Reducing medical spending of the publicly insured: the case for a cash-out option*

Svetlana Pashchenko† Ponpoje Porapakkarm‡

October 28, 2018

Abstract

Individuals' medical spending has both necessary and discretionary components which are not, however, separately observable. This paper studies ways to improve upon existing public health insurance policies by using a framework where both the discretionary and necessary components of medical spending are explicitly modeled. First, using a simple theoretical framework the paper shows that the key to reducing discretionary medical spending is to introduce a trade-off between non-medical and medical consumption. Next, using a rich quantitative life-cycle model the paper shows that this trade-off can be successfully implemented by introducing an option to substitute public health insurance with cash transfers.

Keywords: medical spending, health insurance, optimal taxation, life-cycle models, ex-post moral hazard
JEL Classification Codes: D52, D91, E21, H53, I13, I18

*We thank Charles Brendon, Mariacristina De Nardi, Eric French, Mike Golosov, Oleg Itshoki, John Kennan, Jean-Marie Lozachmeur, Yoki Okawa, Nicola Pavoni, Matthew Shapiro, Ali Shourideh, Motohiro Yogo, two anonymous referees and all seminar participants at the Asian Meeting of the Econometric Society in Hong Kong, CIREQ Macro Conference, IFS, Keio-GRIPS Macroeconomics and Policy Workshop, Midwest Macro Meeting in Kansas City, QSPS Summer Workshop, and World Congress in Montreal for their comments and suggestions. All errors are our own.

†University of Georgia; Email: svetlana@uga.edu

‡National Graduate Institute for Policy Studies (GRIPS); Email: p-porapakkarm@grips.ac.jp

1 Introduction

It is well-known that individuals' spending on medical care is not solely comprised of necessary expenses but is to some extent discretionary. What this means for health insurance is that the insurable events (medical expenses) are not completely exogenous: although an illness itself is a risk, individuals can control how much to spend on a cure. In the first-best world, individuals should be fully insured against necessary medical expenses (which represent a risk) and fully responsible for discretionary medical spending (which represents consumer choice). However, the composition of medical spending in terms of necessary and discretionary parts is not (perfectly) observable. As pointed out by Arrow (1963), this makes full insurance no longer optimal: it protects individuals against financial loss but creates excessive discretionary medical consumption. In this paper, we construct a framework where both discretionary and necessary components of medical spending are explicitly modeled in order to understand how we can improve upon existing public health insurance policies.

Our analysis has two parts: theoretical and quantitative. Our theoretical analysis builds on the optimal taxation literature pioneered by Mirrlees (1971). In the environment considered by this literature, a social planner who does not have information about individual types has to choose incentive-compatible policies: i.e., such policies that individuals do not have incentives to lie about their type.

In a similar spirit, we construct a theoretical model in which individuals differ in their (unobservable) medical needs. Medical need requires some unavoidable medical consumption, but individuals may choose extra (discretionary) medical care on top of that because it increases their utility. In this environment, a social planner allocates bundles of medical and non-medical consumption conditional on an individual's self-reported medical needs. Because an individual who reports having a high medical need is allocated higher medical consumption and medical consumption is valuable, everyone has an incentive to report his medical need as being high. We show that to correct incentives, the planner has to offer high medical consumption in combination with low non-medical consumption. Moreover, the medical consumption of individuals who report having the lowest medical need should be undistorted.

We quantitatively evaluate the effect of this type of policy with an application to non-elderly Medicaid beneficiaries. Medicaid is a means-tested public health insurance program in the US that charges its beneficiaries no premiums and provides them with nearly free health care. Medicaid represents a challenging case for studying the trade-off between moral hazard and risk-protection. On the one hand, the publicly insured face almost zero price for their medical care and this can potentially increase their discretionary medical consumption. In the data, Medicaid beneficiaries' average medical spending per person is substantially

higher than that of either the privately insured or the uninsured (Figure 1). This may be partially due to the selection of unhealthy people into Medicaid (see Pashchenko and Porapakkarm, 2017 for a discussion on the composition of Medicaid beneficiaries). However, there is evidence that people who enroll in Medicaid increase their medical care utilization even though this does not improve their subsequent health outcomes (Baicker et al., 2013). On the other hand, policies that address the moral hazard problem can substantially increase risk exposure of Medicaid beneficiaries and, consequently, be detrimental to the welfare since most beneficiaries are relatively poor.

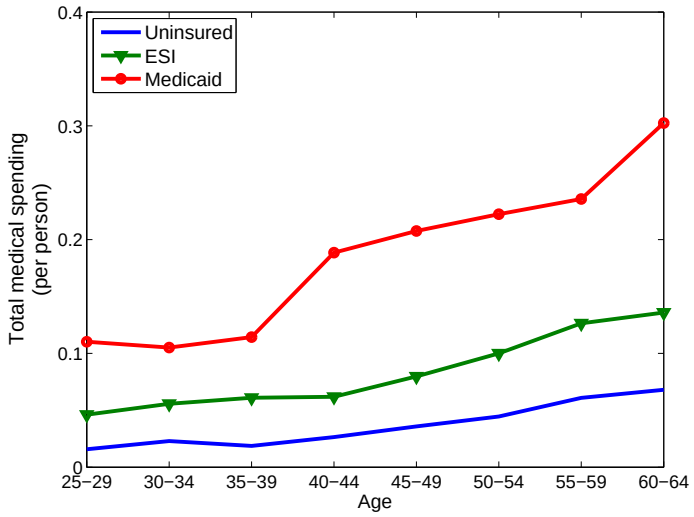

Figure 1: Total medical expenses by insurance status (normalized by average income). Source: Medical Expenditure Panel Survey (MEPS), 1999-2012. Sample: heads of households aged 25 to 64 years old. (More details on sample selection are available in Section 5.1). ESI stands for "employer-sponsored insurance".

For our quantitative analysis, we construct a rich structural life-cycle model that reflects key institutional aspects of the US health insurance system and captures the selection of unhealthy people into Medicaid. More specifically, individuals who are heterogeneous in income, health and medical need shocks choose their labor supply and how to allocate their resources between non-medical consumption, medical consumption and savings. Medical consumption is composed of non-discretionary spending (generated by medical need) and discretionary spending.

The model features both private and public insurance. Private insurance is available from employers for workers with relatively high-income. Low-income individuals can enroll in Medicaid (if eligible) or remain uninsured (self-insure through savings). However, Medicaid imperfectly targets low-resource individuals: even though there is an income test and an asset test, a high-productivity individual can still enroll by decreasing his labor supply so that he passes the income test. This is important for capturing the potential changes in the

composition of Medicaid beneficiaries in response to changes in the program design. We also include a means-tested cash assistance program to capture other parts of the US safety net.

We calibrate/estimate the model using the Medical Expenditure Panel Survey dataset (MEPS) by targeting many important features of the data. In particular, the model reproduces the following life-cycle profiles for each health group: the mean, median and variance of medical expenses, the fraction of people with zero medical expenses, average labor income, employment and insurance take-up rates. We identify discretionary versus non-discretionary medical spending by matching the difference in medical spending profiles between the uninsured and the privately insured, while controlling for the different composition of these two groups in terms of income and health.

