a+1}(N_{a+1}^*(\theta_1) - \epsilon, \ldots, N_{a+1}^*(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon; k_{a+1}^*).$$

This is a contradiction. Therefore, at the optimal allocation we must have $h_a^*(\theta_1) \leq h_a^*(\theta_I)$. Now suppose, $h_a^*(\theta_1) = h_a^*(\theta_I)$. Consider the same allocation as above with ϵ very small.

$$\begin{aligned}
\tilde{h}_a(\theta_1) &= P_a^{-1}\left(P_a(h_a^*(\theta_1)) - \frac{\epsilon}{N_a(\theta_1)}\right) \\
\tilde{h}_a(\theta_I) &= P_a^{-1}\left(P_a(h_a^*(\theta_I)) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\frac{\epsilon}{N_a(\theta_I)}\right)
\end{aligned}$$

Define the function $F(\epsilon)$ as

$$\begin{aligned}
F(\epsilon) &= N_a(\theta_1)\pi(\theta_1)\left(\tilde{h}_a(\theta_1) - h_a^*(\theta_1)\right) + N_a(\theta_I)\pi(\theta_I)\left(\tilde{h}_a(\theta_I)) - h_a^*(\theta_I)\right) \\
&= N_a(\theta_1)\pi(\theta_1)\left[P_a^{-1}\left(P_a(h_a^*(\theta_1)) - \frac{\epsilon}{N_a(\theta_1)}\right) - h_a^*(\theta_1)\right] + \\
&\quad N_a(\theta_I)\pi(\theta_I)\left[P_a^{-1}\left(P_a(h_a^*(\theta_I)) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\frac{\epsilon}{N_a(\theta_I)}\right) - h_a^*(\theta_I)\right]
\end{aligned}$$

$F(\epsilon)$ is the extra resources that the alternative allocation uses (relative to the optimal allocation). Note that

$$F'(0) = -\pi(\theta_1)\frac{1}{P_a'(h_a^*(\theta_1))} + \pi(\theta_I)\frac{1}{P_a'(h_a^*(\theta_I))} = 0.$$

Therefore, if $h_a^*(\theta_1) = h_a^*(\theta_I)$, the perturbation has no first order effect on cost of health. However, we know there is a first order effect on the next period value function (again, since $V_{a+1}(N_{a+1}^*(\theta_1), \ldots, N_{a+1}^*(\theta_I); k_{a+1}^*) < V_{a+1}(N_{a+1}^*(\theta_1) - \epsilon, \ldots, N_{a+1}^*(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon; k_{a+1}^*)$). This is

a contradiction. Hence at the optimum we must have $h_a^*(\theta_1) < h_a^*(\theta_I)$.

So far we have shown that if

$$V_{a+1}\left(N_{a+1}(\theta_1) - \varepsilon, ..., N_{a+1}(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\varepsilon; k_{a+1}\right) > V_{a+1}(N_{a+1}(\theta_1), ..., N_{a+1}(\theta_I); k_{a+1}),$$

then $h_a^*(\theta_1) < h_a^*(\theta_I)$. To complete the proof we need to show that, this also implies

$$V_a\left(N_a(\theta_1) - \varepsilon, ..., N_a(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\varepsilon; k_a\right) > V_a(N_a(\theta_1), ..., N_a(\theta_I); k_a),$$

Consider the following tilde allocation

$$\begin{aligned}
\tilde{N}_a(\theta_1) &= N_a(\theta_1) - \epsilon \\
\tilde{N}_a(\theta_I) &= N_a(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon \\
\tilde{N}_a(\theta) &= N_a(\theta) \quad \text{for all other } \theta \\
\tilde{c}_a(\theta) &= c_a^* \quad \text{for all } \theta \\
\tilde{h}_a(\theta) &= h_a^*(\theta) \quad \theta = \theta_2, \ldots, \theta_{I-1} \\
\tilde{l}_a(\theta) &= l_a^*(\theta) \quad \theta = \theta_1, \ldots, \theta_{I-1}
\end{aligned}$$

We choose $\tilde{l}_a(\theta_I)$ such that the total output in period a is unchanged, i.e.

$$\tilde{l}_a(\theta_I) = \frac{N(\theta_I)\pi(\theta_I)w_a(\theta_I)l_a^*(\theta_I) + \epsilon\pi(\theta_1)w_a(\theta_1)l_a^*(\theta_1)}{(N(\theta_I)\pi(\theta_I) + \epsilon\pi(\theta_1))\, w_a(\theta_I)}.$$

Note that

$$\tilde{N}(\theta_I)\pi(\theta_I)w_a(\theta_I)\tilde{l}_a(\theta_I) + \tilde{N}(\theta_1)\pi(\theta_1)w_a(\theta_1)\tilde{l}_a(\theta_1) = N(\theta_I)\pi(\theta_I)w_a(\theta_I)l_a^*(\theta_I) + N(\theta_1)\pi(\theta_1)w_a(\theta_1)l_a^*(\theta_1).$$

