Optimal Healthcare Spending with Redistributive Financing

May 6, 2013

Mark Shepard

Department of Economics, Harvard University

Katherine Baicker

Harvard School of Public Health, and NBER

Jonathan Skinner

Department of Economics and Geisel School of Medicine, Dartmouth College, and NBER

PRELIMINARY – NOT FOR QUOTATION OR CITATION

We are grateful to Amy Finkelstein and Jeffrey Clemens for helpful suggestions, to IPUMS for the use of the NHIS data, and to the National Institute on Aging (PO1 AG19783) for financial support. Shepard gratefully acknowledges funding from the National Institute on Aging, through Grant No. T32-AG000186 to the National Bureau of Economic Research; and from the National Science Foundation Graduate Research Fellowship.

1 Introduction

Health care spending in the United States as a percentage of GDP is higher than in any other country and is expected to grow to nearly one-quarter of GDP by 2037 (CBO 2012b) and more than one-third of GDP by 2063 (CBO 2007).[1] Despite a recent slowdown in overall healthcare spending growth, total expenditures (including out-of-pocket expenditures) for insured households still increased by an average of 6.9 percent between 2011 and 2012, to $20,720 per insured household (Millman, 2012). The most worrisome manifestation of this trend is a dramatic rise in public health care spending on programs like Medicare, Medicaid, and new insurance subsidies that threatens to swamp state and federal government budgets. While federal health care spending is already 5.5 percent of GDP (or one-quarter of the federal budget) today, the CBO projects spending to rise to 12 percent of GDP over the next 50 years. By 2063 federal spending on entitlement programs is predicted to be as large as the entire federal budget today.

Widespread concern about these trends has led policymakers to seek reforms to slow the growth of health spending, if only to avoid dramatic hikes in future tax rates (Baicker and Skinner 2011; Leonhardt, 2009). But recent economic research suggests a very different interpretation of rising spending. Cutler (2004) and Nordhaus (2002) note that better health over the 20th century has yielded enormous benefits, while Murphy and Topel (2006) estimate health gains have amounted to half of total GDP . Hall and Jones (2007) present the key theoretical point underlying this argument: As income rises, health becomes increasingly valuable (relative to other consumption goods) because it extends the time and ability to enjoy the higher standard of living that rising income affords. Thus, they argue that health care spending in the U.S. should *optimally* grow much faster than income, reaching at least 30% of GDP by 2050. In this view, slowing the growth of health spending could therefore be misguided. More recently, Fonseca et al. (2009) have developed a lifecycle calibrated model of longevity and healthcare spending, and emphasized the importance of technology growth in healthcare, which they predicts will have a larger impact on healthcare expenditure growth than does rising income.

In this paper, we develop a new model that can both reconcile these views and explain two otherwise puzzling empirical regularities: (1) The more rapid growth of healthcare spending relative to GDP in the U.S. relative to other developed countries, and (2) the remarkably modest income-elasticity of healthcare spending at a point in time. Our model differs from Hall and Jones (2007) in two main ways. First, we use a more realistic modeling of healthcare productivity that rules out the ability of the very wealthy to buy near-eternal longevity. Second (and more importantly), we specify a model that includes a role for government to finance

[1] While the CBO has recently scaled back its healthcare spending growth predictions, the long-term implications for growth are still quite similar in magnitude; the 2007 long-term projection predicts that 25 percent of GDP would be spent on health in 2035, while the more recent 2012 report projects nearly 25% by 2037.

health care, motivated by the high share of public financing in every developed country. We specify a population with heterogenous income and health and a government that can engage in taxation to fund both cash and in-kind health insurance benefits. We assume that the social welfare function features *health-specific altruism*, based on the classic model of commodity-specific egalitarianism (Olsen 1981). Essentially, society cares specifically about the health of the poor, beyond just their overall utility. Because people do not internalize this altruistic externality when making health care choices, society wants to ensure that everyone has access to decent health insurance, even if they would choose to purchase other goods if given the cash equivalent. These egalitarian social preferences overcome the standard dominance argument for cash rather than in-kind transfers, making a mix of the two optimal in our model. Indeed, a calibrated version of our model is able to endogenously match the high share of U.S. health spending that is publicly financed.

Government financing changes the economics of health spending growth relative to the results of Hall and Jones (2007). Financing higher public health care spending involves both a resource cost *and* an efficiency cost from the excess burden of higher tax rates. Because excess burden rises with the square of marginal tax rates, the marginal efficiency cost of growth in government health spending rises as health care becomes more expensive (necessitating higher taxes to finance). This force puts a brake on optimal government health spending growth relative to optimal private spending growth. This slows the *overall* optimal growth of health care spending over time.

To show this result formally, we use a calibrated simulation model with heterogeneous incomes and a government that chooses taxes and transfers to maximize an egalitarian social welfare function. We calibrate the model parameters to replicate the result of Hall and Jones (2007) that with purely private health spending (i.e., no in-kind health transfers), optimal health spending as a share of GDP rises sharply, reaching over 30% in 2050. We then show how these conclusions change when we introduce stylized health transfers, modeled after the Medicare program. In this more realistic setting, tax-financed public health insurance that redistributes resources to the poor, disabled, and elderly entails large efficiency costs (Browning and Johnson 1984). Calibrating this model to match both the U.S. and European countries with higher tax-to-GDP ratios, we demonstrate that Europeans countries should have experienced optimally slower growth in healthcare expenditures relative to GDP since 1980, given the efficiency costs of raising the tax-to-GDP ratio above 50 percent. Furthermore, the same (optimal) slowdown in healthcare spend is observed for the U.S., starting around 2030, as the tax hikes necessary to fund additional healthcare spending leads to higher distortionary tax effects approaching those in Europe. Unlike the conventional model, the new parameterization of healthcare technology yields cross-sectional income elasticities of about 0.3, consistent with empirical evidence.

Beyond modeling optimal health spending in a future with continued growth in Medicare expenditures,

the model also yields insights into the consequences of different forms of public insurance. Medicare in the United States provides a uniform benefit to enrollees – a "one size fits all" approach to coverage. The average annual Medicare spending per beneficiary of about $9,000 represents almost half of the median income of beneficiaries ($21,000) and a fifth of the median total savings ($53,000) (Kaiser, 2011). Incorporating income heterogeneity into the model allows us to gauge the consequences of having a generous uniform benefit that might match the preferences of higher income households but be substantially more than lower-income houses would choose to spend on health instead of other goods. That is, Medicare currently covers proton beam therapy for prostate cancer – at a cost of $25,000 more than other treatments such as prostatectomy or radiation therapy and with little evidence of better outcomes than those alternatives (see Baicker et al., 2012). Low-income beneficiaries might be better off with an insurance plan that didn't cover proton beam therapy, but instead provided more financial resources to buy food or rescue their home from foreclosure.

We also consider an alternative policy; a basic insurance plan for everyone, leaving higher-income households free to use private funds to "top up" their coverage to include a wider range of less cost-effective care. This arrangment is not dissimilar from voucher-type premium support suggested by Emanuel and Fuchs (2005), Aaron and Reischauer (1995), and Ryan (2012), in which the government provides enough money to purchase basic insurance and enrollees choosing more extensive coverage pay the incremental cost themselves. Reinhardt (2012), concerned as well about the financial viability of Medicare, suggested a three-tier system again broadly consistent with our model: High-income households are eligible for "the sky's the limit" insurance, largely unsubsidized by the government, while middle-income households receive a "reference pricing" insurance policy that reimburses only for the lowest-cost regional provider, and low-income enrollees receive Medicaid-like public care under strict cost control rules.

Many countries in Europe already offer bifurcated plans, whether de facto, as in England, or de jure, as in the Netherlands (van de Ven and Schut, 2008). The "basic" plan we study differs from these proposals, however, in that that the baseline coverage does not raise copayments and deductibles, which undermines the financial protection of low-income households. Instead, the basic plan saves money by limiting coverage to those services with documented and reasonably effective health benefits; thus the basic plan would not cover proton beam therapy, nor would it cover treatments such as Provenge, a $93,000 drug that extends lifespan for patients with metastatic prostate cancer by a few months. This less costly basic coverage does not imply less public spending on redistribution overall, but it would imply a different composition of redistribution; a greater subsidization of housing and food, and less of health care. In the next section, we consider the emprical patterns of healthcare spending, both at a point in time across income groups, and over time and across countries, hat we wish to match up with our simulation model.

2 Empirical Patterns of healthcare utilization

We first consider two broad empirical regularities that we believe models of healthcare utilization and growth should be able to replicate. The first relates to the cross-sectional relationship between income and healthcare utilization within and across countries, while the second relates to the growth in healthcare spending between the U.S. and European countries during the past several decades.

First, commonly used utility functions (such as Hall and Jones, 2007) predict dramatic differences in healthcare utilization by income. If individuals have constant relative risk aversion of 2, for example, diminishing marginal utility suggests that someone with (non-healthcare) consumption of $100,000 would demand 100 times more healthcare than one with $10,000. Yet we do not observe such steep differences in the data. High income households in the U.S. spend a much smaller share of their income on health care than lower income households. For example, Burtless and Svaton (2010) show that the dollar amount spent per person on health is essentially *flat* across income groups. As a result, the top income decile of non-elderly households devotes 2% of its gross money income to health insurance and health care compared with 65% for the bottom decile; the top income quintile consumes only about 10% or $300 more health care per person, while earning 1500% more.

It is difficult to make inferences about the pure income elasticity of demand from U.S. data, given the importance of insurance coverage, and particularly the impact of Medicaid and Medicare. Indeed, in our model, narrowing inequality in health spending is a key purpose of government-financed health care. Hence, we draw upon evidence from other settings where public health care is less generous. Evidence from India suggests a much larger income elasticity of health spending. Banerjee, Deaton, and Duflo's 2004 survey of rural Udaipur, India, shows that the top 1/3 of the income distribution spends 11% of its budget on health care, while the typical household spends 7.3%; and Banerjee and Duflo (2007) find modest elasticities across income groups in many different emerging economies. Similarly Das, Hammer, and Sanches-Paramo (2011) find a pronounced positive income elasticity of spending – even without controlling for the poorer health of low-income respondents – in urban areas of India.

We can also consider the income-elasticity of health spending among the elderly in the U.S. prior to the introduction of Medicare and Medcaid in 1966. Following Finkelstein and McKnight's (2008) analysis of the 1962-1963 Survey of Health Services Utilization and Expenditures, we regress log health expenditures on log family income after controlling for demographic covariates and self-reported health status among respondents over age 65. The estimated coefficient is 0.297 (standard error 0.104).[2] This represents a much

[2]Because most individuals are retired, this reduces the importance of bias from worse health causing both lower income (due to inability to work) and higher health spending. In addition, because the log specification is sensitive to very low values, we trim the sample at the bottom and top 1% of incomes. Symmetric trimming at other levels (2% or 5%) produces nearly identical results, but without trimming, the income-health spending relationship is much weaker and not significant. We interpret this

larger income-health spending gradient than in the Burtless and Svanton (2010) data from recent decades.

There is also a steeper relationship between income and health spending across regions or countries compared to the modest elasticities at the individual level within countries. Getzen (2000) summarizes empirical research suggesting much higher elasticities (often in excess of 1) across countries, moderate elasticities across regions within countries, and much more modest (often around 0) elasticities at the individual level within countries. For example, an examination of health spending from 20 OECD countries from 1980-2006 shows an elasticity of healthcare expenditure growth with respect to GDP growth of 0.74 (t = 3.9). Acemoglu, Finkelstein & Notowidigdo (2009) use increases in regional income in oil-producing areas of the U.S. when oil prices rise to identify an income elasticity of health spending of around 0.7.

The second empirical regularity is the differential growth in the healthcare to GDP ratio between the United States and other OECD countries. For example, in 1980 U.S. healthcare spending was 9.0 percent of GDP. In the same year, three European countries had healthcare spending similar to the U.S.: (West) Germany (8.4 percent), Denmark (8.9 percent) and Sweden (8.9 percent). The ratio of healthcare spending to GDP in the U.S. grew by 7 percentage points between 1980 and 2008 (from 9 to 16 percent), but the corresponding ratio in the three other countries grew by just 1.1 percentage points; across all OECD countries, the corresponding average increase was 2.9 percentage points. [3] In the Hall and Jones (2007) model, one might explain the higher growth in health care expenditures by a more rapid growth rate of income, but annual U.S. real per capita GDP growth during this period was 1.8 percent, only slightly higher than the corresponding rates for the three European countries, ranging from 1.4 percent to 1.7 percent. Nor can one explain the difference in expenditure growth by technological innovations, since new technologies were equally available in the U.S. and elsewhere. One must explain therefore why European countries with equivalent, or even higher rates of income growth experienced such slow growth in the share of GDP devoted to healthcare. We next turn to a model to explain these empirical patterns.

3 A Model of Private and Public Health Spending

We begin with a standard model of health care demand based on the model in Hall and Jones (2007) and Murphy and Topel (2006). We modify these models in several ways. First, we move away from the representative agent approach, instead modeling a population with heterogeneous sickness probabilities and wages and adding a labor supply decision. We show, as in previous work, that optimal health care spending rises with wages and income.

as bias from a few people with very low incomes obtaining significant amounts of health care.

[3] See Chandra and Skinner (2012). German health care cost growth was derived by chaining together spending growth in West Germany until just prior to unification, and unified Germany growth post-unification.