We first use our model to construct the full information benchmark in which the government can observe the medical need of each Medicaid beneficiary. In this case, it is possible to fully insure individuals without creating excessive discretionary medical consumption. To do this, the government covers 100% of non-discretionary spending and the remaining Medicaid budget is allocated as lump-sum transfers. Under this arrangement Medicaid enrollees are fully protected against medical need shocks while facing full price for their discretionary medical consumption.

Next, we conduct an extensive policy analysis where the success of each policy is measured by benchmarking it against the full information case. Importantly, all policies that we consider are revenue-neutral, i.e., the size of the welfare budget is unchanged. We first show that just an increase in coinsurance is an ineffective tool to achieve the full information benchmark. Following the intuition developed in our theoretical analysis, we then show that an outcome close to the full information case can be achieved by introducing a trade-off between medical and non-medical consumption. This can be done by giving Medicaid-eligible individuals a choice between the following two options: i) regular or traditional (in-kind) Medicaid benefits, i.e., health insurance; ii) lump-sum cash transfers, the amount of which is adjusted to preserve revenue-neutrality. This trade-off induces individuals with low medical need to self-select into the cash subprogram of Medicaid, where they face full price for their medical care, thus substantially decreasing their discretionary medical spending.

It is important to note that several states have experimented with partial cashing-out of benefits in their Medicaid programs (Cash and Counseling Demonstrations). In particular, Arkansas, Florida and New Jersey conducted randomized experiments in which in-kind provision of home health care benefits was partially converted into cash transfers for some beneficiaries. Lieber and Lockwood (2017) find that those participants who were randomized to receive in-kind benefits had significantly higher consumption of formal home health care, which is consistent with our findings.

The rest of the paper is organized as follows. Section 2 reviews the literature. Section 3 details our theoretical analysis. Section 4 describes our quantitative model. Section 5 describes the data and estimation. Section 6 compares the performance of the model with the data. Section 7 discusses the results, and Section 8 concludes.

2 Related literature

Our paper is related to several strands of literature. The first strand includes studies on ex post moral hazard in the use of medical care.[1] Two major questions in this field are i) how quantitatively important is moral hazard in health insurance, and ii) how to design an insurance contract that mitigates the incentive problem.

A common approach to measuring the prevalence of moral hazard is to use the price elasticity of medical consumption.[2] Although the exact measure of this price elasticity varies by study, a common finding is that a change in cost sharing affects the demand for medical care (for a detailed review of related studies see Zweifel and Manning, 2000). The "gold standard" in this literature is the RAND Health Insurance experiment, where the reported price elasticity was of the order of -0.20 (Keeler and Rolph, 1988, Manning et al., 1987, see also Aron-Dine et al., 2013 for a recent reevaluation).

On the normative side, an important implication of the existence of ex post moral hazard is that full insurance is not optimal. This was first pointed out by Arrow (1963) and later formalized by Spence and Zeckhauser (1971). The latter two authors also recognized the similarity between the optimal health insurance problem and the optimal taxation problem: in both cases, a principal (insurance company or the government) aims to provide insurance to agents (against health shocks or income) while realizing that insurance affects incentives (either to use medical care or to work). A number of studies have examined the trade-off between moral hazard and risk protection in health insurance in the context of linear co-insurance functions (e.g., the theoretical study of Besley, 1998, or the quantitative studies of Buchanan et al., 1991, Feldman and Dowd, 1991, Manning and Marquis, 1996). Blomqvist (1997) formulates a more general problem of optimal health insurance contract that allows for an arbitrary non-linear relationship between the coinsurance rate and medical spending. In his numerical simulations, he shows that the coinsurance rate should vary considerably with the level of medical spending.

[1] The term ex post moral hazard refers to a situation where an insured individual chooses how much to spend on medical care and his insurance company cannot observe which part of this spending is necessary versus discretionary.

[2] Vera-Hernandez (2003) provides an alternative measure based on the correlation between medical spending observed by an insurance company and unobserved health shocks.

Our paper is related to both the positive and normative questions outlined above. First, we use the Mirrlees approach to characterize optimal health insurance policies in an environment with both discretionary and non-discretionary spending and then quantitatively evaluate the impact of these policies. Second, using the calibrated life-cycle model we construct the full information case with no moral hazard. Comparing this case with the baseline economy provides another angle to measure existing moral hazard.

Methodologically, we relate to life-cycle structural models featuring health and medical expenses uncertainty. In these studies, two approaches exist to model medical spending. The first approach is to treat medical spending as an exogenous shock, i.e., assume that all medical spending is non-discretionary or necessary (Capatina, 2015, De Nardi et al., 2010, French, 2005, Nakajima and Telyukova, 2012, Pashchenko and Porapakkarm, 2013). The second approach is to assume that all medical spending is discretionary. A common modeling framework for the latter approach is to assume medical spending represents investments in health (Fonseca et al., 2009, Ozkan, 2013, Scholz and Seshadri, 2010).[3] Our paper bridges the gap between these two approaches by explicitly modeling medical spending as a combination of discretionary and non-discretionary components, and estimating the relative importance of these two components using our structural model.[4]

It is also important to note that our study allows for endogenous medical spending, endogenous labor supply and endogenous health insurance purchase decisions. With the exception of Jung and Tran (2016) and Jung et al. (2017), to the best of our knowledge, no structural life-cycle model has all of these features. This is, however, an important methodological development for two reasons. First, in the US institutional framework, health insurance and employment decisions are linked, e.g., individuals typically cannot buy employer-sponsored health insurance if they are not working. Second, many working-age individuals switch between different types of health insurance (private and public) over their life course or spend some time being uninsured. Policies that aim to decrease distortions in medical care consumption can also affect labor supply and insurance decisions, and it is important to consider a framework where these decisions are endogenous.

[3]One exception is De Nardi et al. (2016), who assume that medical spending affects utility but not future health and is entirely discretionary while the marginal utility from medical consumption is exogenous and stochastic.

[4]A somewhat similar approach is implemented by Ameriks et al. (2017) who model spending on long-term care as consisting of necessary and discretionary components.

3 Theoretical model

Consider the following static model where individuals differ in their medical need η_i, $i = \{L, H\}$, $\eta_L < \eta_H$. The measure of individuals of type L is π and of type H is $1 - \pi$. Individuals derive utility from non-medical consumption c_i and discretionary medical consumption defined as the difference between total medical consumption m_i and medical need: $m_i - \eta_i$. We denote utility from non-medical consumption as $u(c)$ and from discretionary medical consumption as $v(m - \eta)$. We assume that both $u(\cdot)$ and $v(\cdot)$ are strictly increasing and strictly concave: $u'(\cdot) > 0$, $u''(\cdot) < 0$, $v'(\cdot) > 0$, $v''(\cdot) < 0$. In addition, $v(m-\eta) \longrightarrow -\infty$ if $m < \eta$, i.e., total medical consumption cannot be lower than medical need. A social planner has a fixed amount of resources B that he allocates to maximize aggregate welfare. We assume that $\pi\eta_L + (1 - \pi)\eta_H < B < \eta_H$, i.e., B is enough to cover medical need of all individuals but not enough to provide all individuals medical consumption equal to η_H.