Finally, we choose $\tilde{h}_a(\theta_1)$ and $\tilde{h}_a(\theta_I)$ such that the new allocation is feasible and the total number in the population in period $a+1$ is unchanged. Moreover, defne $\tilde{h}_a(\theta_1)$ and $\tilde{h}_a(\theta_I)$ such that $P_a(\tilde{h}_a(\theta_1))(N_a(\theta_1) - \epsilon) < N_{a+1}^*(\theta_1)$ and $P_a(\tilde{h}_a(\theta_I))(N_a(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon) > N_{a+1}^*(\theta_I)$. Let

$$\begin{aligned}
P_a(\tilde{h}_a(\theta_1)) &= \frac{N_{a+1}^*(\theta_1)}{N_a(\theta_1) - \epsilon} - \delta_1 \\
P_a(\tilde{h}_a(\theta_I)) &= \frac{N_{a+1}^*(\theta_I)}{N_a(\theta_I) + \frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon} + \delta_I
\end{aligned}$$

for some positive δ_1 and δ_I. To make sure the total number of people in period $a+1$ is

unchanged we must choose δ_1 and δ_I such that

$$\pi(\theta_1)(N_a(\theta_1)-\epsilon)\delta_1 = \pi(\theta_I)\left(N_a(\theta_I)+\frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon\right)\delta_I$$

Moreover, note that $\frac{N^*_{a+1}(\theta_1)}{N_a(\theta_1)-\epsilon} > P_a(h^*_a(\theta_1))$ and $\frac{N^*_{a+1}(\theta_I)}{N_a(\theta_I)+\frac{\pi(\theta_1)}{\pi(\theta_I)}\epsilon} < P_a(h^*_a(\theta_I))$. Therefore, we can choose δ_1 and δ_I small enough so that $P_a(\tilde{h}_a(\theta_1)) > P_a(h^*_a(\theta_1))$ and $P_a(\tilde{h}_a(\theta_I)) < P_a(h^*_a(\theta_I))$. Note that, $P_a(\tilde{h}_a(\theta_1))$ and $P_a(\tilde{h}_a(\theta_I))$ have the same average as $P_a(h^*_a(\theta_1))$ and $P_a(h^*_a(\theta_I))$ and are more concentrated. Also, note that $P_a(\cdot)$ is an strictly concave and monotone function. Therefore, $P_a^{-1}(\cdot)$ is a strictly convex function. Since $P_a(\cdot)$ is strictly increasing and concave, $P_a^{-1}(\cdot)$ is a strictly convex function. This function has a lower average over $P_a(\tilde{h}_a(\theta_1))$ and $P_a(\tilde{h}_a(\theta_I))$ than $P_a(h^*_a(\theta_1))$ and $P_a(h^*_a(\theta_I))$, i.e.,

$$\tilde{N}(\theta_I)\pi(\theta_I)\tilde{h}_a(\theta_I) + \tilde{N}(\theta_1)\pi(\theta_1)\tilde{h}_a(\theta_I) < N(\theta_I)\pi(\theta_I)h^*_a(\theta_I) + N(\theta_1)\pi(\theta_1)h^*_a(\theta_1).$$

Therefore, the alternative allocation is feasible. Note that from our assumption we know

$$V_{a+1}(N_a(\theta_1)P_a(\tilde{h}_a(\theta_1)),\ldots,N_a(\theta_I)P_a(\tilde{h}_a(\theta_I));k^*_{a+1}) > V_{a+1}(N^*_{a+1}(\theta_1),\ldots,N^*_{a+1}(\theta_I);k^*_{a+1}).$$

Using an argument similar to what we used for the problem in age $a_{ret}-1$ we can show that

$$\sum_\theta \pi(\theta)\tilde{N}_a(\theta)v(1-\tilde{l}_a(\theta)) < \sum_\theta \pi(\theta)N_a(\theta)v(1-l^*_a(\theta)).$$

Therefore the proof is complete. ■

B Estimation of the Survival Production Function

The structure of the health production function in (9) that we estimate is:

$$x_{a,t,j} = A_a(z_t h_{a,t,j} s_{a,t})^{\eta_a},$$

where $x_{a,t,j}$ is the inverse of non-suicide, non-homicide mortality rate for agents of age a at time t and in group j; $h_{a,t,j}$ are health expenditures, z_t is aggregate productivity at time t and $s_{a,t}$ are other sources that influence mortality different from health expenditures. Our objective is to estimate A_a and η_a, for each age a. The estimation procedure follows Hall and Jones (2007) closely. One key difference are the observations used for the estimation. Hall and Jones (2007) use variation in mortality and expenditures across years using data from 1950 to 2000 (at five year intervals). In our case, we have cross-sectional data over time only for years 1996 to 2009.[35] To introduce additional variation we partition our sample in groups conditional on observable demographic characteristics. These include race, gender

[35] In the estimation we also include data for 1987 taken from the National Medical Expenditure Survey 1987 (NMES87). This is a precursor to the MEPS discussed in this paper and shares the same structure. Additional information on NMES87 is at: http://wonder.cdc.gov/.