Second, we modify the health production function of Hall and Jones (2007) to make it more suited to modeling the cross-sectional health spending distribution. In particular, we show that their prediction that health spending is a luxury good – higher as a share of income for the rich than the poor – arises because the health production function is unbounded – any level of longevity can be purchased with sufficient spending. We instead specify a bounded health production function that captures more closely the clinical realities that, at some point, there is no amount of money that will buy better health.

Third, we introduce a role for government through a social welfare function that makes redistribution desirable. In addition to the standard utilitarian rationale for cash transfers, we include *egalitarian* social preferences for better health, particularly for the poor. As a result, our model generates endogenous cash and in-kind health transfers financed by distortionary income taxation.

3.1 Individual Model

Consider a stylized model of individuals' labor (L), consumption (C), and medical care (M) choices, similar to the basic model of Hall and Jones (2007) with a labor choice added. Individuals derive flow utility from consumption and disutility from labor, and medical spending determines their life expectancy, $\lambda(M)$. The individual problem is:

$$\begin{aligned} V &= \max_{C,M,L} \lambda(M) \cdot u(C,L) \\ &\text{s.t.} \quad wL + y^u = C + M \end{aligned} \tag{1}$$

where w is wage and y^u is unearned income, both of which vary across individuals, and $\lambda' > 0$, $\lambda'' < 0$. Here, M, C, and L are annual values, and $u(C,L)$ is annual flow utility. Utility of death is normalized to zero, so lifetime utility equals life expectancy, $\lambda(M)$, times the annual flow utility.

The main prediction of this model is that medical spending rises with wages and income. This can be seen from the first-order condition for M, which can be rearranged to be:

$$Value_{LY}(C,L) \equiv \frac{u(C,L)}{u_C(C,L)} = \frac{\lambda(M)}{\lambda'(M)} \equiv MCost_{LY}(M) \tag{2}$$

The right-side is the marginal cost of extending life by one year,[4] which rises with M because of diminishing returns in the production function. The left-side is the (monetized) value of a life-year. The basic insight of Hall and Jones (2007) is that the value of life rises steeply as income increases. As a result, optimal medical spending is increasing with income. Relative to the poor, the rich proceed further up the marginal

[4] Increasing M by \$1 annually means an expected lifetime cost of $\$\lambda(M)$ and yields an increase in life expectancy of $\lambda'(M)$. So the marginal cost per life-year is $\lambda(M)/\lambda'(M)$.

cost curve, spending more on less valuable services, until $MCost_{LY}(M)$ equals their much higher value of a life-year.

In practice, these income differences in the value of a life-year tend to be extremely large. To see this, temporarily ignore labor[5] and consider a numerical example using a constant relative risk aversion utility specification both we and Hall and Jones (2007) use: $u(C) = b + \frac{1}{1-\gamma}C^{1-\gamma}$. As in their work, the constant b is a free parameter: it is calibrated to match an exogenously specified value of a life year based on empirical estimates from outside of the model. For instance, we might calibrate a value of a life year of $\overline{LY}$ =\$100,000 (Cutler 2004) at consumption C_0 =\$30,000. After imposing this calibration, it is easy to show that the value of a life-year grows at least as fast as γ times the growth in consumption, implying extremely large cross-sectional differences. For instance, with $\gamma = 2$ (as Hall and Jones use), an upper-income \$100,000 consumer values a life-year at *\$1.34 million* – an order of magnitude larger than the values used by health economists. In the other direction, a low-income person consuming \$10,000 has a life-year value of just \$4,400, less than half their annual income. Even with the lower value of $\gamma = 0.5$ that we will use, the differences are large: \$4,300 for the low-income person and \$273,000 for the high-income person.

Given the large income heterogeneity across the population (even after taxes and transfers), an efficient health care system would (in theory) apply widely varying cutoffs for cost-effectiveness by income group. The fact that health economists (e.g., Cutler, Rosen and Vijan 2006) and official government guidelines usually apply a uniform value of life across wide income groups implies a strong egalitarian preference that appears to preserve an equitable redistribution of resources. However, as we discuss in Section 3.3, focusing solely on one dimension of equality, healthcare, while ignoring other dimensions of equality such as housing or food consumption, can potentially lead to inefficiency in redistribution, even with egalitarian preferences for health equality.

3.2 The Health Production Function

The implications for health care spending of these wide variations in the value of life depend on the shape of the health production function. If most care were either cheap and effective or expensive and ineffective (mathematically, a highly concave production function), variations in the value of life would be irrelevant for spending. But in reality, medicine has both effective, expensive treatments and gray-area treatments whose effectiveness are heterogeneous and uncertain (Chandra and Skinner 2012). Because of its importance in the welfare implications (and predictions) of the model, we pay special attention to specifying a flexible and realistic health production function. Our starting point is the constant-elasticity specification used by Hall

[5]Later, we will include a separable labor disutility term that does not materially affect this line of reasoning.

and Jones (2007):[6]

$$\text{Constant Elasticity PF:} \qquad \lambda(M) = A \cdot M^{\theta} \tag{3}$$

where A is a constant that captures technology and health differences over time and across age groups (but is invariant across income groups). As noted above, with this production function and a coefficient of relative risk aversion, γ, exceeding one, healthcare becomes a luxury good; even very wealthy households will spend an increasing fraction of their income on health care. But even when$\gamma < 1$, so that healthcare is a necessity, the constant elasticity production function still generates unrealistic health spending at high incomes. Figure 1 shows the case for $\gamma = 0.5$ and $\theta = 0.15$ (near the middle of the Hall and Jones estimates across ages). The left graph confirms that health spending is now a necessity, with its income share falling. But the share falls slowly at high incomes, and it is possible to prove that as income grows large, the health share of income is bounded below by $\frac{\theta}{1+\theta}$ (see Appendix A). This implies that (for our estimate of $\theta = 0.15$) health spending among households will aways exceed 13 percent of income for *any* income level. As the right figure shows, health spending is almost \$50,000 per year for \$200,000 earners and shows no sign of leveling off with income.

In the appendix, we show that the problem lies in the *unboundedness* of the health production function, the property that any longevity can be achieved with sufficient money. With an unbounded production function, the elasticity of longevity with respect to health spending does not decline quickly enough to produce reasonable health spending predictions for high-earners. The constant elasticity production function is unbounded, but so are most other commonly used production functions. Indeed, unboundedness is a natural property when modeling economic output. But it does not work well in modeling health care technology at a point in time.

Further, there is good clinical evidence that unboundedness does not hold in reality. At some point, there just isn't anything physicians can do to treat a patient without introducing even more risk and potentially harming the patient.[7] At higher levels of expenditures, there is no consistent association between spending and health outcomes (for a review, see Skinner, 2012). And even studies that show better outcomes associated with higher rates of spending, there are clearly diminishing returns; Doyle et al. (2012) for example found

[6] Their specification is technically for "health status," a variable that is proportional to our life expectancy function $\lambda(M)$.

[7] An interesting anecdote illustrates this point. An ICU physician, Goetz (2004) stated in a letter to *Health Affairs*: "Here is an example I have used when teaching medical students and residents: You are taking care of a patient in the ICU. You have done every test and procedure you know to do and have done everything that all the consultants have recommended. I now tell you that you must spend another \$5,000 (originally I used \$1,000) to improve the patient's quality of care. What would you do with the money? By this point the student or resident is in a bit of a quandary because they are not quite sure how to use the additional money. If there were a continuing positive linear relationship, it should be reasonably easy to suggest more things that result in improved patient care. Generally the suggestions are more, or repeated, tests and procedures. I respond to the common answers with a statement that if you do more tests or procedures, you could in fact make the patient worse. How? If you do more tests, all tests have false positives and negatives. How will you use results that contradict earlier tests? With again more tests, and the subsequent potential for much more confusion. If you repeat or do another procedure, how do you interpret the results? Also, procedures generally have potential side effects or complications, so again you have a very high risk of NOT improving quality or outcome with more money."

average beneficial effects of being taken by ambulance to an emergency room in a more expensive hospital, except for the most expensive 17 percent of those hospitals, where more spending was associated with no better or even worse outcomes – a result consistent with a bounded health production function.

Based on this evidence, we develop a bounded health production function to use for our model. Consider an individual who lives over a series of periods. In each period, she becomes acutely ill with probability σ and is healthy otherwise. While healthy, survival is certain. (This abstracts from preventive care and factors like exercise that could affect the probability σ.) Conditional on an acute illness, survival occurs with probability $s(m)$, where m is medical spending per acute event. We model the survival probability $s(m)$ as having a minimum value s_0 that can be obtained with no medical treatment and as increasing up to some technological frontier s_{max}. Our key assumption that ensures boundedness of the production function is that $s_{max} < 1$, so mortality can never be reduced to zero. Formally, our survival function is:

$$s(m) = s_0 + F(m) \cdot (s_{max} - s_0)$$

where $F(0) = 0$, $\lim_{m\to\infty} F(m) = 1$, and $F' > 0$, $F'' < 0$. A simple choice for $F(.)$ that satisfies all of these properties is the exponential CDF: $F(m) = 1 - \exp(-\alpha m)$ for $m \geq 0$ (where α is a shape parameter).

In this specification, the per-period mortality rate equals $\sigma \cdot (1 - s(m))$. We assume a stationary setting so that life expectancy equals the inverse of mortality. We also assume costs are fully insured so expected health spending is $M = \sigma \cdot m$. The health production function is therefore:

$$\text{Max Survival PF:} \qquad \lambda(m, \sigma) = \frac{1}{\sigma(1 - s(m))}$$

Note that because $s(.)$ is bounded above by s_{max}, $\lambda(m, \sigma)$ is bounded by $[\sigma(1 - s_{max})]^{-1}$. For instance, starting at age 65 with $\sigma = 0.3$ (approximately the annual elderly hospitalization rate) and $s_{max} = 0.85$, the life expectancy upper bound is 22.2 years (or age 87.2). Note that this is a bound on *expected* longevity that can be purchased with better health care, not on maximum possible longevity. We discuss how we calibrate the parameters in this model in Section 4 below.

Figure 2 displays the predictions of this production function for health spending. More realistically, $200,000 earners only spend about $7,700 per person on health care, and if we extended the graphs, someone earning $5 million would spend $10,900. Health care is again a necessity, but now its income share declines to zero as earnings grow large, the key property necessary to ensure reasonable health spending by the rich. Recall that this implied level of healthcare spending corresponds to a hypothetical case without public health insurance. For moderate income levels, the simulated income elasticity of healthcare spending is not

inconsistent with the empirical estimates of roughly 0.3 among those without insurance, as noted above.[8]

How can we reconcile health care's status as a necessity (income elasticity below 1) in the cross-section and more of a luxury in the time-series? The nature of technological change provides a natural answer. In our model, technological change takes two forms,[9] as can be seen in Figure 3. First, technological improvement can increase s_{max}, leading to an increase in both the average and marginal returns to medical care. Increases in s_{max} therefore raises optimal spending. This form of technical change tends to exacerbate inequality in a world with private insurance – only the wealthy can afford to approach s_{max}. This sort of technical change reflects the view that between 1950 and today, expensive treatments developed that, absent public insurance, low-income people would not have been able to afford (Nyman, 1999).

The second form of technological change is an increase in α. A higher α also raises survival at any given m but does not affect the technological frontier survival, as shown in the bottom graph of Figure 3. This can be thought of as efficiency-improving innovation or inexpensive new treatments that substitute for expensive older treatments. One example would be the use of beta blockers after heart attacks; as shown in Skinner and Staiger (2009), increased use of these drugs raised marginal returns at low levels of health spending but reduce marginal returns at high levels, increasing the concavity of the production function. This form of technical change tends therefore to reduce inequality and may even reduce total health expenditure if sufficient people start as high spenders.

3.3 Social Welfare and Government Problem

Even with the max survival health production function, Figure 2 shows that there are significant income gradients in health care spending (and even greater inequality in consumption) in the absence of health insurance. Primarily to improve equity in healthcare and consumption, governments often play a central role in health insurance markets. In most developed countries, relatively generous public health benefits are available, either to all citizens or to the poor and aged. These in-kind benefits enable the poor to obtain expensive care beyond what they could afford on their own and are therefore critical to integrate into a model of health care spending.

In this section, we present a model that incorporates a role for government health benefits and illustrates the tradeoffs involved in setting them optimally. In doing so, we are forced to confront the longstanding puzzle of why a government would ever use in-kind transfers, rather than transfering the same resources in cash. The argument against in-kind transfers is simple: cash costs the same to taxpayers but raises recipients' utility more because it does not constrain their choice sets. Though powerful, this logic is not

[8] It is perhaps more difficult to make the connection to developing countries where incomes are generally much lower, since the production function for health in (e.g.) neighborhood clinics will also be shifted in.

[9] In addition, increases in s_0 capture improving health due to lifestyle factors such as better nutrition and reduced smoking.

airtight, and a large theoretical literature has identified reasons why in-kind transfers can be socially optimal (see Currie and Gahvari 2008 for a review). Several rationales for in-kind health benefits that emerge from this literature include: social egalitarian preferences, individual optimization failures, market unravelling due to adverse selection, and tagging. Though conceptually distinct, these ideas share a common implication that the government should finance at least a minimum level of health insurance for everyone.[10] Thus, we view all of these rationales as complementary ways of generating endogenous in-kind health benefits. Nonetheless, we base our social welfare function solely on the rationale of egalitarianism. We do so both for simplicity and because egalitarianism best captures the idea that society specifically wants to provide better health care for the poor than they would purchase on their own. It also captures the idea that society may not normatively accept the large variation across income groups in the revealed preference value of a life-year that we discussed in Section 3.1 above.