Full information case We start by considering the full information case where the social planner observes the type of each individual. The social planner's problem can be written as follows:

$$\max_{\{c_i, m_i\}_{i=L,H}} \pi \left[u(c_L) + v(m_L - \eta_L)\right] + (1 - \pi) \left[u(c_H) + v(m_H - \eta_H)\right] \tag{1}$$

s.t.

$$\pi \left[c_L + m_L\right] + (1 - \pi) \left[c_H + m_H\right] = B \tag{2}$$

Denoting the Lagrange multiplier on the resource constraint Eq.(2) as λ, we can write the Lagrangian as follows:

$$\begin{aligned}\mathcal{L} = \pi \left[u(c_L) + v(m_L - \eta_L)\right] + (1 - \pi) \left[u(c_H) + v(m_H - \eta_H)\right] \; + \\ \lambda\big(B - \pi \left[c_L + m_L\right] - (1 - \pi) \left[c_H + m_H\right]\big)\end{aligned}$$

This results in the following first-order conditions:

$$u'(c_L) = v'(m_L - \eta_L) = \lambda$$

$$u'(c_H) = v'(m_H - \eta_H) = \lambda$$

From these conditions, it follows that $c_L = c_H$ and $(m_L - \eta_L) = (m_H - \eta_H)$, i.e., the planner equalizes non-medical and discretionary medical consumption of individuals.

Asymmetric information case Next, we assume that the social planner cannot observe the medical need of an individual. Instead, he allocates non-medical/medical consumption bundles based on individuals' self-reported types. To ensure that individuals do not lie about their types, we need to add incentive compatibility constraints (ICC) to the social planner's problem. Note that the H-type will never choose to lie about his type.[5] Because of this we have only one ICC for the L-type:

$$u(c_L) + v(m_L - \eta_L) \geq u(c_H) + v(m_H - \eta_L) \tag{3}$$

The social planner in the asymmetric information environment solves Problem (1) subject to the constraints (2) and (3). Denoting the Lagrange multiplier on Eq.(3) as μ, we can write the Lagrangian as follows:

$$\begin{aligned}\mathcal{L} = {} & \pi\left[u(c_L) + v(m_L - \eta_L)\right] + (1-\pi)\left[u(c_H) + v(m_H - \eta_H)\right] \; + \\ & \lambda(B - \pi\left[c_L + m_L\right] - (1-\pi)\left[c_H + m_H\right]) \; + \\ & \mu(u(c_L) + v(m_L - \eta_L) - u(c_H) - v(m_H - \eta_L))\end{aligned}$$

This results in the following first-order conditions:

$$u'(c_L) = \frac{\pi\lambda}{\pi + \mu} \tag{4}$$

$$u'(c_H) = \frac{(1-\pi)\lambda}{1 - \pi - \mu} \tag{5}$$

$$(\pi + \mu)v'(m_L - \eta_L) = \pi\lambda \tag{6}$$

$$(1-\pi)v'(m_H - \eta_H) - \mu v'(m_H - \eta_L) = (1-\pi)\lambda \tag{7}$$

Denoting the optimal allocation as $(c_L^*, c_H^*, m_L^*, m_H^*)$, the solution to the social planner's problem can be characterized as follows:

$$v'(m_L^* - \eta_L) = u'(c_L^*) \tag{8}$$

$$v'(m_H^* - \eta_H) = u'(c_H^*)\left(\frac{u'(c_L^*) + \frac{v'(m_H^* - \eta_L)}{u'(c_H^*)}\pi(u'(c_H^*) - u'(c_L^*))}{u'(c_L^*) + \pi(u'(c_H^*) - u'(c_L^*))}\right) \tag{9}$$

[5]This happens because if the H-type does lie, he receives medical consumption m_L. However, due to the assumption that $B < \eta_H$ it must be that $m_L < \eta_H$. Otherwise the total spending of the social planner on medical consumption for both types will be equal to $\pi m_L + (1-\pi)m_H > \pi\eta_H + (1-\pi)\eta_H = \eta_H$. Thus, if the H-type chooses to mimic the L-type, he will receive medical consumption below his medical need η_H resulting in infinite negative utility.

It is important to stress several results. First, since $m_L^* < m_H^*$, from Eq.(3) it follows that $c_L^* > c_H^*$. In other words, individuals who report low medical need are rewarded with higher consumption of non-medical good ($c_L^* > c_H^*$). Second, the non-medical versus medical consumption allocation of individuals with low medical need is undistorted (Eq. (8)). Third, the ratio in the bracket on the right hand side of Eq.(9) is less than one, meaning that compared to the first-best case, H-type individuals' decisions are distorted towards consuming more medical consumption and less non-medical consumption.[6]

Implementation Next, we consider an implementation of the optimal allocation in a decentralized setup. Consider the insurance system offering individuals two options, each option is characterized by a bundle $(T_i, q_i(m), i = \{1, 2\})$, where T_i is cash transfers and $q_i(m)$ is the price of medical services for an individual who chooses option i. These transfers and prices are determined as follows:

$$T_1 = c_L^* + m_L^*$$
$$q_1(m) = 1 \qquad \text{for any } m$$

$$T_2 = \ c_H^* + q_2(m_H^*)m_H^*$$

$$q_2(m) = \begin{cases} \dfrac{u'(c_L^*) + \frac{v'(m_H^* - \eta_L)}{u'(c_H^*)}\pi(u'(c_H^*) - u'(c_L^*))}{u'(c_L^*) + \pi(u'(c_H^*) - u'(c_L^*))} < 1 & \text{if } m \geq m_H^* \\ 1 & \text{if } m < m_H^* \end{cases}$$

We now show that if $T_1 \geq T_2$, the insurance system $(T_i, q_i(m), i = \{1, 2\})$ implements the optimum. The condition $T_1 \geq T_2$ requires some restrictions on the parameters of the model, which we discuss further in Appendix A. Note that the L-type who chooses the first option solves the following problem:

$$\max_{c_L, m_L} \ u(c_L) + v(m_L - \eta_L) \tag{10}$$

$$\text{s.t.} \qquad c_L + m_L = T_1,$$

[6]This can be seen as follows: $1 = \frac{v'(m_L^* - \eta_L)}{u'(c_L^*)} > \frac{v'(m_H^* - \eta_L)}{u'(c_H^*)}$. The first inequality follows from Eq.(8) and the second from the fact that $m_L^* < m_H^*$ and $c_L^* > c_H^*$ and the fact that both $u(\cdot)$ and $v(\cdot)$ are concave functions. Since $\frac{v'(m_H^* - \eta_L)}{u'(c_H^*)} < 1$, we have $\dfrac{u'(c_L^*) + \frac{v'(m_H^* - \eta_L)}{u'(c_H^*)}\pi(u'(c_H^*) - u'(c_L^*))}{u'(c_L^*) + \pi(u'(c_H^*) - u'(c_L^*))} < 1$.

while the problem of the H-type who chooses the second option is as follows:

$$\max_{c_H, m_H} \quad u(c_H) + v(m_H - \eta_H) \tag{11}$$

$$\text{s.t.} \qquad c_H + q_2(m_H)m_H = T_2.$$

By construction, (c_L^*, m_L^*) and (c_H^*, m_H^*) are feasible for Problem (10) and (11) respectively. Total spending on transfers satisfies the aggregate resource constraint because

$$\pi T_1 + (1-\pi)\left[T_2 + (1 - q_2(m_H^*))m_H^*\right] = \pi\left[c_L^* + m_L^*\right] + (1-\pi)\left[c_H^* + m_H^*\right] = B.$$

We need to show that (c_L^*, m_L^*) solves Problem (10) and (c_H^*, m_H^*) solves Problem (11). Based on Eq.(8), (c_L^*, m_L^*) satisfies the first-order conditions and therefore solves the optimization problem for the L-type.