and census region of residence. Taking logs of the previous equation we get

$$\log \tilde{x}_{a,t,j} = \log A_a + \eta_a \Big[\log z_t + \log h_{a,t,j} + \log s_{a,t} \Big] + \epsilon_{a,t,j}, \quad \forall a, t, j.$$

To identify A_a and η_a we remove the effects due to z_t and $s_{a,t}$. Let $\log s_{a,t} = g_{s,a}t + \gamma_{a,t}$. We estimate $g_{s,a}$ looking at the average health expenditure growth in our time periods. Substituting in the above we have

$$\log \tilde{x}_{a,t,j} = \log A_a + \eta_a \Big[\log z_t + \log h_{a,t,j} + g_{s,a}t \Big] + \widehat{\epsilon}_{a,t,j}, \quad \forall a, t, j. \tag{15}$$

Where $\widehat{\epsilon}_{a,t,j} = \epsilon_{a,t,j} + \eta_a \gamma_{a,t}$. Differencing over time (15) and taking expectations we have

$$E_t[\Delta \log \tilde{x}_{a,t,j}] = \eta_a E_t \Big[\Delta \log z_t + \Delta \log h_{a,t,j} + g_{s,a} \Big], \quad \forall a, t, \tag{16}$$

where for all a, t, j we have assumed that $E_t[\widehat{\epsilon}_{a,t+1,j} - \widehat{\epsilon}_{a,t,j}] = 0$. The key identifying assumption is that

$$(1-\mu) E_t[\Delta \log \tilde{x}_{a,t,j}] = g_{s,a}\eta_a, \qquad \forall a, t, j.$$

With $\mu = 2/3$. Substituting into (16) and taking expectations over time we have for all a:

$$g_{s,a}\eta_a = (1-\mu)\eta_a E_t \Big[\Delta \log z_t + \Delta \log h_{a,t,j} + g_{s,a} \Big] \rightarrow g_{s,a} = \frac{(1-\mu)}{\mu} E_t \Big[\Delta \log z_t + \Delta \log h_{a,t,j} \Big].$$

Given the value of $g_{s,a}$ we can estimate A_a and η_a.

C The MEPS Data Set

The Medical Expenditure Panel Survey (MEPS) is a large-scale panel survey administered by the U.S. department of Health & Human Services since 1996. The survey is designed to be representative of the non-institutionalized population in the U.S.. The dataset collects information on what health services are used, how frequently they are used and what expenses are incurred in using these services. The MEPS data is one of the main sources for the National Health Expenditure Accounts (NHEA). Our main source is the household component of the data set (MEPS-HC) which is designed with an overlapping structure. Each household is interviewed 5 times over a 2½-year period. Each wave contains roughly 30,000 individuals for a total of around 12,000 families. Data is available from year 1996 to 2009. To make observations comparable across years, all variables in the data set are deflated to 2005 dollars using the recommended price indices.[36] For each individual in the MEPS we are provided with frequency weights designed to make the population comparable with the Current Population Survey. We aggregate the MEPS data using these weights.

In each interview, data is collected on: demographic variables, income variables, hours worked, various measures of self reported health and specific questions on health conditions.

[36] Refer to: http://www.meps.ahrq.gov/mepswebabout_meps/Price_Index.shtml.

Year	Observations	Year	Observations
1996	22,601	2003	34,124
1997	34,550	2004	34,399
1998	21,107	2005	33,957
1999	24,614	2006	34,142
2000	25,090	2007	30,962
2001	33,554	2008	33,058
2002	39,162	2009	36,849

TABLE 4: Observations available per year in MEPS data.

In addition the data set includes detailed information on: health services usage, charges incurred and expenditures (including out of pocket) on these services.[37] This data is complemented with information obtained from the individual's medical providers. In particular the following payment sources are included in the MEPS:[38]

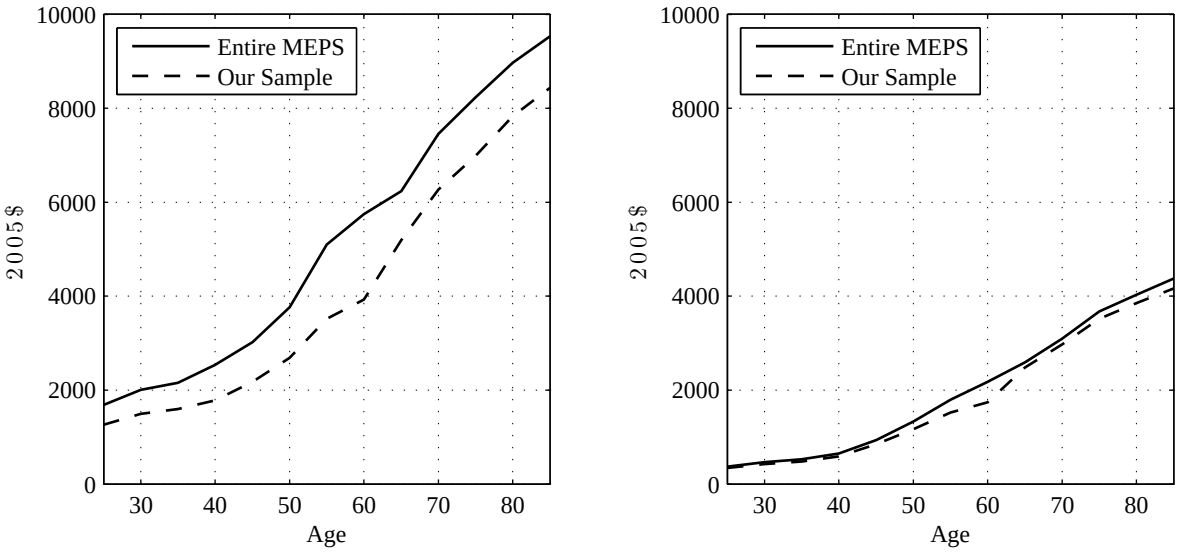
1. Out of pocket by patient or patient's family;
2. Medicare and Medicaid;
3. Private Insurance including automobile, homeowner's, liability insurances;
4. TRICARE and Veterans' administration; other federal, state and local sources;
5. Workers compensation.