We base our social welfare function on the classic formulation of commodity-specific egalitarianism by Olsen (1981). This model captures the idea that the public may be especially concerned when the poor cannot or do not obtain adequate food, shelter, and other basic commodities. In the model, people are altruistic but instead of valuing others' utility, they value their consumption of a specific commodity. We set the egalitarian commodity to be life expectancy $\lambda(.)$, meaning people get altruistic utility when others live longer.[11] Formally, we define egalitarian welfare for an individual i as:

$$W_i = \lambda(m_i, \sigma_i) \cdot u(C_i, L_i) + \hat{u} \cdot \left[\frac{1}{N-1} \sum_{j \neq i} \lambda(m_j, \sigma_j) \right] \tag{4}$$

The first term is people's private utility, as in the model above. The second term is the egalitarian addition. We assume that people place value $\hat{u}$ on raising the life-expectancy of the remainder of the population by one year.[12] Summing across individuals, this yields a simple expression for the social welfare function:

$$SW \equiv \sum_i W_i \quad = \quad \sum_i \lambda(m_i, \sigma_i) \cdot [u(C_i, L_i) + \hat{u}] \tag{5}$$

[10] Egalitarianism, in which society specifically values individuals' health rather than just their utility, is the classic rationale for in-kind benefits (Currie and Gahvari 2008), since it implies that individuals privately underconsume health care. Mandating coverage is the classic policy response to adverse selection, and mandates are also sensible if individuals are non-optimally foregoing insurance, perhaps because of biased beliefs about the risk of a health shock. However, a mandate alone is a lump-sum tax, so optimal tax theory implies that the mandated benefits should be government-subsidized, at least for the poor. As a result, both adverse selection and individual optimization failures rationalize government funding for the poor for the minimally mandated coverage. Similarly, government-financed health insurance benefits make sense under the tagging rationale because health benefits efficiently redistribute to the poor because the poor are more likely to be sick and need care.

[11] People may also value others' health outcomes more generally, but in our simplified model, increased life expectancy is the only output of health care. Because life expectancy is a function of health care spending in our model, this is equivalent to setting health care itself as the egalitarian commodity with a value function equal to the health production function.

[12] We use a fixed value per life year for simplicity, but our main argument would be robust to incorporating diminishing returns in the valuation of life years. In addition, the specification could be mapped to one in which each individual cared especially about the longevity of a few others (e.g., close friends and relatives), so long as this externality is not fully internalized through Coasian bargaining.

There are two things to note about this social welfare function. First, as opposed to the private utility of life, the egalitarian utility of life $\hat{u}$ does not vary with income or other characteristics. This captures the sense that people may not want to apply a higher value of life for those with higher incomes, despite the logic in Section 3.1 above. It also accords with the way economists conduct cost-effectiveness analysis by valuing the life years of all people equally (e.g., at $100,000), rather than placing a higher value on the those of the rich. Second, the health of society is like a large public good in this model. Improving the health of individual i yields altruistic utility $\frac{1}{N-1}\hat{u}$ to everyone else in society in a way that is non-rival and non-excludable. Even if this altruism is weak, it is the sum of the altruistic marginal utilities, $\sum\left(\frac{1}{N-1}\hat{u}\right) = \hat{u}$, that matters for social welfare, as in the Samuelson (1954) rule for public goods. As with other public goods, people will invest too little in their health in a purely private market. This "under-consumption" of life will be most severe for lower-income groups, since they start out spending less in a private market and health spending is subject to diminishing returns. Therefore, a natural policy response is for the government to provide or mandate a minimum level of health insurance benefits.

Based on this, we define the government's problem as choosing a schedule for taxes $T(.)$ and a level of cash benefits (R) and in-kind health benefits when sick (m_{pub}) to maximize social welfare subject to a government budget constraint:

$$\begin{aligned} SW &= \max_{\tau, R, m_{pub}} \sum_i \lambda(m_i, \sigma_i) \cdot [u(C_i, L_i) + \hat{u}] \\ &\text{s.t.} \quad \sum_i [T(w_i L_i + y_i^u) - R - \sigma_i m_{pub} - E] = 0 \end{aligned} \tag{6}$$

where E is exogenous government spending (e.g., national defense, roads, schools), which we assume scales with the population. The government budget constraint requires that the sum of lifetime taxes must equal the sum of lifetime benefits and extra costs over all individuals. The variables $\{m_i, C_i, L_i\}$ are determined by the individual problem in (1) above, with the budget constraint modified to include taxes and public health benefits m_{pub}:

$$(w_i L_i + y_i^u) - T(w_i L_i + y_i^u) + R = C_i + \sigma_i(m_i - m_{pub}) \tag{7}$$

We consider two stylized ways to structure the health insurance program, which impose different constraints on individuals' choices:

1. **Uniform Insurance:** Government finances equal health care for everyone: $m_i = m_{pub}\ \forall i$
2. **Basic Insurance:** The government benefit is a floor on individuals' health care: $m_i \geq m_{pub}\ \forall i$

Note that we have modeled uniform insurance as one in which individuals cannot top-up their benefits. Few

countries place restrictions on private markets in health insurance; certainly in the U.S. one can purchase services with cash, particulalry for "concierage" or luxury services. We characterize the current Medicare program, however, as a uniform service because so few people actually seek treatment outside of Medicare, because (as we discuss below) its coverage is commensurate with a what a private plan for high-income individuals would be, in the sense of covering nearly any possible treatment regardless of cost-effectivenss. This plan best enacts the egalitarian ideal for healthcare, with all citizens receiving (by the standards of the world) gold-plated healthcare. We acknowledge that Medicare does not cover all of the overall spending for the elderly, but the point is that it covers essentially all treatments and providers.[13] Further, 90 percent of the elderly have Medicaid or a private plan (retiree health insurance, Medigap, or Medicare Advantage) to cover the Medicare's out-of-pocket expenses, so that on net, there is little meaningful variation in health insurance coverage generosity among the elderly.

By contrast, basic insurance permits some health care inequality by allowing individuals to "top-up" the basic insurance benefit. Note that for large values of m_{pub}, the two systems are nearly identical, since few people will choose to top-up the generous public coverage (as we noted above). Therefore, the two systems materially differ only when m_{pub} is modest – i.e., covering mainly the "basics" (hence the name). In this case, higher-income people will top-up while the poor will not, and there will be inequality in insurance coverage.

In the stylized framework above, basic insurance strictly dominates uniform insurance: uniform coverage imposes a ceiling on medical coverage that constrains the utility of the very rich without helping anyone else. In reality, uniform insurance is often enacted because it is administratively simpler (a single payer versus many payers) and avoids adverse selection and other market failures that may plague the market for "topped-up" insurance. To model this, we assume that under basic insurance, a share ϕ of health spending is lost to these non-health care costs. So for basic insurance, we replace values m_i and m_{pub} with $(1+\phi)\, m_i$ and $(1+\phi)\, m_{pub}$ in the government and individual budget constraints (but not in the health production function). This ensures that uniform insurance is no longer dominated, and as we will show, is actually optimal in many cases.

We will discuss most of this model's implications using simulations with calibrated parameters. However, we make a few observations in the special case of uniform insurance with homogenous illness rates ($\sigma_i = \sigma$ $\forall i$), based on the government's first-order condition for m_{pub} :

$$\frac{\overline{u\left(C,L\right)} + \hat{u}}{\alpha} = \frac{\sigma \cdot \lambda\left(m_{pub}, \sigma\right)}{\lambda'\left(m_{pub}, \sigma\right)} \tag{8}$$

where α is the government's marginal cost of funds (the Lagrange multiplier on its budget constraint) and

[13] The main exception is long-term care, which is covered by Medicaid, although a recent court ruling may lead to a larger role for Medicare in covering such benefits.

$\overline{u(C,L)} = \frac{1}{N}\sum u(C_i, L_i)$ is the average value of utility over the population. Equation (8) is analogous to individuals' FOC in equation (2). The right-side is the marginal cost of saving a life-year, just as before (with a slightly different form because of the change of variables from M to $m = M/\sigma$). The left-side is now the *social* value of a life-year – social utility of a life-year divided by the marginal cost of funds.[14]

Equation (8) shows one of the key points in this paper: the optimal public health benefit depends heavily on the marginal cost of funds, α. The marginal cost funds, in turn, increases as tax rates rise because of increasing marginal excess burden (since excess burden is proportional to the square of marginal tax rates).[15] Therefore, as taxes increase – either to fund expensive medical care or other programs – optimal health benefits decline. Effectively, the government sets public health benefits to balance gains in health equity (arising from the egalitarian externality $\hat{u}$) against the efficiency costs of taxation. As health technology improves and taxes rise to fund expensive new care, efficiency costs rise. Therefore, the rising excess burden of taxation puts a natural brake on optimal growth of public health care costs.

3.4 High Deductibles versus Limited Benefits

While one might interpret the basic coverage insurance plan as a conventional "high-deductible" plan with broad coverage but significant cost-sharing, this is not what we mean by basic insurance. In our model, all health care spending is fully insured without cost sharing. The variation in insurance coverage is therefore along a different dimension: what is covered (and perhaps also which providers are covered). In our simple framework, individuals (or the government) choose an overall spending level m that is covered in the event of illness. This equivalently sets a cost-effectiveness threshold for which treatments/providers are covered. Treatments with cost-effectiveness below this threshold would not be covered, or equivalently covered at a "reference price" of the cost-effective treatment.[16]

This basic plan would therefore resemble the English health insurance coverage under the National Institute for Health and Clinical Excellence (NICE), which uses a cost-effectiveness hurdle to judge which treatments are publicly covered, and which are not. Importantly, individuals wanting more extensive coverage of treatments and/or providers could purchase top-up coverage using private funds. In that respect, it would resemble the premium support plan of Ryan (2012), which provides a public amount for purchasing insurance and lets people top up that amount with their own money to purchase more generous coverage.

The distinction between high deductibles versus limited benefits is important because of the different

[14] By the government's FOC for R, α equals a longevity-weighted average of marginal utilities of consumption. So the social value of a life-year again equals (average) utility divided by (average) marginal utility of consumption.

[15] Kaplow (2008) argues that the marginal cost of funds is always 1.0 when the government can offset any distributional effects of the tax-and-expenditure program. As is discussed below in Section 6, Kaplow's special case does not hold in our model because the government sector does not "undo" any distributional effects.

[16] For example, the system might pay for intensity-modulated radiation therapy (IMRT), robotic prostatectomy, and proton beam therapy for prostate cancer patients at the rate one would pay for a standard open prostatecomy.

predictions of economic theory for these two dimensions of insurance. As our analysis and that of Hall and Jones (2007) show, *health care* is a normal good, with the rich demanding and being able to afford more of it. By contrast, conditional on an amount of spending when sick, *health insurance* is an inferior good, given the usual assumption of utility with decreasing absolute risk aversion. Intuitively, a given size risk is more costly for a poor person for whom this risk constitutes a larger portion of income.[17] Our model abstracts from the cost-sharing dimension and considers only the amount of health care purchased. But a more general theory would predict that a basic insurance plan designed for the needs of the poor should have low cost-sharing and limited coverage of cost-ineffective services/providers – similar to our basic plan and exactly the opposite of high-deductible plans.

4 Simulation Model and Calibration

We will illustrate most of our results using a simulated version of the two-level government and individual problem described in the previous section. This section describes how we specify a number of functions and calibrate their parameters to create a model that is consistent with empirical data.

4.1 A Mixed Public-Private Health Care Model

The U.S. health care system is financed by a mixture of public and private insurance. The elderly all get public Medicare, which means that the structure of health benefits may affect everyone's life cycle consumption and labor choices. Before age 65, however, only the poor and disabled are publicly covered (through Medicare, Medicaid, and similar programs), with most people obtaining private coverage. We adjust our model from the previous section to capture these realities, while retaining its stylized nature.

Recall that our model above specified a stationary problem with probability of illness σ and an amount of medical spending covered upon illness, m. We now divide lifetime medical care choices into two intervals, with m^y for the young (before age 65) and m^o for the old (above 65). We also set separate illness rates for these two periods, with $\sigma^y < \sigma^o$. Elderly spending is funded publicly and subject to the public system's

[17] To see this formally, note that with a fixed risk M that occurs with probability p, the gain in utility from obtaining full insurance at premium $\pi = pM(1+\eta)$ (which has markup η) is

$$\begin{aligned} \Delta U_{insurance} &= u(y-\pi) - [(1-p)u(y) + pu(y-M)] \\ &\approx u'(y-\pi)\left[-\pi \cdot \frac{\eta}{1+\eta} + \frac{1}{2}\frac{\pi}{1+\eta} M\left(1-p(1+\eta)^2\right)\left(\frac{-u''(y-\pi)}{u'(y-\pi)}\right)\right] \end{aligned}$$

which is positive whenever:

$$\frac{\frac{1}{2}CARA \cdot (M-\pi)}{1+\frac{1}{2}CARA \cdot (M-2\pi)} > \frac{\eta}{1+\eta}$$

where $CARA = -u''(y-\pi)/u'(y-\pi)$. The left hand side is increasing in CARA, so insurance is more likely to have positive value for individuals with high CARA. Thus, assuming declining absolute risk aversion, insurance coverage is an inferior good.

constraints – i.e., $m^o = m_{pub}$ with uniform insurance, and $m^o \geq m_{pub}$ with basic insurance. By contrast, non-elderly spending is funded privately and not subject to these constraints. For simplicity, we do not separately model public health care funding for the non-elderly.