To see that (c_H^*, m_H^*) solves the optimization problem for the H-type agent, note that based on Eq.(9) he will never choose to deviate from this point in the direction to increase m (and decrease c). Conversely, if he chooses to deviate in the direction to decrease m, he would face the price of medical consumption equal to one. Assuming a small deviation and given that his cash transfers $T_2 = c_H^* + q_2(m_H^*)m_H^*$, this would result in a new bundle, $\approx (c_H^*, q_2(m_H^*)m_H^*)$, which is clearly dominated by (c_H^*, m_H^*) since $q_2(m_H^*) < 1$.

Next, we need to show that the L-type will not choose the second option and the H-type will not choose the first option. For the latter case, the H-type will never choose the first option because $T_1 < c_H^* + m_H^*$.[7] Consider now the problem of the L-type agent who chooses the second insurance option:

$$\max_{c_L, m_L} \quad u(c_L) + v(m_L - \eta_L)$$

$$\text{s.t.} \qquad c_L + q_2(m_L)m_L = T_2$$

Note that if the L-type individual chooses $m_L < m_H^*$, he is weakly worse off than under the first option because he still faces a price of medical care equal to one but receives fewer cash transfers (because $T_1 \geq T_2$). The question is whether choosing $m_L \geq m_H^*$ will bring the L-type agent higher utility than the bundle (c_L^*, m_L^*). Because the ICC is binding at the optimum, the utility from consuming bundle (c_L^*, m_L^*) is the same as the utility from consuming bundle (c_H^*, m_H^*). The latter bundle also maximizes the utility of the H-type choosing the second insurance option. Because at any given m, the marginal utility of medical

[7] This can be shown as follows: Suppose $c_H^* + m_H^* < T_1 = c_L^* + m_L^*$, thus, from the aggregate resource constraint $B > c_H^* + m_H^* > \eta_H$. The latter inequality follows from the fact that $m_H^* > \eta_H$. However, this contradicts the assumption that $B < \eta_H$.

expenses of the L-type is less than that of the H-type, choosing $m_L > m_H^*$ cannot bring the L-type agent higher utility than (c_H^*, m_H^*). Thus, the first option (only cash transfers) is at least as good as the second option (cash transfers+insurance).[8]

Overall, these results show that one way to achieve the optimal allocation is to allow individuals to choose between a cash option and an insurance coverage (accompanied by smaller cash transfers). We will use this intuition in our quantitative policy analysis below.

4 Quantitative Model

The theoretical model described in the previous section highlights the importance of the trade-off between medical and non-medical consumption in providing efficient insurance. The analytical tractability of the model comes at the cost of making some strong assumptions, but these can be relaxed in the quantitative analysis presented in this section.

Before describing the quantitative model, however, it is important to outline what features of reality are missing in the theoretical analysis but can be included in our quantitative analysis. First, our theoretical model has no saving decisions which can be important as they allow individuals to self-insure. Second, medical/non-medical consumption choice is the only decision individuals make in the theoretical model; however, health insurance arrangements may affect other decisions such as labor supply and savings. Third, the only (ex-ante) risk in the model is medical need realization; more generally, high medical spending can be a consequence of bad health which also affects other outcomes such as earnings, employment, and life expectancy. Finally, the theoretical model abstracts from the life-cycle dimension which is important as medical spending changes considerably with age.

Our quantitative model presented below overcomes all of these limitations while maintaining the key feature of the theoretical model: medical spending has both discretionary and necessary components. We will use this quantitative model to examine the implications of our theoretical analysis in a richer setup. Even though our focus is public health insurance, we model in detail various institutional features that are not directly related to public health insurance. We do this for the following reasons.

First, unlike in other developed countries, in the US there is substantial heterogeneity in terms of insurance coverage among working-age adults. We use this observation to identify discretionary versus non-discretionary medical spending by matching the difference in medical spending profiles between the privately insured and the uninsured, while controlling for

[8] Note that this is not the case if the price of medical care for those who choose the second option is equal to $\dfrac{u'(c_L^*) + \frac{v'(m_H^* - \eta_L)}{u'(c_H^*)}\pi(u'(c_H^*) - u'(c_L^*))}{u'(c_L^*) + \pi(u'(c_H^*) - u'(c_L^*))}$ when $m < m_H^*$.

the composition of these two groups. Second, the source of insurance coverage in the US varies significantly over the life-cycle and is, to a certain extent, endogenous. Thus, when studying changes in public health insurance, it is important to capture how people self-select into different types of insurance.

4.1 Households

4.1.1 Demographics and preferences

The economy is populated by overlapping generations of individuals. An individual lives for a maximum of N periods. During the first $R-1$ periods of life, an individual can choose whether to work or not; at age R, all individuals retire.

At age t, an agent's health condition h_t can be either good ($h_t = 1$) or bad ($h_t = 0$). His health condition evolves according to an age-dependent Markov process, $\mathcal{H}_t(h_t|h_{t-1})$. Health affects one's available time, productivity and survival probability. Agents discount the future at rate β and survive until the next period with conditional probability ζ_t^h, which depends on age and health.

An individual is endowed with one unit of time that can be used for either leisure or work. Labor supply (l_t) is indivisible: $l_t \in \left\{0, \bar{l}\right\}$. Working brings disutility modeled as a fixed cost of leisure. This disutility depends on health status: healthy individuals have fixed cost of working ϕ_w, and unhealthy individuals have a higher fixed cost equal to $\phi_w + \phi_t^{UH}$, where the latter component can increase with age. We assume a Cobb-Douglas specification for preferences over consumption and leisure:

$$u(c_t, l_t, h_t) = \frac{\left(c_t^{\chi}\left(1 - l_t - \phi_w \mathbf{1}_{\{l_t>0\}} - \phi_t^{UH}\mathbf{1}_{\{h_t=0,\, l_t>0\}}\right)^{1-\chi}\right)^{1-\sigma}}{1-\sigma},$$

where $\mathbf{1}_{\{\cdot\}}$ is an indicator function equal to one if its argument is true. Here, χ is a parameter determining the relative weight of consumption, and σ is the risk-aversion coefficient over the consumption-leisure composite.

Each period an individual chooses medical consumption m_t that brings utility $v(m_t, \eta_t^h)$, where η_t^h is an age- and health-dependent medical need shock. The medical need shock represents non-discretionary medical spending, i.e., total medical consumption m_t cannot be less than medical need ($m_t > \eta_t^h$). The parametrization of the utility of medical consumption is discussed in more details in Section 4.2.

It is important to note that we do not explicitly model the link between medical spending and future health, and in this approach, we follow De Nardi et al. (2016). In our model, the utility that individuals get from additional discretionary medical consumption can be

considered as a reduced form representation of all potential benefits that this spending can bring. In the empirical literature, the extent to which medical spending can improve future health is often found to be insignificant.