MEPS Sample vs. Selected Sample

There are a total of 438,259 observations in MEPS. In Table 4, we summarize the number of raw observations available for each year in our data. From the sample we drop individuals with missing self reported health status and/or missing age. We also limit our analysis to individuals older than 20 years old who report positive total income. We call the resulting sample 'Entire MEPS.' This sample sample has a total of 262,986 observations. As outlined in the text, we perform our analysis on a filtered sample. We drop individuals who report being in poor health in all their interviews as well as individuals with extremely high health expenditure in each age group (those with expenditure higher than the top 1%). We also drop working age individuals (age 20 to 65) whose wage income is less than 10% of their total income, or are reported to be unemployed in all their 5 interviews. This selection drops 35,195 observations, leaving us with a sample size of 227,791 observations which we call 'Our Sample.' Table 5 shows some comparisons between the 'Entire MEPS' and 'Our Sample.' The main difference is that in 'Our Sample' average health spending is lower and average

[37] Information on over the counter medicines is not included in the MEPS definition of expenditures.

[38] Uncompensated care–since is not linked to any payer–is not included as an expenditure in MEPS-HC. Hence an ER visit that is unpaid will be registered as a zero in the data (this is not the case if the visit is paid, for example by MEDICAID). Second, ER visits will also be listed as generating zero expenditures if the patient is subsequently admitted in the hospital.



(a) Average health spending by age.

(b) Median health spending by age.

FIGURE 14: MEPS Data: Comparison Between Entire MEPS Sample and Our Selected Sample.

total income is higher. This is expected since: first, we have excluded some extreme health spending observations; second, we have excluded some low income observations (mostly unemployed individuals).

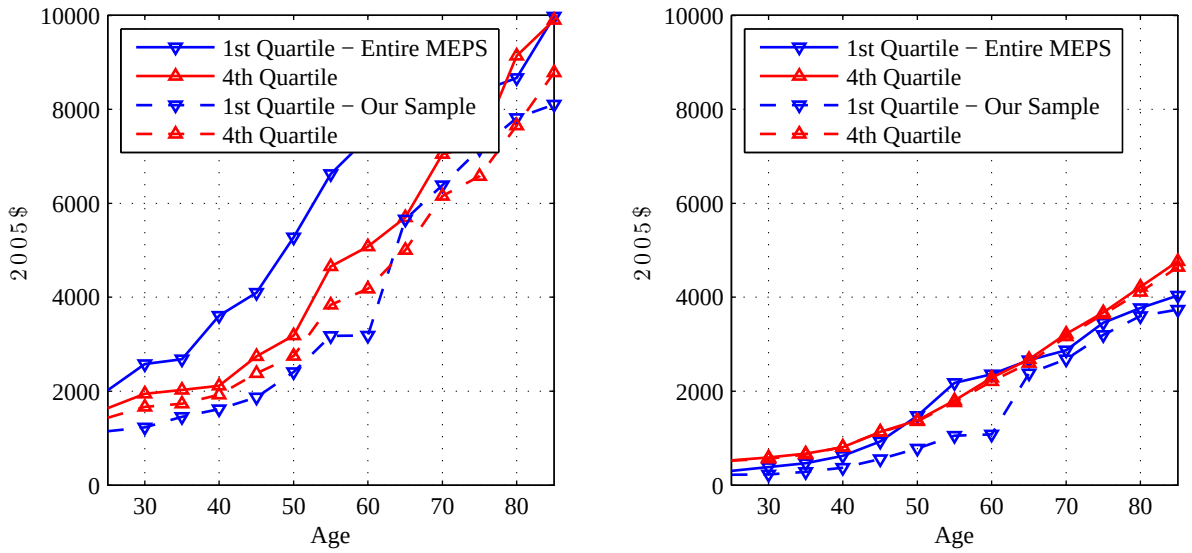
	Entire MEPS	Our Sample	Omitted Observations			
			All	Unemployed	Poor Health	Extreme Health Expenditure
Observations	262,986	227,791	35,195	30,028	3,977	2,665
Avg. Health Spending	$3,676.6	$2,732	$9,790	$6,484.3	$17,929.7	$65,186.2
Avg. Tot. Inc.	$30,762.3	$32,986.4	$14,781	$13,405.1	$15,036.5	$23,674
Avg. age	47	46.6	49	47.6	58.7	47.1
% insured (public)	16.7	13.3	38.7	39.3	57.2	33.8
% insured (private)	67.2	70.2	48.1	46.1	34.5	62.3
% zero spending	17.7	19	9.3	10.4	2.6	0
% in 4th inc. quartile	21.4	20	6	4.4	6.2	14.7
% in 1st inc. quartile	25.7	23	63.1	68	59.4	39.8