Given this setup, people experience a constant mortality rate in each interval of life $I = \{y, o\}$ of $\mu^I = \sigma^I \left(1 - s\left(m^I\right)\right)$. Constant mortality implies that, conditional on starting the interval alive, the probability of surviving an additional t years equals $\exp\left(-\mu^I t\right)$. Integrating over this function, life expectancy equals $\int_0^{65} \exp\left(-\mu^y t\right) dt + \exp\left(-\mu^y \cdot 65\right) \int_{65}^{\infty} \exp\left(-\mu^o \left(t - 65\right)\right) dt$. Simplifying this expression yields:

$$\begin{aligned} \lambda_i\left(m^y, m^o\right) &= \left(1 - \exp\left(-\mu_i^y \cdot 65\right)\right) \cdot \frac{1}{\mu_i^y} + \exp\left(-\mu_i^y \cdot 65\right) \cdot \frac{1}{\mu_i^o} \\ &\equiv \lambda^y\left(m^y, \sigma_i^y\right) + \lambda^o\left(m^y, \sigma_i^y, m^o, \sigma_i^o\right) \end{aligned} \tag{9}$$

We use this function in place of the one-interval function $\lambda\left(m, \sigma\right)$ used in the exposition of the model above. Just as above, we set $s\left(m\right) = s_0 + F^I\left(m\right) \cdot \left(s_{max}^I - s_0\right)$ with $F^I\left(m\right) = 1 - \exp\left(-\alpha^I m\right)$, where α and s_{max} now vary between the young and old intervals.

Finally, we adjust the government and individuals' problems to take account of this new structure. We assume that everyone retires after age 65 (so L can be interpreted as the share of total work effort exerted before 65, and there is no labor disutility after that) but that cash benefits go to people of all ages. We assume perfect annuity markets so that the problem can be described with a single lifetime budget constraint.

In summary, the final individual problem is:

$$\max_{C, L, m^y, m^o} \lambda_i\left(m^y, m^o\right) \cdot u\left(C, L\right) \tag{10}$$

subject to the lifetime budget constraint:

$$\begin{aligned} 0 &= \lambda_i^y\left(m^y\right)\left[w_i L + y_i^u - T\left(w_i L + y_i^u\right) + R - C - \sigma_i^y m^y\right] \\ &\quad + \lambda_i^o\left(m^y, m^o\right)\left[R - C - \sigma_i^o\left(m^o - m_{pub}\right)\left(1 + \phi\right)\right] \end{aligned} \tag{11}$$

plus any constraints on m^o applied by the public system ($m^o = m_{pub}$ for uniform insurance; $m^o \geq m_{pub}$ for basic) and where $\phi = 0$ with uniform insurance and > 0 with basic insurance. The government then solves:

$$\begin{aligned} SW &= \max_{\tau, R, m_{pub}} \sum_i \lambda_i\left(m_i^y, m_i^o\right) \cdot \left[u\left(C_i, L_i\right) + \hat{u}\right] \\ \text{s.t.} &\quad \sum_i \left\{\lambda_i^y\left(m_i^y\right) T\left(w_i L_i + y_i^u\right) - \lambda_i^o\left(m_i^y, m_i^o\right) \sigma_i^o m_{pub}\left(1 + \phi\right) - \lambda_i\left(m_i^y, m_i^o\right)\left(R + E\right)\right\} = 0 \end{aligned} \tag{12}$$

where C_i, L_i, m_i^y, and m_i^o are the individuals' optimal choices given government policy.

We parameterize the health production function as follows. We start by setting the rates of acute illness (σ^y and σ^o), which we proxy for using average hospitalization rates for the young and old. (We do not model variations in acute illness rates within age groups.) Using the number of overnight hospitalizations per person-year for people 0-64 and 65+ reported in the 2010 National Health Interview Survey, we set $\sigma^y = .100$ and $\sigma^o = .258$. We next set s_0^y and s_0^o, the baseline survival probabilities in an acute illness without any medical spending. Because individuals always obtain at least some health care (both in our model and in reality), these values are difficult to identify separately from the remainder of the health production parameters. We therefore set them to generate what seem like reasonable values for life expectancy at birth (55 years) and at age 65 (5 years) if health care were unavailable; given the values of the σ's, this implies $s_0^y = 0.897$ and $s_0^o = 0.225$.

Finally, we set s_{max}^I, and α^I for $I = \{y, o\}$ to match empirical moments of life expectancy and medical spending to their corresponding model values. We calibrate them in the setting where there is only private health care for both non-elderly and elderly ($m_{pub} = 0$), since this more closely matches the specification used by Hall and Jones (2007). Solving the optimal government and individuals' problems above (using the other parameter values specified in this section), we search for the $\left\{s_{max}^I, \alpha^I\right\}_{I=\{y,o\}}$ vector that best matches the moments. We use four sets of moments to identify these four parameters. First, we match life expectancy at birth and at age 65 from SSA period life tables (Bell and Miller 2005) to the corresponding values in our model. This helps most to identify s_{max}^y and s_{max}^o, since in our model, most people obtain enough health care to get close to the frontier survival. Next, we use private and publicly financed medical spending as a share of GDP from the National Health Expenditure Accounts. These help pin down the α's, since higher values of α imply a more concave health production function and therefore lower health spending overall. We match nonelderly spending as a share of GDP in our model to *total private* spending in the NHEA and elderly spending in our model to *total public* spending in the NHEA. While this is not quite correct (in reality, some non-elderly spending is private and some elderly spending public), our main goal is to match overall public and private health spending to ensure realistic estimates of the excess burden of tax-financed care. In our model, the elderly proxy for government-financed health spending, and the non-elderly for privately financed spending. Therefore, we adopt the current strategy as an approximation.

4.2 Utility Function Parameterization

We use a standard formulation for our flow utility function. Utility is separable in consumption and the disutility of labor, so $u(C, L) = u(C) - \psi(L)$. Consumption utility takes the constant relative risk aversion

form used by Hall and Jones (2007): $u(C) = b + \frac{1}{1-\gamma}c^{1-\gamma}$, where γ is the coefficient of relative risk aversion and b is a constant that sets the utility of life relative to death. We calibrate b so that the value of a life-year equals \$100,000 for a typical person consuming $C_0 = \$30,000$ and working 20% of the time.

Labor disutility takes a standard form often used in the optimal tax literature (e.g., Saez 2001): $\psi(L) = \frac{\psi}{1+1/\varepsilon}L^{1+1/\varepsilon}$. This form is used because in special cases it ensures a constant wage elasticity of labor supply. For instance, with CRRA utility and no unearned income (or individual health care costs), the uncompensated labor supply elasticity is $\frac{1-\gamma}{1+\gamma\varepsilon} \cdot \varepsilon$, and the Hicksian elasticity is $\frac{1}{1+\gamma\varepsilon} \cdot \varepsilon$.[18] Thus, both γ and ε affect the labor supply elasticities and must be calibrated jointly. The parameter ψ is a scale parameter primarily affecting the level of labor supply but not its elasticity. We set it at $\psi = 4450$, which generates average labor supply of $L = 0.20$ per working year (about 35 hours per week) for the population of workers.

In our base calibration, we set the coefficient of relative risk aversion equal to $\gamma = 0.5$. Although lower than the typical range between 1.0 and 3.0 used in the life cycle literature, we choose this value to generate reasonable implications for labor supply. The coefficient of risk aversion determines the size of income effects, and if income effects are too large, labor supply curves no longer slope upward but are backward bending. Chetty (2006) shows that with utility separable in labor, upward sloping labor supply requires $\gamma < 1$, and that $\gamma < 2$ even with fairly high substitutability between consumption and leisure. Drawing upon many labor supply and income effects estimates from the literature, he finds a central estimate for γ of 0.71, with a range of 0.15 to 1.78. We choose a slightly lower value of γ than 0.71 because of evidence of relatively small income effects in response to tax changes (Gruber and Saez 2002).[19]

After fixing γ, we set ε to match an estimated Hicksian elasticity of 0.6, consistent with the sum of the extensive and intensive margin elasticity estimates in Chetty (2012). Once we include unearned income and endogenous health spending, labor supply elasticities are no longer constant across income groups. We set $\varepsilon = 0.903$, which generates an income-weighted average Hicksian wage elasticity of 0.60 and an uncompensated elasticity of 0.34 at our base parameter values.

4.3 Population Wage Distribution, Tax System, and Government Costs

To model a realistic wage distribution for a tax and transfer system, we use data from the IRS Statistics of Income public use file of year 2007 tax returns. We define earned income as the sum of wages and salaries, Schedule C business income, self-employment income, and farm income. We set unearned income as the

[18] This uncompensated elasticity formula is easy to derive from the labor FOC $wC^{-\gamma} = \psi L^{1/\varepsilon}$ and the budget constraint $C = wL$. We have not been able to demonstrate the Hicksian elasticity formaul analytically, but it holds true in our simulations. In an intertemporal setting, ε would be the Frisch elasticity, but our model does not have an intertemporal component so ε itself is not directly relevant.

[19] In the special case above (which turns out to be approximately correct), the uncompensated elasticity equals $(1-\gamma)$ times the Hicksian elasticity. Thus, a value of $\gamma = 0.5$ ensures that our uncompensated elasticity is about half as large as our compensated elasticity, rather than only 29% $(= 1 - 0.71)$ as large.

residual of adjusted gross income above earned income; this measure of unearned income represents 12% of AGI on average. We exclude retirees (proxied by receiving Social Security benefits) and the small number of non-retirees (~3%) with zero or negative earned income. Because our model is at the individual level, we convert income into per-adult units by dividing by two for households that are married, filing jointly.[20] Finally, we scale the whole income distribution proportionately so that the resulting mean income matches U.S. real GDP per capita in 2010.

To solve the model, we discretize the income distribution and solve the individual's problem at these limited number of points. We choose these points carefully to capture both the low incomes important for redistribution and the high incomes important for tax revenue. To capture low incomes, we select the first ten points as the average earnings in each of the bottom 10 vintiles (1/20ths) of the income distribution. To capture high incomes for tax revenue, we first divide the earnings distribution into 20 groups such that each group pays 5% of total income taxes. The bottom half of the earnings distribution pays less than 10% of the taxes, so they are contained entirely in the first two tax groups. The remainder of the discretization points are defined as average earnings in each of the rest of the tax groups. The final values of earnings and unearned income at the discretization points are shown in Appendix B. Using these values of earnings, we set wages endogenously so that $w \cdot L^*$ matches the value of earnings at the discretization point.[21] In our simulations, wage earnings are endogenous but unearned income is assumed to be exogenous.[22]

We use a tax schedule that is a scaled approximation to the progressive U.S. code. Specifically, we specify a smooth function, $T(.)$, approximating the shape of the U.S. federal income, payroll, plus state taxes in 2007. This function smooths over jumps in marginal tax rates (which simplifies computation) and ensures globally increasing marginal tax rates, to avoid problems associated with non-convex budget sets when tax rates fall at the Social Security taxable maximum. Appendix Figure A.1 provides details and shows the resulting marginal and average tax rate schedule. Given this smooth function $T(.)$, the government then chooses the level of taxes through a single parameter $\tau \in [0, 1]$ – normalized to equal the marginal tax rate at the top of the income distribution. In the estimation, τ scales the entire tax schedule proportionately, raising or lowering taxes to ensure budget balance.

The remaining parameters in the government's problem are set as follows. The egalitarian externality $\hat{u}$

[20] Note that our measure of annual income probably overstates inequality, since annual income variation includes temporary fluctuations. In future drafts, we will explore shrinking the income distribution to better match the level of lifetime earnings inequality.

[21] For this exercise, we use the baseline tax schedule (designed to approximate the actual U.S. code), $R = 2000$ (close to the value of the standard deduction and personal exemption), and uniform insurance with $M = \$10,000$.

[22] In a dynamic life-cycle model, unearned income would also be endogenous. We did not use a life cycle model, however, because of dynamic instability arising from future tax rates that are expected to rise because of healthcare cost increases. In a life cycle model, individuals looking ahead work more today, while marginal rates are low (leading to even lower tax rates in a balanced-budget fiscal setting), and work much less when marginal rates rise in the future. This in turn leads to implausible swings in marginal tax rates over time.

is set to equal the utility of a life-year for a person consuming $c_0 = \$30,000$ and working $L_0 = 0.2$, with the value of c_0 growing with wages in our simulations over time. The basic insurance inefficiency term ϕ is set at a baseline value of 5%, around which we will do sensitivity analysis. Finally, exogenous government spending is set based on 30-year historical averages of federal discretionary spending (8.6% of GDP); net interest (2.2% GDP); mandatory spending other than Social Security, Medicare, Medicaid, and income security (as categorized by CBO, 2.2% GDP); and state and local government spending (10.6% GDP), less state and local health care spending (1.2% GDP).[23] The total of these factors is 22.4% of GDP. Because GDP is endogenous in our model (since taxes reduce labor supply), we set exogenous government spending to equal 22.4% of the value GDP would take at our baseline tax schedule (with a top MTR of 44%).