Two important pieces of evidence come from the RAND Health insurance experiment and the more recent Oregon Health Insurance experiment. In the first experiment, as summarized in Newhouse et al. (1993), individuals with access to free health care increased their use of medical care but did not have significantly different health outcomes (using a variety of measures) compared to individuals who faced out-of-pocket costs for their care. In the Oregon experiment, as summarized by Baicker et al. (2013), individuals who got access to Medicaid through a lottery did not exhibit statistically significant improvements in physical health measures despite an increase in health care use. Importantly, the latter study also finds that individuals who got access to Medicaid had improved mental outcomes (reduced depression). This can be interpreted as an additional utility from health care consumption, which is consistent with our formulation.

4.1.2 Health insurance

Every period with some probability $Prob_t$, a working-age agent receives an offer to buy employer-sponsored health insurance (ESHI).[9] The variable g_t characterizes the status of the offer: $g_t = 1$ if an individual receives an offer, and $g_t = 0$ if he does not. We assume that an individual who has an offer always buys health insurance.[10] Participants of the employer-based pool face an out-of-pocket premium $\overline{p}$.[11]

Low-income individuals of working age can obtain free health insurance from Medicaid. An individual is eligible for Medicaid if his total income is below the threshold y^{cat} and his assets are less than k^{cat}.[12]

[9] In the data, unhealthy people are less likely to work at firms that offer ESHI. In addition, people with higher incomes are more likely to be offered ESHI. Based on these facts, we allow $Prob_t$ to depend on an individual's health condition and income.

[10] In our sample from the MEPS, 96% of people with an offer of ESHI take it.

[11] We refer to $\overline{p}$ as the out-of-pocket premium because the actual employer-based premium is higher but the employer contributes a significant fraction of it, so employees only pay the remainder of the cost.

[12] The institutional framework described in the model corresponds to the situation before the Affordable Care Act (ACA) became effective. Prior to the ACA, federal regulations only required states' Medicaid programs to cover certain categories of the population - individuals with dependent children and low-income disabled individuals. We abstract from these eligibility criteria because many states had additional pathways for childless adults through Medicaid expansion programs. In 2008, 23 states and the District of Columbia operated programs for low-income childless adults (Klein and Schwartz, 2008). The financing of these programs comes from state funding or through Medicaid §1115 waivers. As a result, introducing a tight link between Medicaid eligibility and family/disability status may significantly underestimate the extent to which this program is available to other categories of the population. In addition, modeling family or disability status even in a simplistic way would require us to introduce an additional state variable which would significantly increase our computational costs.

We denote by i_t an individual's health insurance status. If an individual is uninsured, $i_t = U$; if an individual has ESHI, $i_t = G$; if an individual has Medicaid, $i_t = M$. All types of insurance contracts - both private and public - cover only a portion of each individual's medical spending. We denote by ded^{ESI} and ded^{MCD} the deductibles for employer-based insurance and Medicaid, respectively, while denoting by q^{ESI} and q^{MCD} the fractions of medical expenditures above the deductibles paid by private and public insurance, respectively.

All retired households are enrolled in the Medicare program. The Medicare program charges a fixed premium p_{MCR} and covers a fraction q^{MCR} of medical costs above the deductible level ded^{MCR}.

4.1.3 Labor income

Individual earnings are equal to $z_t^h l_t$, where z_t^h is the idiosyncratic productivity that depends on age (t) and health status (h_t). The latter modeling assumption is motivated by the observation that the average labor income of unhealthy workers is much lower than the average income of healthy workers in the data.

4.1.4 Taxation and social transfers

Individuals pay three types of taxes. First, there is an income tax $\mathcal{T}(y_t)$, where taxable income y_t includes both labor and capital income. Second, there is a consumption tax τ_c. Third, working individuals pay payroll taxes: Medicare tax (τ_{MCR}) and Social Security tax (τ_{ss}). The latter tax does not apply to earnings above the level $\overline{y}_{ss}$. We also model several health-related deductions existing in the US tax code. In particular, out-of-pocket medical expenditures exceeding 7.5% of households' income and ESHI premium ($\overline{p}$) can be excluded from taxable income.

There are two types of cash transfers. First, all retired households receive Social Security benefits ss. In practice these payments depend on an individual's history of earnings, but to avoid the computational cost of introducing an additional state variable, we model ss as fixed payments which depend on average income in the economy.

Second, poor individuals can rely on the safety-net program, T_t^{SI}. This program guarantees each individual a minimum subsistence level equal to $\underline{c}$ by giving transfers to people with low disposable resources. This minimum consumption floor is a stylized representation of means-tested public transfer programs such as food stamps, the Medically Needy part of Medicaid, Supplemental Security Income, disability insurance, and uncompensated medical care.

4.1.5 Timing in the model

The timing in the model is as follows. In the beginning of the period, an individual learns his productivity, health status, ESHI offer, and medical need shock. Next, an individual chooses his labor supply (l_t). An individual who chooses not to work cannot access ESHI. An individual without ESHI and whose total resources allow him to be eligible for Medicaid gets public insurance. Then an individual chooses his medical consumption (m_t), non-medical consumption (c_t) and savings (k_{t+1}).

4.1.6 Optimization problem

Working-age individuals ($t < R$) The state variables for working-age individuals' optimization problem at the beginning of each period are capital ($k_t \in \mathbb{K} = R^+ \cup \{0\}$), health status ($h_t \in \mathbb{H} = \{0, 1\}$), medical need shock ($\eta_t^h \in R$), idiosyncratic labor productivity $\left(z_t^h \in \mathbb{Z} = R^+\right)$, ESHI offer status ($g_t \in \mathbb{G} = \{0, 1\}$), and age ($t \in \mathbb{T} = \{1, 2, ..., R-1\}$).

The value function of a working-age individual can be written as follows:

$$V_t\left(k_t, h_t, \eta_t^h, z_t^h, g_t\right) = \max_{c_t, l_t, m_t, k_{t+1}} u\left(c_t, l_t, h_t\right) + v(m_t, \eta_t^h) + \beta \zeta_t^h E_t V_{t+1}\left(k_{t+1}, h_{t+1}, \eta_{t+1}^h, z_{t+1}^h, g_{t+1}\right) \tag{12}$$

subject to

$$k_t\left(1 + r\right) + z_t^h l_t + T^{SI} = k_{t+1} + \left(1 + \tau_c\right) c_t + Tax + P_t + X(m_t) \tag{13}$$

$$P_t = \begin{cases} 0 & ; \text{ if } i_t \in \{U, M\} \\ \overline{p} & ; \text{ if } i_t \in \{G\} \end{cases} \tag{14}$$

$$T_t^{SI} = \max\left(0, \underline{c} + Tax + P_t + X(\eta_t^h) - k_t\left(1 + r\right) - z_t^h l_t\right) \tag{15}$$

$$Tax = \mathcal{T}\left(y_t\right) + \tau_{MCR}\left(z_t^h l_t - \overline{p}\mathbf{1}_{\{i_t = G\}}\right) + \tau_{ss} \max\left(z_t^h l_t - \overline{p}\mathbf{1}_{\{i_t = G\}}, \overline{y}_{ss}\right) \tag{16}$$

$$y_t = \max\left(0, k_t r + z_t^h l_t - \overline{p}\mathbf{1}_{\{i_t = G\}} - \max\left(0, X(m_t) - 0.075(k_t r + z_t^h l_t)\right)\right) \tag{17}$$

$$X(m_t) = \begin{cases} m_t & \text{if } i_t = \{U\} \text{ or } m_t \leq ded^{i_t} \\ ded^{i_t} + (1 - q^{i_t})(m_t - ded^{i_t}) & \text{if } i_t = \{M, G\} \text{ and } m_t > ded^{i_t} \end{cases} \tag{18}$$