TABLE 5: Comparing MEPS with Our Sample.

Next we look at the fraction of insured individuals or fraction with zero health spendings the sample are very similar. However, as we see in columns 4 to 6, the omitted observations are very different form both samples. We next look at mean and median health spending by age in the 'Entire MEPS' and 'Our Sample.' As we see in Figures 14(a) and 14(b) in 'Our Sample' both mean and median is lower for all ages. Also, median health spending is significantly lower than mean spending in both samples, pointing to a very skewed distribution of spending across individuals.

The most significant difference between 'Entire MEPS' sample and 'Our Sample' is the relationship between income and health spending. Figures 15(a) and 15(b) show mean and median health spending for top and bottom income quartiles over age. As we see in Our

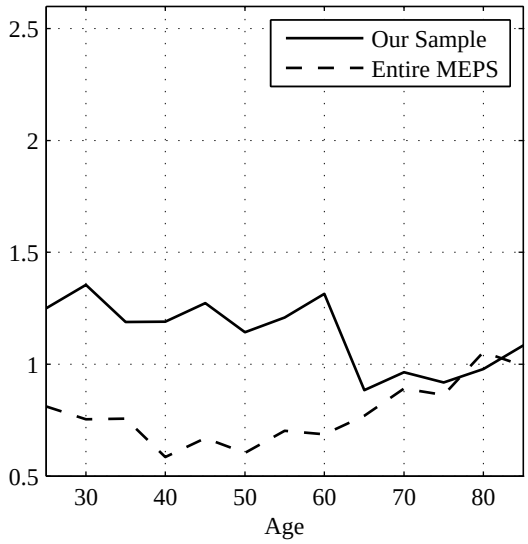
Sample for individuals younger than 65, mean spending in the top income quartile is higher than that in the bottom income quartile. However, if we look at mean spending in the top and bottom income quartiles in the 'Entire MEPS' sample, we see the opposite relationship. However, when looking at medians, spending in the top income quartile is higher than that in the bottom income quartile in both samples (however, the difference is larger in 'Our Sample'). Figures 16(a) and 16(b) illustrate these facts in terms of the ratio of mean/median spending of the top income quartile relative to that of the bottom income quartile.



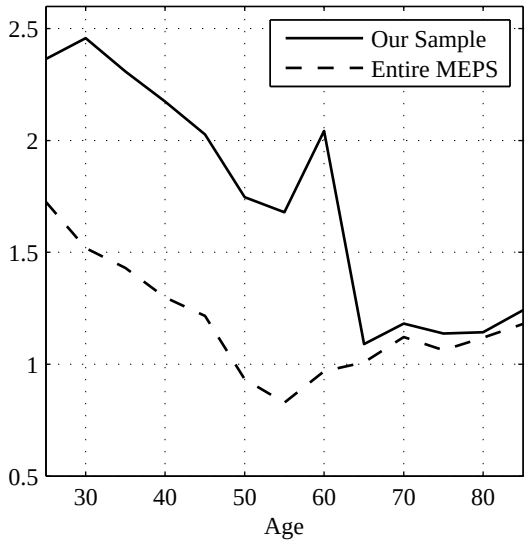
(a) Average health spending by age and income.

(b) Median health spending by age and income.

FIGURE 15: MEPS Data: Comparison Between Entire MEPS Sample and Our Selected Sample.



(a) Ratio of mean health spending: top to bottom income quartile.