4.4 Simulations over Time

We simulate how the optimal policies and outcomes in our model evolve over time as incomes and health costs grow. To do so, we solve the problem above at a series of different parameter values, which are calibrated to match GDP, health spending, and life expectancy patterns in decade intervals from 1960 to 2010, and projected forward to 2060. Between each decade, there are two main sets of parameter changes: (1) Health production parameters $\{\alpha^y, \alpha^o, s^y_{max}, s^o_{max}\}$ change, with each decade's parameters calibrated to match non-elderly and elderly longevity and public and private health spending moments, as described in Section 4.1 above; and (2) wages grow proportionately at the growth rate of GDP per capita. Thus, the forces underlying the changes in our model are wage growth and technological improvements in health care which improve longevity and create expensive new treatments. In addition, we adjust three more minor parameters to account for wage growth: the labor disutility term ψ grows to prevent an unrealistic secular rise in labor supply;[24] extra government spending rises to stay at approximately 22.4% of GDP; and the egalitarian externality $\hat{u}$ is increased over time to account for rising incomes, by setting it as the utility of life at consumption c^t_0, with $c^{2010}_0 = \$30,000$ and this consumption value growing proportionately with wages over time.

Table 1 shows the moments used for our parameter calibration in each year; note that we aggregate the young and old to match aggregate expenditures relative to GDP. Up to 2010, we use actual values of real GDP per capita, health spending, and period life expectancies. After 2010, we project these values forward as follows. Per capita GDP is projected forward assuming a 1.5% real annual growth rate, based on the

[23] The federal budget figures are averages for fiscal years 1973-2012, from CBO's historical budget tables. The state and local government spending number is an average for 1973-2010 from OMB's Historical Budget Tables. The state and local health spending number is an average for 1973-2010 from CMS's National Health Expenditure Accounts publication.

[24] The rise in ψ can be thought of as proxying for the increased value of leisure because of the invention of new goods, e.g. computers and iPhones. This is one potential explanation for the macro puzzle of why labor supply curves can slope up in the short run but labor supply has not significantly increased (or decreased) with large wage increases over time.

OECD long-term projections from 2011-2060 (OECD, 2012). For life expectancies, we use the projections of the SSA actuaries (Bell and Miller 2005). For health spending, we assume that total health spending grows at a constant decadal rate so that it reaches 35% of GDP in 2050, approximately the level predicted by Hall and Jones (2007) in their base specification. We assume that the government share of health spending grows to 49% in 2020, following the projections of CMS in the National Health Expenditures publication. After 2020, we assume that the government's share of health spending continues its upward trend averaging 3.7% points per decade from 1970-2020. Thus, government spending reaches 64 percent of total health spending by 2060.

5 Results

We first present results for a representive agent model where healthcare expenditures are chosen optimally under perfect insurance markets and no redistribution. As shown in Figure 4, the pattern of growth from 1960 to 2010 follows closely actual growth in healthcare expenditures (relative to GDP). Projected optimal healthcare expenditures continues to grow rapidly as income rises and technological innovations continue, with expenditures growing rapidly through 2050, when it approaches one-third of GDP. That is, our new health production function that limits healthcare spending at a point in time does not change the basic insight by Hall and Jones (2007) and Fonseca et al. (2009) that rising income and technology growth should lead to a continued rise in healthcare spending.

We next expand the model to include households with heterogeneous earnings, and illustrate, in Figure 5, the difference in healthcare expenditures by income group for different degrees of insurance coverage. The curved black line demonstrates the steep income elasticity of demand for private health insurance by income group for very low income levels, but at higher incomes the elasticity is much lower; an increase in income from $100,000 to $200,000 raises healthcare expenditures by only about $2,000 annually. As noted above, we believe that our model better captures the curvature of health insurance demand observed in the data. The straight blue line is uniform coverage, determined optimally (given that it is offered to everyone), while in the case shown here, the basic insurance guarantees a minimum level of healthcare that is not much less than the uniform level, but which is "topped up" by households with income groups over $70,000, so as to more closely resemble spending in the absence of redistribution.

Figure 6 uses the simulation model to attempt to explain the differential growth in healthcare expenditures relative to GDP between 1980 and 2010. The "low-tax" country is modeled on the United States, and exhibits similar growth patterns during these three decades (albeit with slightly higher spending relative to GDP in

2010).[25] We hypothesize that the initial tax burden – that is, taxes used to pay for non-healthcare government programs – could help to explain the divergent pattern we observe. We therefore make just two changes in our economic model – we shift the U.S. mixed government-private healthcare system to one funded entirely through the government, and increase as well the overall tax-to-GDP ratio to a level commensurate with the corresponding ratios in Germany, Sweden, and Denmark, which in 1979 were nearly 50 percent. (Thus some of the additional tax collection goes to replacing private healthcare spending, while the rest goes towards non-health government programs.) Once again, the uniform level of healthcare expenditures is determined optimally given the marginal resource cost of redistribution. The high-tax country begins slightly above the low-tax country in 1980, but ends up spending just over 12 percent of GDP by 2010; a rise of less than 2 percent of GDP. In other words, the distortionary impact of taxation can in theory explain the sharply divergent paths in healthcare spending between the U.S. and other countries.

Looking forward to 2050, Figure 7 focuses on spending for three different scenarios. In the first case, "No Health Benefits," health insurance is funded privately, so redistribution takes place only through non-health government programs. And while even these expenditures are quite income elastic and rise over time, in the absence of a large-scale government program, healthcare spending for the elderly is predicted to have remained below 10 percent of GDP by 2020, and just over 20 percent of GDP in 2050.

The corresponding level of expenditures under a more realistic Medicare-style "Uniform Public Insurance" progam is shown by the blue line. Because we consider here only spending among the aged, the initial level of expenditures in 2010 relative to overall GDP is lower than for all healthcare spending. But the simulations indicate that, by 2040 when spending is near 25 percent of GDP, the distortionary effects of the taxes necessary to pay for this public insurance program restrains the growth in healthcare expenditures quite sharply. Finally, the basic insurance program is shown to lead to much lower growth in government spending on the basic insurance policy – flattening out at around 15 percent of GDP by 2050 – but with a continued rise in private healthcare expenditures, so that by 2050, overall spending is still predicted to be over 30 percent of GDP, with the difference that much of it is funded privately by high-income households. Not shown here is the difference in non-health redistribution, which is larger in the case of basic insurance; by contrast non-health redistribution evaporates by 2040 under the uniform benefit.

Figure 7(b) shows the evolution of the top marginal tax rate under a hypothetical tax structure calibrated to 2010 (note that we do not back-cast the model to capture the very high top marginal tax rates in the 1960s paid by a small fraction of taxpayers); the point here is that through 2010, the top marginal rate is about 50 percent, but by 2020 these rates begin to diverge; the taxes necessary to pay for rising costs in the uniform insurance plan rise to nearly 65 percent, while the marginal tax rate for the basic insurance plan

[25] Note that the empirical data stops in 2006 to abstract from the drop in GDP occuring during the great recession.

has flattened out and even declined somewhat given the stability of the basic plan cost relative to GDP. By contrast, the top marginal tax rate for the "No Health Benefit" scenario has declined as the demand for non-health-related redistribution moderated as economic growth continued. (Recall that we assume all earnings rise at the same rate over time, thus leading to lower absolute levels of poverty.)

6 Conclusion

Rising health care costs have created tremendous budgetary pressures for the U.S. and other countries around the world. There is an inherent conflict between policymakers' desire to avoid restrictions on public insurance benefits and their desire to avoid raising the taxes necessary to pay for the growth of existing benefits. In this paper, we have developed a simple stylized model of health care productivity and spending that incorporates publicly-financed insurance programs with redistribution from rich to poor. By incorporating heterogeneity of income and preferences for health care redistribution in our model, as well as the distortionary impact of taxation, we are able to model aggregate patterns of optimal public (and total) healthcare spending, and to explore the economic impact of changing the basic structure of publicly provided insurance program.

The model explains several puzzling empirical regularities. First, it helps to explain the empirical phenomenon that within-country income gradients in health care consumption are typically quite modest, while time-series and cross-country income gradients are much steeper. Demand-side models can easily generate the latter facts, since they predict that the value of health and longevity grows rapidly with income. But they have more trouble explaining why, even after controlling for health status, American high school dropouts obtain nearly as much health care as college graduates. These modest income gradients largely reflect the explicit provision of public insurance coverage that appears to be benchmarked to reflect the demands of higher-income individuals. Second, it can also help explain why health spending as a share of GDP grew so much faster in the U.S. than in some other countries. The model predicts that countries with a higher initial tax burden experience slower growth in public spending because of the rapidly rising efficiency cost of tax-finance.

The model is also used to assess the efficiency costs of continued growth in Medicare. As innovation drives up potential medical spending, the cost of providing a uniform benefit geared to the preferences of higher income residents becomes substantially costlier: treatments that may pass cost-effectiveness hurdles for higher income beneficiaries may not be worth the opportunity cost of other forms of redistribution to lower-income beneficiaries. More interesting for our purposes are new treatments that are expensive and effective, at least for some patients. This group includes both treatments considered highly effective and those whose effectiveness is more heterogeneous, but where treatments are provided up to the point

where incremental medical benefits are small or even zero, leading to very high cost-effectiveness ratios. Administrators of such an insurance plan are stuck with a dilemma – whether to cover treatments that entail very small benefits relative to cost, and thus lead to ever-increasing healthcare costs (with little improvement in survival and functioning), or whether to disallow treatments, and anger those high-income enrollees who believe that the treatment is worth the cost.

An alternative "basic" form of public insurance benefit might provide a more basic plan to all beneficiaries, leaving higher income groups free to "top up" their coverage with privately-financed policies. This is similar to "premium support" type models (and shares some features of Medicare Advantage and Medicare Part D prescription drug coverage today) but with the lower cost of the basic benefit achieved not through higher copayments or deductibles (which undermine the insurance value of the plan, particularly for lower-income beneficiaries), but through limitation of coverage of lower-value care – and with reduced spending on public insurance redirected to other redistributive programs. Setting the value threshold for coverage under the minimum plan involves a tradeoff between the government's goal of providing health care to the poor and the efficient cost of the taxes needed to finance the care. Of course, identifying ex ante which care is of sufficiently high value is no simple task, but public and private insurance plans are already experimenting with mechanisms for tailoring coverage in this way (Fendrick and Chernew, 2009; Gibson et al., 2011).

Past estimates of the marginal cost of funds have made a similar point to our model; that the tax financing of government programs entails efficiency costs that should attenuate the optimal provision of those government services (Fullerton, 1991). Yet a more recent literature seems to suggest no incremental hurdle rate associated with the provision of government public goods, which might also appear to our health insurance program. For example, Kaplow (2008) has argued that the marginal cost of funds should be one because the government can, through the tax code, undo any redistribution that takes place in providing a public good.[26] As Slemrod and Yitzhaki (2001) and more recently Gahvari (2006) have shown, this result is sensitive to strong assumptions about what the government can (and will) do to undo any restributive effects of the government program. In our model, we avoid this controversy because we simulate a fully-specified model in which our government sector does what governments do – redistribute from one group to another through the provision of government services such as health insurance.

That we find such large efficiency costs of taxation is consistent with the intuition underlying the Browning and Johnson (1984) result – that transferring one dollar in government services from rich to poor requires rotating the entire budget line, so even the large share of middle-income individuals don't actually receive any net transfer – just an increase in their marginal tax rate. And while our in-kind health insurance program

[26] Alternatively, Jacobs (2010) argues that the marginal cost of funds is one given appropriate rescaling so that the relevent reference value is the social marginal value of private income.

may appear to be less distortionary because of the in-kind benefit (rather than cash), there is a further distortionary effect of tax-financed uniform health insurance, which is the crowding-out of private health insurance by public insurance. Crowding out occurs in our model when public insurance provided in-kind substitutes for private purchase, and so it ends up acting as a de facto cash transfer to the highest-income groups.

Our model thus highlights the tradeoffs in different approaches to providing public insurance as health care costs rise – including the current approach of providing a uniform benefit that increasingly crowds out other programs to a less egalitarian model that guarantees only a basic benefit with some redistribution redirected towards other programs. The advent of new technologies with wildly varying cost-effectiveness (Chandra and Skinner, 2012) on top of the aging of the population adds urgency to the evaluation of the sustainability of current programs. Our analysis suggests that the policy of providing a uniform government health insurance benefit that pays for nearly any treatment, regardless of effectiveness, may be increasingly untenable as the burden of paying for such a program rises over time.

References

Baicker K, Chandra A, Skinner J. Saving Money or Just Saving Lives? Improving the Productivity of the U.S. Health Care Spending. Ann Rev Econ. 2012;4:33-56.

Baicker K, Skinner JS. Health Care Spending Growth and the Future of U.S. Tax Rates. In: Brown J, editor. Tax Policy and the Economy. Chicago: University of Chicago Press; 2011. p. 39-67.

Browning EK, Johnson WR. The Trade-Off between Equality and Efficiency The Journal of Political Economy, 1984;92(2):175-203.

Chandra A, Skinner JS. Technology Growth and Expenditure Growth in Health Care. JEL. 2012;50(3):645-80.

Chetty, R. A New Method of Estimating Risk Aversion. American Economic Review 96(5): 1821-1834, Dec. 2006

Chetty, R. Bounds on Elasticities with Optimization Frictions: A Synthesis of Micro and Macro Evidence on Labor Supply, Econometrica 2012;80(3): 969-1018.