An individual is enrolled in ESHI ($i_t = G$) if $l_t = \bar{l}$ and $g_t = 1$. An individual is enrolled

in Medicaid ($i_t = M$) if:

$$k_t r + z_t^h l_t \leq y^{cat} \text{ and } k_t \leq k^{cat}$$

The conditional expectation on the right-hand side of Eq.(12) is over $\left\{h_{t+1}, \eta_{t+1}^h, z_{t+1}^h, g_{t+1}\right\}$. Eq.(13) is the budget constraint. Eq.(15) describes how the transfers of cash assistance programs are calculated. In Eq.(16), the first two terms are income taxes and the last two terms are payroll taxes. Eq. (17) defines taxable income. Eq.(18) describes out-of-pocket medical expenses $X(m_t)$ as a function of total medical expenses m_t.

Retired individuals For a retired individual ($t \geq R$), the state variables are capital (k_t), health (h_t), medical need shock $\left(\eta_t^h\right)$, and age (t). The value function of a retired individual is:

$$V_t\left(k_t, h_t, \eta_t^h\right) = \max_{c_t, k_{t+1}} u\left(c_t, 0, h_t\right) + v(m_t, \eta_t^h) + \beta \zeta_t^h E_t V_{t+1}\left(k_{t+1}, h_{t+1}, \eta_{t+1}^h\right) \tag{19}$$

subject to:

$$k_t\left(1+r\right) + ss + T^{SI} = k_{t+1} + \left(1+\tau_c\right) c_t + \mathcal{T}\left(y_t\right) + p_{MCR} + X(m_t) \tag{20}$$

$$T_t^{SI} = \max\left(0, \underline{c} + \mathcal{T}\left(y_t\right) + p_{MCR} + X(\eta_t^h) - k_t\left(1+r\right) - ss\right) \tag{21}$$

$$y_t = k_t r + ss - \max\left(0, X(m_t) - 0.075\left(k_t r + ss\right)\right) \tag{22}$$

$$X(m_t) = \begin{cases} m_t & \text{if } m_t \leq ded^{MCR} \\ ded^{MCR} + (1 - q^{MCR})(m_t - ded^{MCR}) & \text{if } m_t > ded^{MCR} \end{cases} \tag{23}$$

4.2 Utility from medical consumption

When specifying the functional form of the utility from medical consumption we require it to satisfy three properties. First, there should be a large disutility from consuming less than one's medical need. This property ensures that medical need represents unavoidable or necessary medical consumption.

Second, it should allow for varying compositions of necessary and discretionary medical spending. The key intuition underlying our identification strategy is that health insurance affects only the discretionary part of medical spending. Consider, for example, two extreme cases. If the entire amount of medical spending is necessary, health insurance does not affect medical spending because the latter is a shock. On the other hand, if medical spending is entirely discretionary, i.e., it is not a shock but a choice, health insurance is important

because it affects its price. We want our utility specification to parsimoniously represent various intermediate scenarios between these two extreme cases.

Third, the utility specification should be such that even though discretionary medical consumption is valuable, the demand for medical care is income inelastic. A number of studies have documented that the elasticity of medical spending with respect to income is very small (for an extensive review, see Ringel et al., 2002, and OECD, 2006) and we require our model to be consistent with this observation.

Based on these considerations, we use a specification which combines CRRA utility over discretionary medical spending $(m_t - \eta_t^h)$ with a quadratic component:[13]

$$v(m_t, \eta_t^h) = \frac{(m_t - \eta_t^h)^{1-\sigma}}{1-\sigma} - \frac{1}{2}m_t^2 + \gamma_{1,t}^h m_t \tag{24}$$

Note that this functional form has the property that the marginal utility of medical consumption is not always positive; i.e., medical consumption can reach a saturation point. We denote by Δ the maximum amount of discretionary medical consumption on top of the medical need that an agent can consume before marginal utility becomes negative (medical consumption reaches the saturation point), i.e., an agent's medical consumption always lies within the bracket $(\eta_t^h, \eta_t^h + \Delta]$. We can rewrite $\gamma_{1,t}^h$ as a function of Δ and η_t^h, so our utility function depends only on these two parameters.[14]

Note that our specification satisfies all three properties listed above. The CRRA component ensures that medical spending is always above medical need. By changing the saturation point we can change the composition of spending into necessary and discretionary parts (this is discussed in Sections 5.4.2 and 6.3). Finally, the fact that the marginal utility of medical spending is not always positive works towards decreasing the income elasticity of medical consumption.[15]

---

[13] An alternative specification is the CRRA function with a multiplier θ: $\theta\frac{(m_t - \eta_t^h)^{1-\sigma}}{1-\sigma}$. This specification was used in Ameriks et al. (2017) with application to spending on long-term care. This specification satisfies the first two properties but not the third one. We have tried to re-estimate our model using this functional form; in Appendix B we show that this alternative model can capture many features of the data but it significantly overestimates the empirical price and income elasticities of the demand for medical care.

[14] Specifically, because at the saturation point

$$\frac{\partial v(m_t, \eta_t^h)}{\partial m_t}|_{m_t = \eta_t^h + \Delta} = 0,$$

we have $\gamma_{1,t}^h = \eta_t^h + \Delta - \Delta^{-\sigma}$.

[15] If the marginal utility of medical consumption is always positive, a fraction of every additional dollar of income will be allocated towards medical spending thereby resulting in high income elasticity.

5 Model parameters estimation

5.1 Data and estimation procedure

The main dataset we use in our estimation is the Medical Expenditure Panel Survey (MEPS). The MEPS is a nationally representative survey of households that collects detailed records on demographics, income, employment, medical costs and insurance. Each individual is interviewed at most five times over a two-year period. The medical spending reported in the MEPS is cross-checked with insurers and providers, which improves the accuracy.[16] We use fourteen waves of the MEPS (1999-2012) to estimate the distribution of medical spending, health dynamics, parameters related to health insurance, labor earnings and employment profiles.

In our sample, we include heads of households who are older than 20 years old and do not have missing information on medical spending. In the MEPS, a family (or a household) is defined based on eligibility for coverage under a typical family insurance plan (refereed to in the MEPS as Health Insurance Eligibility Unit, HIEU). We define the head as the person with the highest income in the HIEU. We choose to include only HIEU heads because in our model we abstract from modeling families.

The MEPS does not contain information about wealth and survival, because of this we also use the Panel Study of Income Dynamic (PSID) and the Health and Retirement Study (HRS) to construct these related moments. The PSID is a nationally representative panel that surveys individuals and their families. We use the PSID to construct the moment for wealth. The PSID collected wealth information every five years before 1997 and every two years after that, and we use the 1994 and 1999-2011 waves.[17]

The HRS is a bi-annual panel that surveys a nationally representative sample of individuals over the age of 50. We use the HRS to estimate the health-dependent survival probabilities.

We estimate/calibrate our model in two steps. In the first step, we set parameters related to demographics, taxes, social security benefits, labor productivity shocks, and health insurance and estimate the health transition probabilities directly from the data. In the second step, we calibrate the remaining parameters using our model to match the targeted moments from the data. We convert nominal values to constant 2003 dollars using the CPI as a deflator.