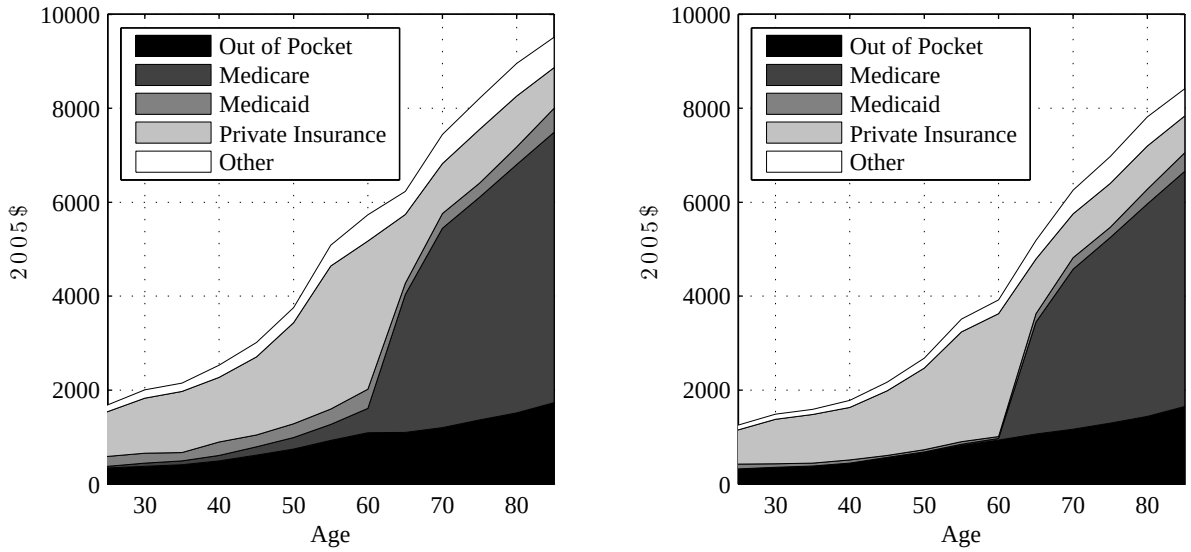


(b) Ratio of median health spending: top to bottom income quartile.

FIGURE 16: MEPS Data: Comparison Between Entire MEPS Sample and Our Selected Sample.

As discussed in the main text, we argue that this pattern in the 'Entire MEPS' sample is due to the effect of income endogeneity: some individuals lose income due to health conditions. Since the panel structure in MEPS is short, we are not able to directly asses the loss of income for an individual for a given health shock. However, we can get a sense of the amount of income lost with two measures: by looking at days lost from work for medical reasons, and by looking at the average years of schooling as a proxy for potential income. We begin by looking at the number of days lost from work due to medical reasons. We look at individuals between the ages of 45 to 50. Individuals removed from our sample for medical reasons (as described above) on average loose 24.7 days of work (this number increases to 56.7 days of work if we condition on at least one day lost for medical reasons). In comparison, individuals in our sample, lose an average of 3.3 days of work (up to 10.5 conditioning on at least one day of work lost). It is clear, in this case, the potential loss of income that might be accrued due to health. We next look at schooling. We look at average years of schooling for individuals in our sample and individuals we have removed from the sample for medical reasons. We look at ages 30 to 65 and focus on the lowest income quartile. We convert the recorded highest degree information encoded in the MEPS into years of schooling. We find that individuals in our sample on average have 10.40 (0.015) years of schooling (standard errors in parenthesis). At the same time, individuals out of our sample have 10.77 (.11) years of schooling. This points to the fact that individuals from the lowest income quartile, that have catastrophic expenditures, based on their schooling might otherwise have earned a higher level of income.

Finally we look at how sources of financing change between 'Our Sample' and 'Entire MEPS' sample. Figures 15(a) and 15(b) shows the average spending by source of funding. The main difference between samples is the reduction in the amount of funding received from Medicare and Medicaid for individuals younger than 65. These amounts are, however, small



(a) Entire MEPS sample.

(b) Our sample.

FIGURE 17: Health Expenditure by Major Source of Funding.

overall. The pattern of spending in both samples is identical otherwise.

MEPS vs. NIPA

We now look at the relationship between the MEPS data set and aggregate U.S. data. Our objective is to compare the total level of health expenditures between the two sources. Our aggregate analogue is Personal Consumption Expenditure of Health Care from NIPA tables. This data is taken from Table 2.4.5u line 168 (deflated to 2005 dollars using it's own deflator provided in table 2.4.4u.) summed with line 27 in Table 3.15.5 (deflated with the deflator in 3.15.4).

MEPS data is not designed to be directly comparable with aggregate measures. The expenditures of nursing homes are not included in MEPS (but they are included in NIPA). Also, expenditures on prescription drugs are included in MEPS but they are not included in Personal Consumption Expenditure of Health Care in NIPA (line 168 in Table 2.4.5u).

To alleviate these discrepancies we make the following adjustment to data in NIPA. First, we subtract expenditures on nursing homes (line 183 in NIPA Table 2.4.5u). Second, we add expenditures on prescription drugs (line 121 in NIPA Table 2.4.5u). Figure 18 shows the aggregate health expenditures as a share of GDP. The dotted line is calculated using the 'adjusted' aggregate health spending in NIPA. The solid line is calculated using the MEPS data. The main observation is that MEPS data systematically underestimates the aggregate health spending relative to aggregate data in NIPA. However there is no trend in the discrepancy. So the MEPS captures the trend in aggregate in health expenditures.[39]

[39]For an additional analysis of this discrepancy refer to Selden et al. (2001) and Hartman et al. (2010). Key differences between MEPS and NHEA are: nursing homes; long term care (greater than 45 days); non-community, non-Federal Hospitals and Alternative care; government spending on health care not ad-