Clements B, Coady D, Gupta S. The Economics of Public Health Care Reform in Advanced and Emerging Economies. Washington: I. M. Fund; 2012.

Congressional Budget Office (CBO). The Long-Term Outlook for Health Care Spending. The Congress of the United States; November 2007.

Congressional Budget Office (CBO). Historical Budget Data-January 2012 Baseline. The Congress of the United States; January 2012a.

Congressional Budget Office (CBO). The 2012 Long-Term Budget Outlook. The Congress of the United States; June 2012b.

Currie J, Gahvari F. Transfers in Cash and In-Kind: Theory Meets the Data. JEL. 2008;46(2):333-83.

Cutler D. Your Money or Your Life: Strong Medicine for America's Healthcare System. New York: Oxford University Press; 2004.

Cutler DM, Reber SJ. Paying for health insurance: the trade-off between competition and adverse selection. QJE. 1998;113(2):433-66.

Cutler DM, Rosen A, Vigan S. The Value of Medical Spending in the United States, 1960-2000. New England Journal of Medicine. 2006;355: 920-927.

Doyle, J., Graves, J.A., Gruber, J., Kleiner, S., Do High-Cost Hospitals Deliver Better Care? Evidence from Ambulance Referral Patterns, NBER Working Paper No. 17936; March 2012.

Emanuel EJ, Fuchs V. Health Care Vouchers – A Proposal for Universal Coverage. N Engl J Med. 2005;352(12):1255-60.

Fendrick AM, Chernew ME. Value Based Insurance Design: Maintaining a Focus on Health in an Era of Cost Containment. Am J Manag Care. 2009;15(6):338-43.

Fonseca R, Michaud PC, Galama T, Kapteyn A, On the Rise of Health Spending and Longevity, IZA Working Paper No. 4622, December 2009.

Fullerton D. Reconciling Recent Estimates of the Marginal Welfare Costs of Taxation. American Economic Review 1991;81:302-8.

Gahvary, Firouz, On the Marginal Cost of Public Funds and the Optimal Provision of Public Goods. Journal of Public Economics 2006;90(6-7):1251-1262.

Getzen TE. Population Aging and the Growth of Health Expenditures. J Gerontol. 1992;17:S98-104.

Getzen TE. Health Care is an Individual Necessity and a National Luxury: Applying Multilevel Decision Models to the Analysis of Health Care Expenditures. Journal of Health Economics. 2000;19(2):259-70.

Gibson TB, Wang S, Kelly E, Brown C, Turner C, Frech-Tamas F, et al. A Value-Based Insurance Design Program at a Large Company Boosted Medication Adherence for Employees with Chronic Illnesses. Health Aff (Millwood). 2011;30(1):109-17.

Gruber, J. and E. Saez, The Elasticity of Taxable Income: Evidence and Implications, Journal of Public Economics 2002; 84(1):1-12.

Gruber, J. and K. Simon, Crowd-out 10 years later: Have Recent Public Insurance Expansions Crowded Out Private Health Insurance? Journal of Health Economics 2008; 27:201–217.

Hall RE, Jones CI. The Value of Life and the Rise in Health Spending. QJE. 2007;122(1):39-72.

Jacobs, Bas. The marginal cost of Public Funds is One. Working paper, Tilburg University. http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1711121.

Kaplow L. The Theory of Taxation and Public Economics. Princeton: Princeton University Press; 2008.

MedPAC. Health Care Spending and the Medicare Program; June 2012.

Murphy KM, Topel RH. The Value of Health and Longevity. JPE. 2006;114(5):871-904.

Nordhaus W. The Health of Nations: The Contributions of Improved Health to Living Standards. NBER Working Papers 8828; 2002.

OECD, Looking to 2060: Long-term global growth prospects, http://www.oecd.org/eco/outlook/2060%20policy%20paper%20FINAL.pdf; 2012.

Reinhardt U. A Fork in the Road for Health Care. The New York Times. 2012 May 25.

Ryan P. The Path to Prosperity: A Blueprint for American Renewal. Washington: Fiscal Year 2013 Budget Resolution; 2012.

Slemrod J and Yitzhaki S. Integrating Expenditure And Tax Decisions: The Marginal Cost Of Funds And The Marginal Benefit Of Projects. National Tax Journal. 2001;54(2):189-201.

Skinner JS. Causes and Consequences of Geographic Variation in Health Care. In: Pauly MV, McGuire TG, Barros PP, editors. Handbook of Health Economics. New York: North Holland; 2012. p. 45-94.

Skinner, J, and Staiger, D., "Technology Diffusion and Productivity Growth in Health Care," NBER Working Paper, April 2009.

Van de Ven, WPMM, Schut FT. Universal Mandatory Health Insurance In The Netherlands: A Model For The United States? Health Affairs 2008;27(3):771-781.

Appendix A: Proofs

Lemma: Assume (as in the text) that utility is calibrated to have a positive value of a life year at some level of consumption. Then as $y \to \infty$, $\frac{M^*}{y-M^*} > \frac{M \cdot \lambda'(M)}{\lambda(M)}$, the elasticity of $\lambda(.)$ w.r.t. M. Therefore, in the constant elasticity production function (where $\frac{M\lambda'(M)}{\lambda(M)} = \theta$), $\frac{M^*}{y} > \frac{\theta}{1+\theta}$.

Proof: Because of the utility calibration, there is a level of consumption C_{min} such that (for the L at which the value of a life-year is calibrated), $u(C_{min}, L) \geq 0$ and therefore, $u(C, L) > 0$ for all $C > C_{min}$. Now, rearranging the FOC for M in (2) generates:

$$\frac{M}{C - C_{min}} = \left(\frac{M \cdot \lambda'(M)}{\lambda(M)}\right) \cdot \frac{1}{C - C_{min}} \cdot \frac{u(C, L)}{u_C(C, L)} \tag{13}$$

Now, by concavity of utility in consumption, $u(C, L) > u(C_{min}, L) + (C - C_{min}) \cdot u_C(C, L) \geq (C - C_{min}) \cdot u_C(C, L)$ for all $C > C_{min}$, and therefore, $\frac{1}{C-C_{min}} \frac{u(C,L)}{u_C(C,L)} > 1$. So equation (13) implies that for all $C > C_{min}$, $\frac{M}{C-C_{min}} > \frac{M \cdot \lambda'(M)}{\lambda(M)}$. But as $y \to \infty$, $\frac{M}{C-C_{min}} \to \frac{M}{C} = \frac{M}{y-M}$. Therefore, for all y sufficiently large, $\frac{M}{y-M} > \frac{M \cdot \lambda'(M)}{\lambda(M)}$, which establishes the result. ■

Proposition: Suppose that $\frac{M^*}{y} \to 0$ as $y \to \infty$. Then $\lambda(M)$ must be bounded above.

Proof: If $M/y \to 0$ as $y \to \infty$, then also $\frac{M}{C-C_{min}} = \frac{M}{y-M-C_{min}} \to 0$. Therefore, the right-side of equation (13) in the Lemma must also approach zero as $y \to \infty$. Because the term $\frac{u(C,L)}{(C-C_{min})u_C(C,L)}$ is increasing in C, the right-side of (13) can only approach zero if the health production function elasticity approaches zero as M grows large. So $\frac{d \log \lambda(M)}{d \log M} \equiv \varepsilon(M) \to 0$ as $M \to \infty$. I claim that this fact implies that $\lambda(M)$ is bounded. To see this, note that the differential equation for the elasticity implies:

$$\begin{aligned}
\log \lambda(M) &= \int_0^M \left(\frac{\varepsilon\left(\tilde{M}\right)}{\tilde{M}}\right) d\tilde{M} \\
&\approx \sum_{n=1}^{M/\Delta} \frac{\varepsilon(M_n)}{M_n} \Delta \qquad \text{for a small } \Delta, \text{ with } M_n = n\Delta \\
&= \sum_{n=1}^{M/\Delta} \frac{\varepsilon(n\Delta)}{n}
\end{aligned}$$

where the second line is by definition of a Riemann sum. But because $\varepsilon(n\Delta) \to 0$, the summand declines to zero faster than $1/n$, and therefore the series converges. Formally, there is some $\delta > 0$ and some N such that $\frac{\varepsilon(n\Delta)}{n} \leq n^{-(1+\delta)}$ for all $n \geq N$, and because $\sum_{n=1}^{\infty} n^{-(1+\delta)}$ converges, this series converges by the bounding test. Therefore, $\log \lambda(M)$ approaches a finite number as $M \to \infty$, which establishes the result. ■

Appendix B: Specifying Smooth Tax Function

We use a smooth approximation to the actual tax schedule in order to simplify computation and because we are not interested in bunching or other properties that arise only with a kinked tax schedule. We first define an "actual" marginal tax rate schedule, based on an approximation to tax laws in 2007. Total marginal tax rates equal the sum of (1) federal income tax rates for a single individual (since our model is at the individual level) in 2007, (2) Social Security and Medicare payroll taxes, and (3) an assumed constant state tax rate of 4%. (We do not model marginal tax rates arising from phase-outs of benefits, credits, or deductions.) For payroll tax rates, we assume half of the Social Security payroll tax is benefit-linked and therefore not a true marginal tax. So we use a payroll rate of 9.1% (= 6.2% Social Security + 2.9% Medicare) up to the 2007 taxable maximum of $97,500 and a rate of 2.9% above this. The resulting schedule for marginal and average tax rates are shown in the black lines in Appendix Figure 1.

We then generate a smoothed schedule by fitting a second-order polynomial in log-income to actual marginal tax rates. Specifically, we run the regression $MTR_y = \beta_0 + \beta_1 \log(y + 2000) + \beta_2 \log(y + 2000)^2 + \varepsilon_y$ for a large number of values of y between $0 and $5 million (the top of our income distribution), imposing the constraints that predicted marginal tax rates are positive and increasing at all incomes in this range. (Here, we use $\log(y + 2000)$ rather than $\log(y)$ because in trials, this significantly improves the fit by avoiding issues with log increasing very rapidly near zero.) The resulting estimates (which are $\beta_0 = -0.542$, $\beta_1 = 0.127$, $\beta_2 = -0.0041$) are shown in the red dashed lines in Appendix Figure 1. This simple model clearly fits the actual tax schedule quite well in the range shown (up to $600,000). Beyond this, the model's marginal tax rates continue increasing slightly (while actual tax rates are flat), but even at $5 million, the difference is small (1.7% points).

Thus, we view this smooth model as a good approximation to the progressive U.S. tax schedule and use it in our modeling. In the optimal policy problem, the government chooses a single parameter $\tau \in [0, 1]$ that scales this entire marginal tax rate distribution, multiplying all of the β-coefficients. We normalize τ so it equals the marginal tax rate at the top of our income distribution ($5 million).[27] To generate the total tax rate, we integrate this marginal tax rate function (which conveniently has an analytic solution), setting the constant so that $T(0) = 0$.

[27] In the original smoothed function, $\hat{MTR}(\$5\,million) = 0.436$. So what we are doing in practice is multiplying the estimated β's times $\tau/0.436$.

Appendix C: Theoretical Propositions

Implication 1: *Assuming the redistributionary motive is not too small and higher income is not correlated with worse health, optimal policy features a public health insurance benefit that is binding for some share of the population.*

To see this, first assume health status is homogenous at σ, public insurance works as a floor plan, and private insurance is actuarially fair (so everyone fully insures). Suppose that absent public insurance, the minimum private health insurance purchase in the population is I_{min}. A public insurance benefit of I_{min} is therefore equivalent to a cash transfer of its actuarial equivalent, σI_{min}. Assuming that $R^*(I_{pub} = 0) > \sigma I_{min}$, then a public benefit of up to I_{min} can be fully offset by reduced cash transfers. Suppose that $I_{pub} = I_{min}$ and that cash redistribution has been set either optimally or too low, so that $\frac{dSW}{dR} \geq 0$. We show that under these conditions $\frac{dSW}{dI_{pub}} > \sigma \frac{dSW}{dR}$. The formula for the latter is:

$$\sigma \frac{dSW}{dR} = \sum_i \sigma \frac{dV_i}{dR} + \sum_i \frac{dM_i^*}{\frac{1}{\sigma} dR} \cdot \lambda'(M_i) \cdot \hat{u} - \alpha \sum_i \left(\sigma - \tau w_i \frac{dL_i^*}{\frac{1}{\sigma} dR} \right) \tag{14}$$

and the formula for $\frac{dSW}{dI_{pub}}$ in this case is:

$$\frac{dSW}{dI_{pub}} = \sum_i \frac{dV_i}{dc_{is}} + \sum_i \frac{dM_i^*}{dI_{pub}} \cdot \lambda'(M_i) \cdot \hat{u} - \alpha \cdot \sum_i \left(\sigma - \tau w_i \frac{dL_i^*}{dI_{pub}} \right) \tag{15}$$

The two policies are identical for everyone with $I_i > I_{min}$. So consider just the group with $I_i = I_{min}$. The first terms in these expressions are equal, since people are fully insured. The second term is strictly larger in (15), since increasing public insurance binds these people to obtain \$1 more health insurance, whereas the actuarially equivalent cash transfer is partly spent on other goods. Finally, as argued in the previous section, $\frac{dL_i^*}{dI_{pub}} > \frac{dL_i^*}{\frac{1}{\sigma} dR}$ because cash transfers reduce labor supply more in-kind transfers because of the income effect. Thus, $\frac{dSW}{dI_{pub}} > \sigma \frac{dSW}{dR} \geq 0$, so optimal public insurance is set at a level $I_{pub} > I_{min}$.