[16] Pashchenko and Porapakkarm (2016) provide more details on the MEPS dataset.

[17] Our measure of net worth controls for family size and year effects.

5.2 Parameters set/estimated in the first step

5.2.1 Demographics and preferences

In the model, agents are born at age 25 and can live to a maximum age of 99. The model period is one year; thus, the maximum lifespan N is 75. Agents retire at the age of 65, so R is 41.

To adjust conditional survival probabilities ζ_t^h for health, we follow Attanasio et al. (2011). More specifically, we first use the HRS to estimate the survival probabilities by health, and we use the MEPS to estimate the fraction of healthy and unhealthy individuals by age. Then we combine these estimates to compute the difference in survival probabilities for people in different health status and use it to adjust the Social Security Administration's male life tables.

We set the consumption share (χ) in the utility function to 0.6 using the estimates of French (2005).[18] The risk aversion parameter σ is set to 3. This corresponds to the risk-aversion over non-medical consumption equal to 2.3, which is within the range commonly used in structural life-cycle and macroeconomic models. The interest rate r is set to 2%. We set labor supply of those who choose to work ($\bar{l}$) to 0.4. Individuals' initial wealth is set to the median wealth of people aged 20 and 25 years old in the PSID.

5.2.2 Government policies

In specifying the tax function $\mathcal{T}(y)$, we use a nonlinear functional form as in Gouveia and Strauss (1994), together with a linear income tax τ_y:

$$\mathcal{T}(y) = a_0 \left[y - (y^{-a_1} + a_2)^{-1/a_1} \right] + \tau_y y$$

In this functional form, a_0 controls the marginal tax rate faced by the highest income group, a_1 determines the curvature of marginal taxes, and a_2 is a scaling parameter. We set a_0 and a_1 to 0.258 and 0.768, respectively, as in Gouveia and Strauss (1994), and the parameters a_2 and τ_y are set to 0.616 and 0.067, respectively, based on Pashchenko and Porapakkarm (2017).

For retired individuals, Social Security pension payment ss is the average labor income over working ages (25-64) multiplied by the replacement rate. The Social Security replacement rate is set to 40%.

The Medicare, Social Security and consumption tax rates are set to 2.9%, 12.4%, and 5.67%, respectively. The maximum taxable income for Social Security is set to $87,000 as

[18] Given that we have indivisible labor supply, we cannot pin down this parameter using a moment in the data.

in 2003.

5.2.3 Health insurance, health, and health transition probability

In the MEPS, the question about the source of insurance coverage is asked retrospectively for each month of the year. We define a person as having employer-based insurance if he reports having ESHI for at least six months of the year; the same criterion is used when defining individuals insured by Medicaid. We reclassify individuals as being insured by Medicaid even if they report another type of coverage (or no coverage) in two cases. First, if they report receiving Supplemental Security Income (SSI). Second, if more than 30% of their total medical spending is covered by Medicaid. We define a person as uninsured if he has no insurance coverage for six months or more out of the year. We exclude from our sample individuals who are covered by individual insurance or any other type of health insurance which cannot be categorized into Medicaid, individual or employer-based insurance.

We construct our measure of health based on self-reported health status. In the MEPS, a person's self-reported health status is coded as 1 for excellent, 2 for very good, 3 for good, 4 for fair, and 5 for poor. We define a person as being in bad health if his average health score over a given year is greater than or equal to 3.

To construct the age-dependent health transition matrix, we start by computing the health transition probabilities for ages 30, 40,...,70. In each case, we use a sample within a 10-year age bracket. For example, to construct the transition probabilities for age 40, we pool individuals between ages 35 and 44. Then we construct the health transition matrix by fitting these estimates with quadratic functions of age.

5.2.4 Insurance policies

The coinsurance for each type of insurance was computed as the median ratio of out-of-pocket medical expenses to total medical expenses for people within a corresponding insurance group who have positive medical spending and positive insurance payouts. The resulting coinsurance rates are 23%, 3%, and 20% for private insurance, Medicaid and Medicare, respectively.

The deductibles for private insurance are set to \$143, which corresponds to the median out-of-pocket costs for people with private insurance whose medical expenses are positive but insurance payments are zero. The deductibles for Medicaid are set to zero since the absolute majority of Medicaid beneficiaries have positive insurance payout and very small out-of-pocket payments. The deductibles for Medicare are set to be the same as the deductibles for private insurance. We cannot compute this number from the data as very few individuals above the age of 65 have positive total medical spending but zero insurance payouts.

We set the out-of-pocket part of ESHI premium $\overline{p}$ to \$508 based on Kaiser (2002). The premium for Medicare p_{MCR} is set to \$704 which was the Medicare Part B premium in 2003.[19]

5.3 Labor income

We specify labor productivity as follows:

$$z_t^h = \lambda_t^h \exp(\omega_t) \exp(\xi) \tag{25}$$

where λ_t^h is the deterministic function of age and health. The stochastic component of productivity consists of the persistent shock ω_t and a fixed productivity ξ:

$$\omega_t = \rho\omega_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0, \sigma_\varepsilon^2) \tag{26}$$

$$\xi \sim N(0, \sigma_\xi^2)$$

Fixed productivity is discretized into two grids with equal measure, referred to as low (ξ^1) and high (ξ^2) fixed productivity types ($\xi^1 < \xi^2$). For the persistent shock ω_t, we set ρ to 0.98 and σ_ε^2 to 0.02 following the incomplete market literature (Storesletten et al. (2004); Hubbard et al. (1995); French (2005)). We set the variance of the fixed productivity type (σ_ξ^2) to 0.242 as in Storesletten et al. (2004). In our computation, we discretize ω_t using the method in Floden (2008).[20] To construct the distribution of newborn individuals, we draw ω_1 in Eq.(26) from the $N(0, 0.352^2)$ distribution, following Heathcote et al. (2010).

To estimate the deterministic part of productivity λ_t^h, we need to take into account that in the data, we only observe labor income of workers and we do not know the potential labor income of non-workers. To avoid the selection bias, particularly among the unhealthy, we estimate the labor income profiles inside the model. The estimation approach is described in Section 5.4.5.

5.4 Parameters estimated inside the model

In this section, we explain how we calibrate the remaining parameters inside our model by targeting the moments from the data.

[19] Most Medicare beneficiaries do not pay a premium for Part A.

[20] We use 9 gridpoints for ω_t, and the grid of ω_t is expanding over ages to capture the increasing cross-sectional variance. Our discretized process for ω_t generates an autocorrelation of 0.98 and 0.0173 for its innovation variance.

5.4.1 Discount factor and institutional parameters

The discount factor β is set to 0.968 to match the ratio of median assets of people aged 60-64 to median assets of people aged 35-39 in the PSID.

The minimum subsistence level $\underline{c}$ is set to \$2,000 to match the average employment among Medicaid beneficiaries. The income eligibility threshold for Medicaid (y^{cat}) is set to 94% of Federal Poverty Line (FPL) and the asset test is set to \$17,000 to match the life-cycle profile of people with public insurance.