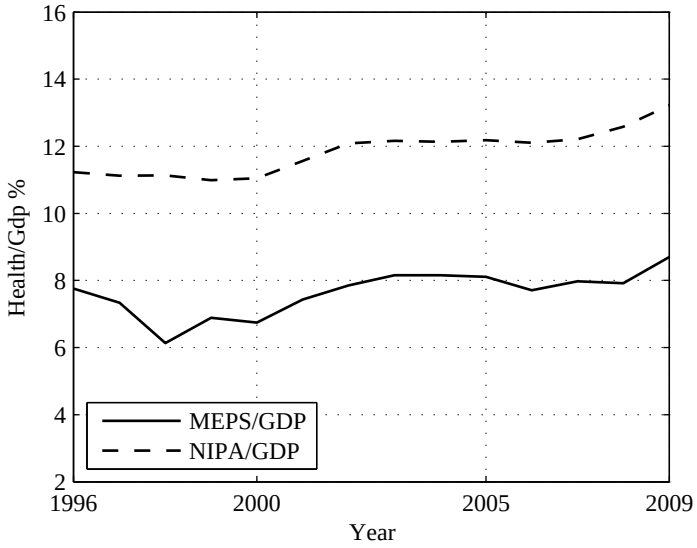


FIGURE 18: Fraction of GDP due to health expenditures as reported by MEPS.

	MEPS (55-64)	CEX (55-64)	MEPS (65-74)	CEX (65-74)
Income before taxes	$65,245	$64,156	$51,640	$45,202
Age of reference person	59.1	59.3	69.3	69.1
Persons in reporting unit	2.04	2.1	1.82	1.9
Out of pocket	$1,995.5	$1,825.64	$2,324.9	$1,823.64
Out of pocket on drugs	$875.9	$712.73	$1,288.5	$956.26

TABLE 6: Comparison MEPS - CEX. Year 2005.

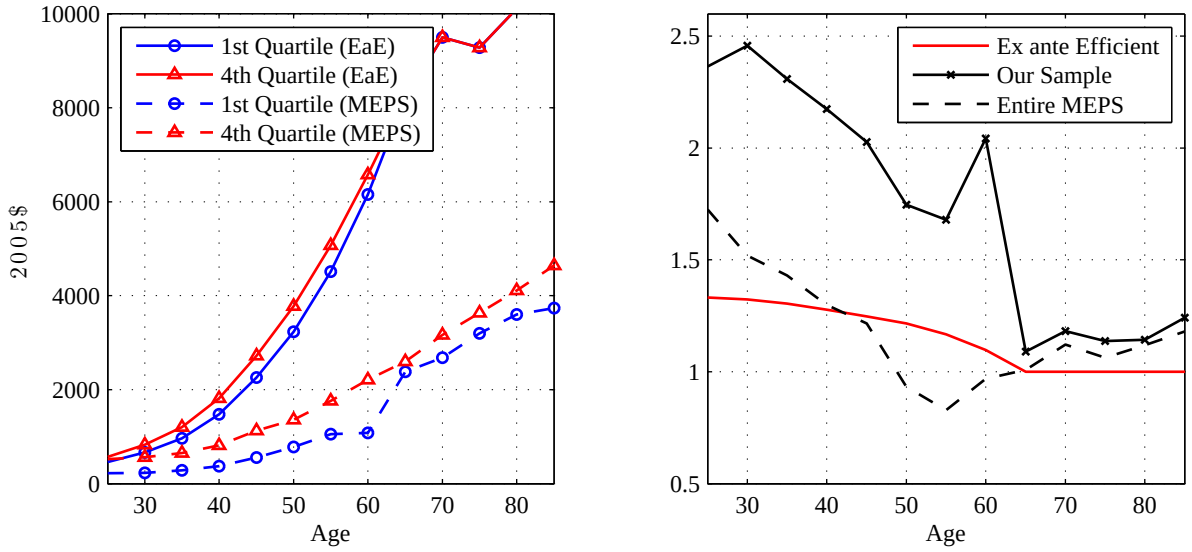
MEPS vs. CEX

As a final comparison for the MEPS we look at another source that reports comparable health expenditure measures. The consumer expenditure survey (CEX) reports out of pocket expenditures incurred by households. Out of pocket expenditures do not include health insurance premiums (see also Duetsch (2008)). In Table 6 we compare the MEPS and the CEX. Data is aggregated at the dwelling unit level (we use the provided family weights for MEPS). In Figure 17(a), we show the average amount of health expenditures financed by major source of funding. This figure makes clear how valuable the data in MEPS is. For example, only looking at out of pocket expenditure on health (which would have been the case if the main data source were to be the Consumption Expenditure Survey) would provide an incomplete picture of total health expenditures for an individual. This picture also highlights how significant Medicare becomes later in life as individuals become eligible.

ministered in hospitals (schools); expenditures for institutionalized population; expenditures of active duty military personnel and long term care in VA hospitals; foreign visitors to the United States.