Now, consider weakening the assumptions above. Suppose that health is not homogenous. Then *non*-binding in-kind transfers (i.e. for $I_{pub} < I_{min}$) are equivalent to a non-linear cash transfer that has either progressive or regressive incidence. As long as higher incomes are not correlated with worse health, their incidence will be progressive, and in-kind benefits will be *more* attractive than in the homogenous model. Therefore, the result still holds. Next, suppose that private insurance is not actuarially fair, so some people partially insure. For these people, the marginal utility of consumption in the sick state is higher than in the healthy state, so $\frac{dV_i}{dc_{is}} > \sigma \frac{dV_i}{dR}$, preserving the result. Finally, we show below that the optimal universal insurance benefit is always higher than with a floor benefit.

Implication 2: *As tax rates rise exogenously (e.g., to fund additional extra government spending E), the optimal public health insurance benefit decreases.*

NOTE: We should be able to prove this from the comparative static $\frac{\partial I_{pub}}{\partial \tau}$, derived from differentiating the FOC (??).

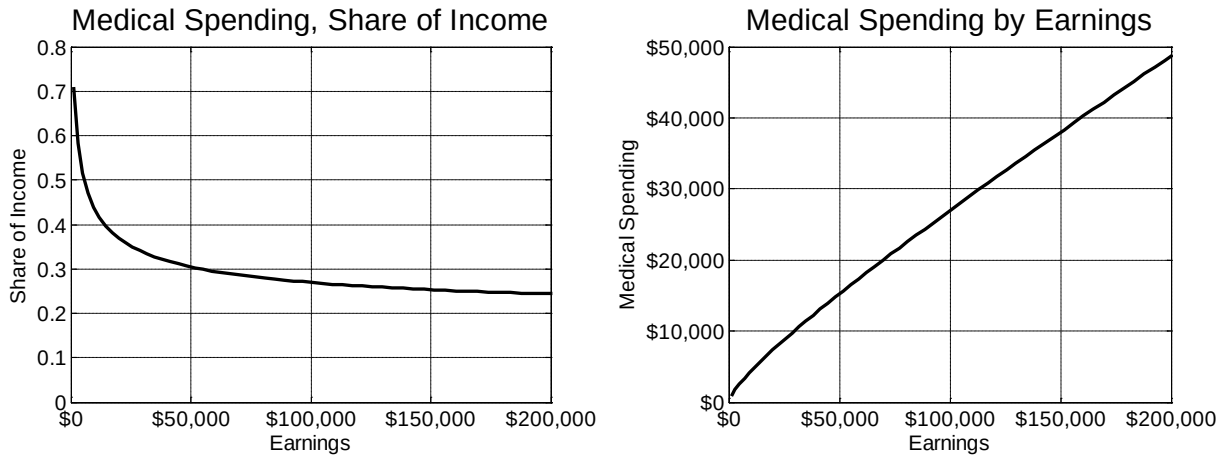
The following fact may be helpful in this proof: Consider the formula for α, which is defined from the government's FOC for τ:

$$\alpha = \left(1 - \frac{\tau}{1-\tau}\varepsilon_{\overline{wL},1-\tau}\right)^{-1} \frac{1}{N}\sum_i \left(\frac{dV_i}{dR} \cdot \frac{w_i L_i}{\overline{wL}} + \frac{\lambda'(M_i)}{\overline{wL}} \frac{dM_i}{d(1-\tau)} \cdot \hat{u}\right)$$

This formula is increasing in the tax rate both explicitly (from the first term) and because higher taxes cause people to cut back on consumption and medical care, raising people's marginal utilities of consumption, $\frac{dV_i}{dR}$, and marginal productivities of medical spending, $\lambda'(M_i)$.

FIGURE 1

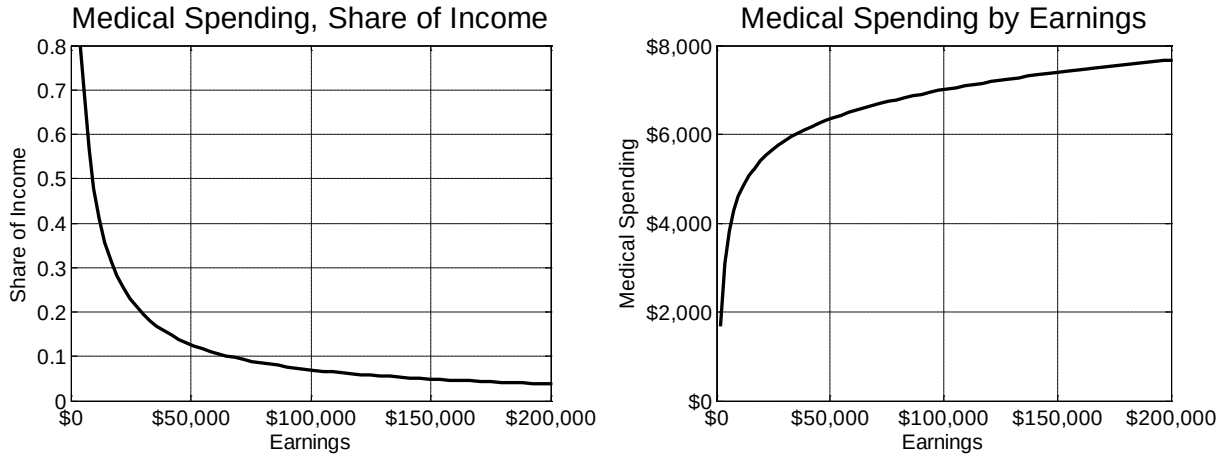
Health Spending with Constant Elasticity Health Production Function



NOTES: These display optimal private medical spending as a share of income (left) and in levels (right) by earnings for the constant elasticity health production function, $\lambda(M) = Const * M^{\theta}$, as discussed in the text. There are no taxes or transfers so earnings equal income. The coefficient of relative risk aversion is set as 0.5, and the production function elasticity θ is set as 0.15, near the middle of the estimates across ages of Hall and Jones (2007), Figure III. The value of *Const* does not affect the health spending estimates so is set arbitrarily.

FIGURE 2

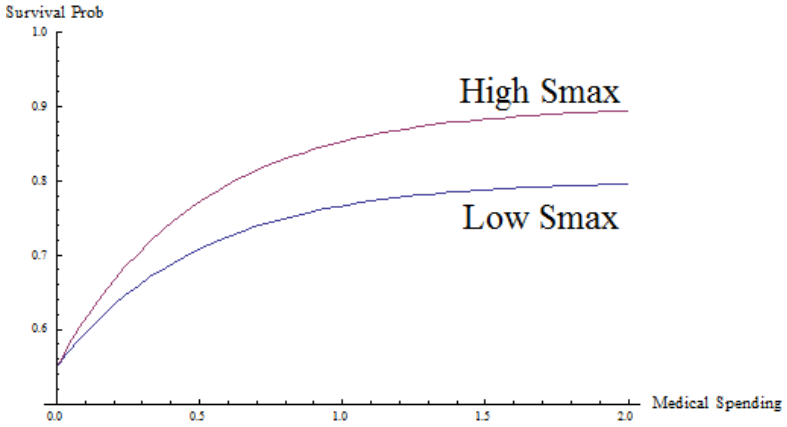
Health Spending with Max Survival Production Function



NOTES: These display optimal private medical spending as a share of income (left) and in levels (right) by earnings for the max survival production function, as discussed in the text. There are no taxes or transfers so earnings equal income. The coefficient of relative risk aversion is set as 0.5. The production function parameters are set to the values used for the year 2010 in our simulations: $\sigma = 0.3,\ \alpha = 0.000275,\ s_0 = 0.1,\ s_{max} = 0.771.$