D Health Expenditures and Income: Other Measures

In this section we look at additional measures of the relationship between health expenditures and productivity. We begin by looking at medians. In the main text, we restrict attention to average health expenditures by age and by income type. However, in the data we observe that the distribution of health expenditure is highly skewed. This is due to two features: first a large fraction of the population reports no expenditures on health. The second is that there is a very small number of observations with very high expenditures. Given the highly skewed distribution of health expenditures we now look at the median of health expenditures at a given age for a given income group. We begin showing medians for the top and bottom quartiles in both model allocations and in the data. Note that in our benchmark model environment there are only slight differences between means and medians. In the data, however, median expenditure levels are significantly lower than model quantities particularly at higher ages. This is shown in Figure 19(a).



(a) Median spending by income: top and bottom quartiles.

(b) Ratio of median health spending of top to bottom income quartile.

FIGURE 19: Median Health Expenditures: Ex ante Efficient and MEPS Data.

We also observe that there is more inequality in the data when measured as the ratio of medians. This is shown in Figure 19(b) where the ratio for the data is around 2 at lower ages. Thus, by this measure, the 'excess' inequality in the data (over the ex ante efficient allocation) is significantly higher. We conjecture that this higher degree of inequality will diminish once we consider an environment where it might be efficient to allocate no health expenditures to individuals as we observe in data.

To determine the relationship between income (y) and health expenditures (h) in the body of the paper we use averages of health expenditures within income subgroups. This procedure has the advantage of being robust to classical measurement error and requires

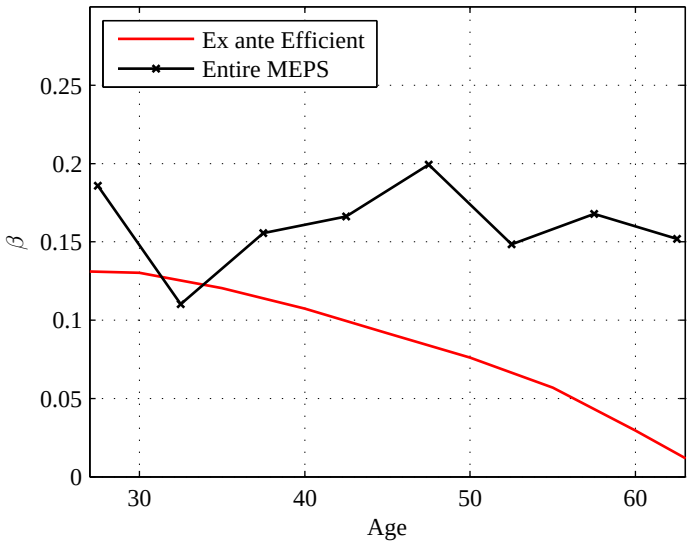


FIGURE 20: Elasticity of health expenditures relative to income. Estimates from equation (17).

little structure to be imposed on the relationship between h and y. The downside of this procedure is that it forces us to look at a very specific aggregate measure of inequality. We now proceed in a different way assuming additional structure on h and maintaining all of our observations. A natural starting point is to assume that h is a log-linear function as follows:[40]

$$\log h_i = \beta_0 + \alpha^{\log} \log y_i + \beta_1^{\log} age_i + \beta_2^{\log} \frac{age_i^2}{100} + \beta_3^{\log} h_i^{rth} + \varepsilon_i. \tag{17}$$

where h_i are health expenditures for individual i, y_i is income of individual i and h_i^{rth} is the annualized level of self reported health and ε_i is classical measurement error. In addition to alleviate the left censoring due to 0 health expenditures lower bound, following Albouy et al. (2009) we also look at a two stage Tobit model:

$$h_i^{data} = \max\{0, h_i\}; \quad h_i = \alpha y_i + \beta X_i + \varepsilon_i. \tag{18}$$

where h^{data} are actual health expenditures observed in the data, h is a latent variable that depends linearly on the covariates and X_i is a vector of observable characteristics. In Table 7, we report our estimates.

Estimation Type	α	Proxy for Productivity
Robust Regression	0.209 (.005)	Income
Tobit	0.189 (.0002)	Income

TABLE 7: Estimation results from equation (17) and (18) on the entire MEPS sample.

For each of the estimates the relationship between productivity and health expenditures

[40] A log-linear specification minimizes difficulties in estimation due to the large number of individuals reporting zero health expenditures and due to the highly skewed distribution of health expenditures.

is positive. This results strengthens the argument in favor of filtering the data as described in the body of the paper. Since only by doing so do we recover a positive relationship between productivity and health expenditures. Summarizing the estimates in Table 7, we observe that a 1% increase in productivity a is associated with approximately a 0.2% increase in spending on health. In Figure 20, we display the regression coefficients from estimating equation (17) for each age group. Both data and model coefficients are displayed. Comparing estimates from the ex ante efficient allocation and the Entire MEPS sample we reach a similar conclusion as in the body of the paper: in the data there is excess sensitivity of health expenditures relative to income than what we observe in the ex ante efficient allocation. However, this excess sensitivity is not large (especially compared to the same estimates for the Laissez Faire allocation which are substantially larger, these estimates are not shown in the figure) and is the largest in pre-retirement ages.