FIGURE 3

Two Forms of Technological Progress



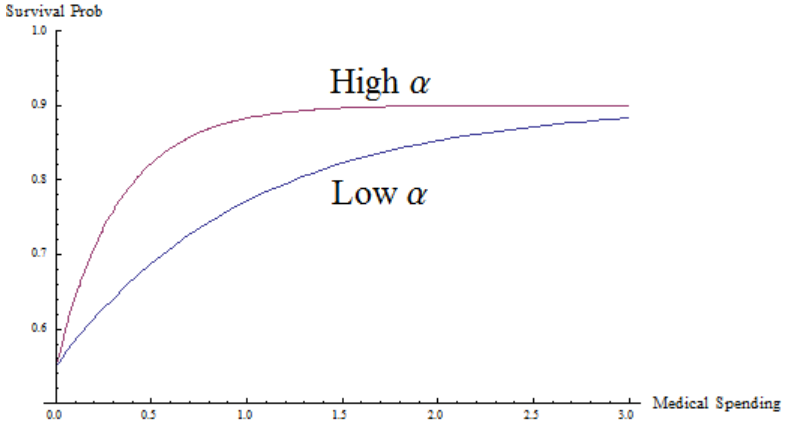


FIGURE 4

Health Care Spending Growth with Private Insurance Only

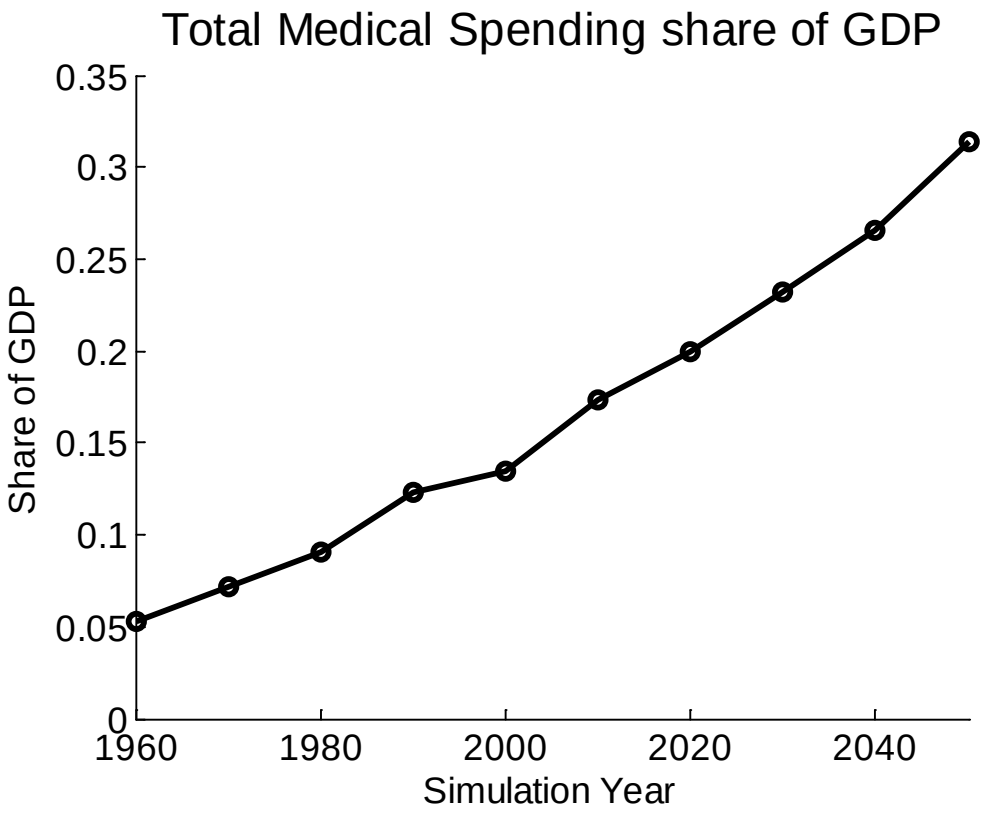


FIGURE 5

Cross-Sectional Health Care Spending Distribution with Different Policies

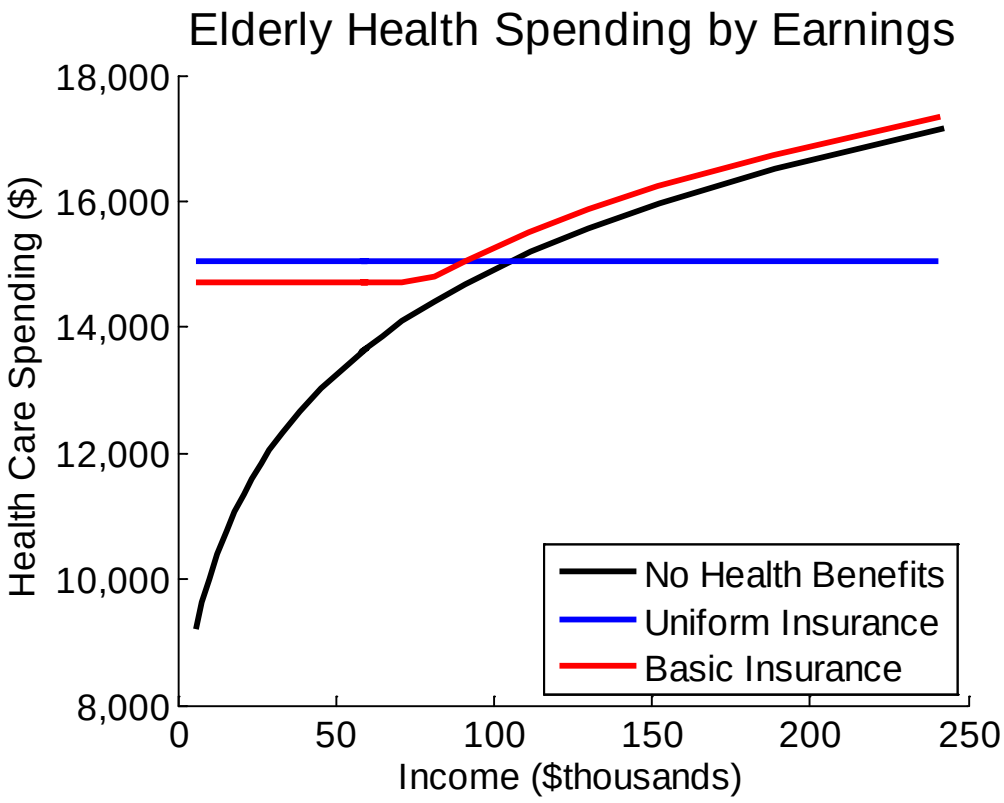


FIGURE 6

International Comparisons

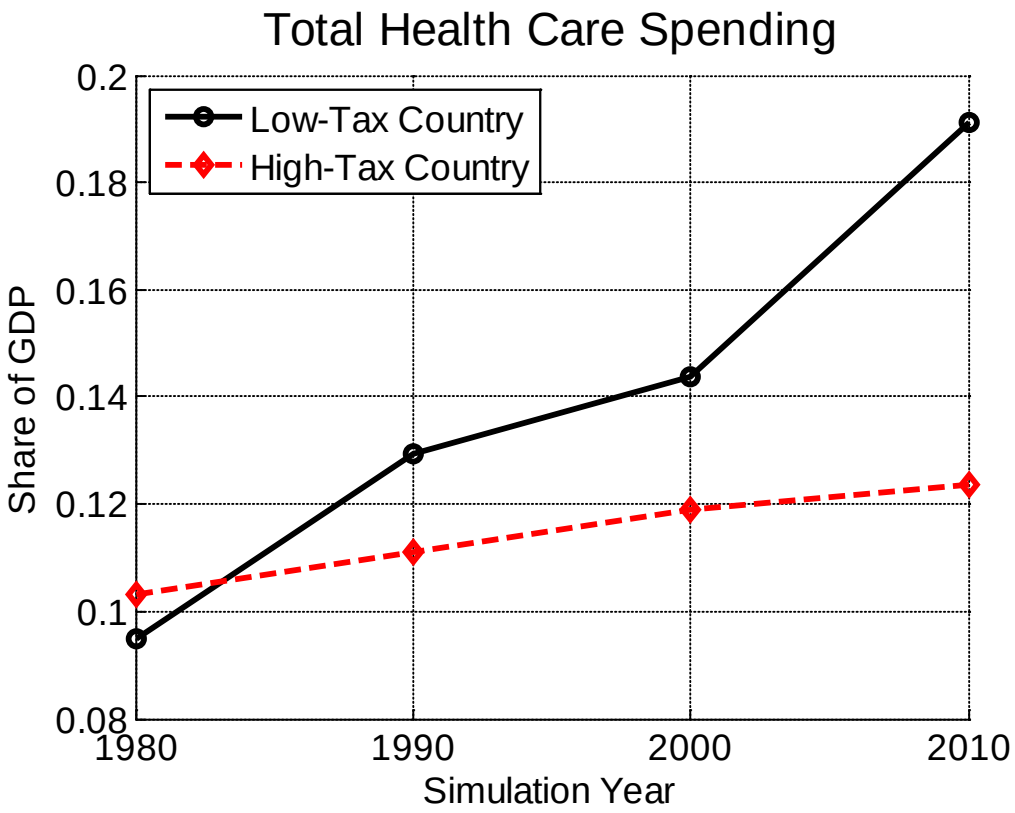


FIGURE 7

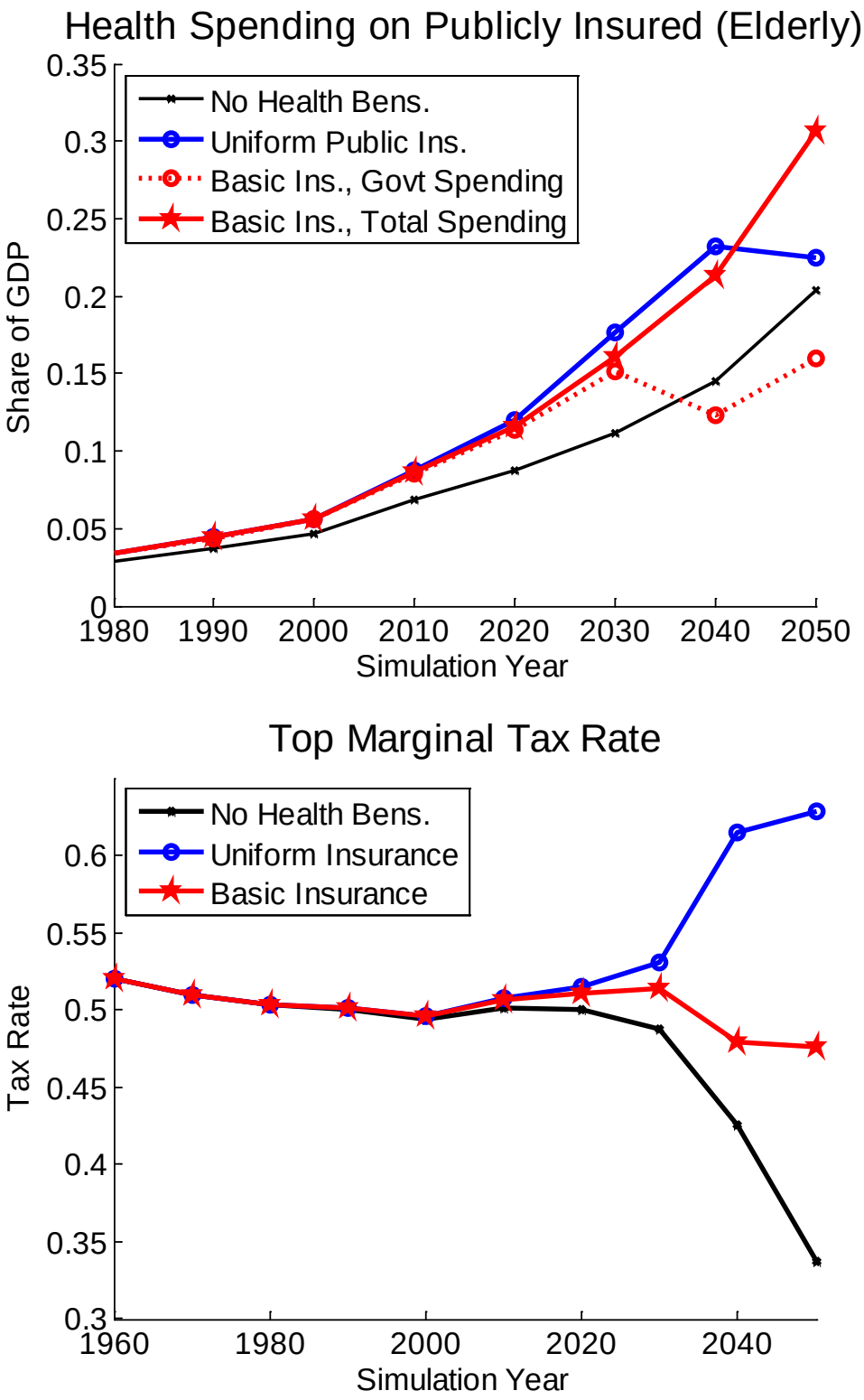

Health Spending on Publicly Insured (Elderly)
No Health Bens.
Uniform Public Ins.
Basic Ins., Govt Spending
Basic Ins., Total Spending
Share of GDP
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
1980
1990
2000
2010
2020
2030
2040
2050
Simulation Year
Top Marginal Tax Rate
No Health Bens.
Uniform Insurance
Basic Insurance
Tax Rate
0.6
0.55
0.5
0.45
0.4
0.35
0.3
1960
1980
2000
2020
2040
Simulation Year

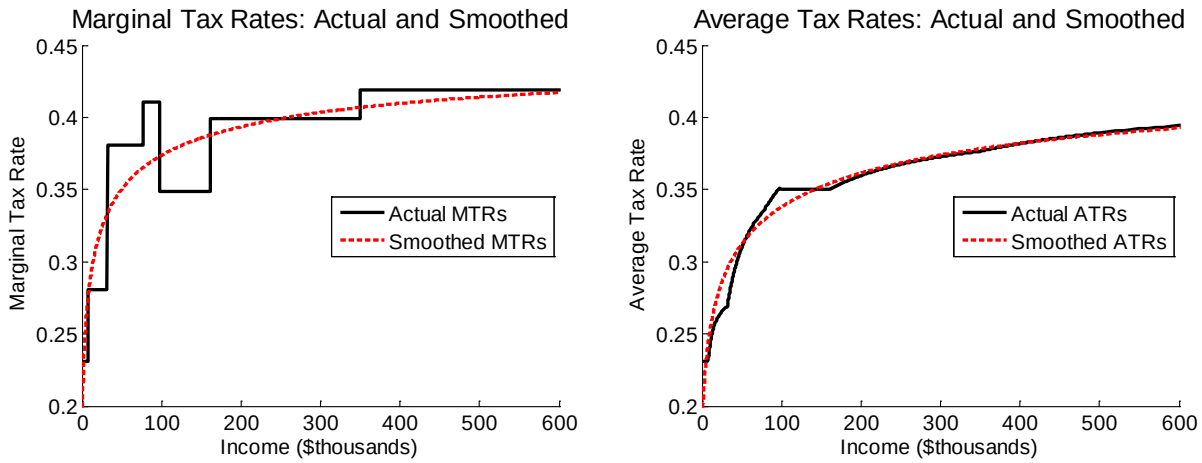
TABLE 1

Moments for Parameter Calibration

Simulation Year	**1960**	**1970**	**1980**	**1990**	**2000**	**2010**	**2020**	**2030**	**2040**	**2050**
Real GDP Per Capita	$17,368	$23,089	$28,434	$35,612	$44,081	$46,844	$54,364	$63,092	$73,221	$84,976
Total Health Spending / GDP	5.2%	7.2%	9.2%	12.5%	13.8%	17.9%	21.2%	25.1%	29.6%	35.0%
Public Health Spending / GDP	0.9%	2.2%	3.3%	4.3%	5.4%	7.9%	10.4%	13.3%	16.8%	21.1%
Private Health Spending / GDP	4.3%	5.0%	5.9%	8.2%	8.4%	10.0%	10.8%	11.8%	12.8%	13.9%
Life expectancy at birth	70.0	71.0	73.7	75.4	76.7	77.7	78.7	79.6	80.5	81.3
Life expectancy at 65	14.5	15.2	16.3	17.2	17.5	17.9	18.5	19.1	19.7	20.2

APPENDIX FIGURE 1:

Actual and Smoothed Tax Schedules



NOTE: These graphs show actual and our smoothed function for marginal tax rates (left figure) and average tax rates (right figure) by income level. The actual tax rates shown equal the sum of (1) federal income tax rates for a single individual (since our model is at the individual level) in 2007, (2) payroll tax rates in 2007, and (3) an assumed state tax rate of 4%. We do not model marginal tax rates arising from phase-outs of benefits, credits, or deductions. For payroll tax rates, we assume half of the Social Security payroll tax is benefit-linked and therefore not a true marginal tax. So the payroll tax rates equal the 2.9% Medicare rate at all incomes plus a 6.2% Social Security tax (half of the full 12.4%) on incomes below the taxable maximum of $97,500. The smoothed tax schedules are constructed by fitting a polynomial in log-income to the marginal tax rate schedule, as described in greater detail in the Appendix.

APPENDIX TABLE 1

Annual Earnings and Unearned Income at Model Discretization Points

Earnings	Unearned Income	Population Share
$1,872	$3,791	5.001%
$4,973	$2,930	5.001%
$8,042	$2,148	5.002%
$10,847	$1,934	5.015%
$13,660	$1,825	5.001%
$16,378	$1,922	4.987%
$19,037	$1,855	4.998%
$21,729	$1,697	5.006%
$24,551	$1,610	4.989%
$27,327	$1,668	5.011%
$30,955	$1,742	7.412%
$36,401	$1,965	9.440%
$42,627	$2,308	7.161%
$47,914	$10,809	4.043%
$52,751	$4,300	4.546%
$58,774	$4,618	3.702%
$65,644	$4,290	3.196%
$73,630	$6,033	2.549%
$83,200	$6,295	2.116%
$95,003	$14,339	1.547%
$110,300	$17,080	1.235%
$129,718	$19,483	0.982%
$157,747	$27,323	0.729%
$197,393	$40,015	0.530%
$257,804	$66,473	0.369%
$361,448	$120,211	0.236%
$581,890	$282,717	0.126%
$1,239,043	$620,854	0.057%
$4,989,363	$1,807,997	0.015%

NOTE: This table shows the values of annual earnings and unearned income at the discretization points used for our income distribution. Earnings are used to set wages in our model (so that wage times optimal labor equals the given value of earnings), and unearned income enters directly into the model as exogenous income.