5.4.2 Saturation point

In our model, the marginal utility of medical consumption becomes zero when medical spending reaches the saturation point, $\eta_t^h + \Delta$. As explained in the next subsection, medical need η_t^h is estimated to match the *total* medical spending in the data. The parameter Δ determines the proportion of *discretionary* medical expenses within the total.

We identify Δ by matching the difference in medical spending profiles between the uninsured and the privately insured. The model where medical expenses are mostly non-discretionary (Δ is rather small) will underestimate this difference, while the model where medical consumption is mostly discretionary ($\Delta \gg \eta_t^h$) will overestimate it.

It is important to point out that our estimation strategy is to match the average and median medical spending by health, while simultaneously matching the health composition of individuals in different insurance groups (as will be shown in Section 6). In addition, we match the average labor income of people with private insurance. Thus, we identify Δ from the difference in medical spending between the privately insured and the uninsured, *which is not accounted for by the difference in composition between these two groups*. In Section 6.3, we illustrate how changing Δ affects the difference in medical spending between the privately insured and the uninsured.[21]

5.4.3 Medical need shock process

We assume that the medical need shock η_t^h has a shifted lognormal distribution

$$\eta_t^h = \exp(\kappa_t^h) - \exp(b_t^h)$$

$$\kappa_t^h = \mu_t^h + \delta_t^h \zeta_t$$

[21] We have also considered the case where Δ is age-dependent, but the resulting estimate varies little by age.

$$\zeta_t = \rho_m \zeta_{t-1} + \varepsilon_t, \quad \varepsilon_t \sim N(0,1)$$

We assume a lognormal distribution to capture the empirical fact that the medical expenses distribution is highly skewed.

Note that in the data, there is a large fraction of people with zero medical spending, especially among the young. Without the shift parameter b_t^h, η_t^h is always positive, and *all* individuals in our model have positive medical expenses, which is counterfactual. We parametrize b_t^h as follows:

$$b_t^h = b_0^h + b_1^h t + b_2^h t^2 + b_3^h t^3 + b_4^h t^4$$

Similarly, we parameterize μ_t^h and δ_t^h as:

$$\mu_t^h = \mu_0^h + \mu_1^h t + \mu_2^h t^2 + \mu_3^h t^3$$

$$\delta_t^h = \delta_0^h + \delta_1^h t + \delta_2^h t^2 + \delta_3^h t^3$$

The process of medical need shock η_t^h is estimated inside the model in the following way. Given Δ, the health-dependent coefficients $\mu_0^h, \mu_1^h, \mu_2^h,$ and μ_3^h are set to match average medical expenses at ages 30, 50, 70, and 85 for each health group. The coefficients $\delta_0^h, \delta_1^h, \delta_2^h,$ and δ_3^h are set to match the variance of medical expenses at ages 30, 50, 70, and 85 for each health group. The coefficients $b_0^h, b_1^h, b_2^h, b_3^h,$ and b_4^h are used to match the fraction of people with zero medical expenses at ages 25, 40, 50, 64, and 97 for each health group. The persistence parameter ρ_m is set to match the autocorrelation of total medical expenses in the data (0.3).[22] Since all of the parameters governing the evolution of medical need depend on Δ, we estimate the parameters b_t^h, μ_t^h and δ_t^h jointly with Δ.

5.4.4 ESHI offer rate

We assume that the probability of receiving an offer of ESHI coverage is a logistic function

$$Prob_t = \frac{\exp(u_t)}{1+\exp(u_t)},$$

where the variable u_t is an odds ratio that takes the following form:

$$u_t = \eta_{0,t} + \varkappa_{1,t}\mathbf{1}_{\{h_{t-1}=0\}} + \varkappa_{2,t}\log(inc_t\,) + \varkappa_{3,t}\log(inc_t\,)\mathbf{1}_{\{h_{t-1}=0\}} + \varkappa_4\mathbf{1}_{\{g_{t-1}=1\}}\mathbf{1}_{\{t>25\}} \quad (27)$$

[22] This number is computed as the correlation of medical spending between two consecutive years for those individuals in our sample that we observe for two years.

Here, $\varkappa_{0,t}, \varkappa_{1,t}, \varkappa_{2,t}$, and $\varkappa_{3,t}$ are age-dependent coefficients, and inc_t is individual labor income. This specification allows us to match the life-cycle profile of ESHI coverage and the average labor income of workers with ESHI. We include dummy coefficients for bad health to capture decreased opportunity to access ESHI for the unhealthy.

In general, estimating Eq.(27) directly from the data can give biased results because of the selection into employment: individuals with an ESHI offer are more likely to work than those without an ESHI offer.[23] To avoid this problem, we follow Pashchenko and Porapakkarm (2013) by estimating Eq.(27) inside the model together with the labor income as described in the next subsection.

5.4.5 Disutility from work and labor income

We estimate the fixed leisure costs of work ϕ_w, the loss of time due to bad health ϕ_t^{UH}, the deterministic part of productivity λ_t^h and the parameters in Eq.(27) using a method similar to French (2005) and Pashchenko and Porapakkarm (2013).[24] Our estimation algorithm searches for these parameters until our model produces the following outcomes: i) the labor income profiles generated by our model are the same as in the data for all workers as well as for only workers covered by ESHI for each health group; ii) the profiles of ESHI coverage and employment in the model are the same as in the data for each health group, iii) the probability of being insured by ESHI in the current period conditional on being insured by ESHI in the previous period is the same in the model and in the data.[25]

5.5 Parameters' values

Table 1 summarizes the parametrization of our model. The top panel lists parameters set or estimated outside the model and the bottom panel lists parameters estimated/calibrated inside the model.

[23] See French and Jones (2010) for an investigation of the effect of employer-based health insurance on decisions to work.

[24] In our estimation, for each health group, we parametrize λ_t^h as a polynomial degree three of age.

[25] In our estimation, we define a person as employed if he works at least 520 hours per year, earns at least \$2,678 per year in base year dollars (this corresponds to working at least 10 hours per week and earning a minimum wage of \$5.15 per hour) and does not report being retired or receiving Social Security benefits. Household heads' labor income is defined as the sum of wage/salary income and 50% of the income from business.

Parameter name	Notation	Value	Source
Parameters set outside the model			
Consumption share	$\varkappa$	0.6	French (2005)
Labor supply	$\bar{l}$	0.4	
Risk aversion	σ	3	
Tax parameters	a_0	0.258	Gouveia and Strauss (1994)
	a_1	0.768	"
	a_2	0.616	Pashchenko and Porapakkarm (2017)
	τ_y	0.067	"
Social Security replacement rate	$-$	40%	
Medicare premium	p^{med}	\$704	Data (2003)
Out-of-pocket ESHI premium	$\overline{p}$	\$508	Kaiser (2002)
Labor productivity			
- Persistence parameter	ρ	0.98	Storesletten, et al. (2004)
- Variance of innovations	σ^2_ε	0.02	"
- Fixed effect	σ^2_ξ	0.24	"
Deductible and cost-sharing			
- ESHI	ded^{ESI}, q^{ESI}	\$143, 77%	MEPS
- Medicaid	ded^{MCD}, q^{MCD}	\$0, 97%	MEPS
- Medicare	ded^{MCR}, q^{MCR}	\$143, 80%	MEPS
Parameters used to match some targets			
Discount factor	β	0.968	Ratio of median assets 60-64 to 35-39
Consumption floor	$\underline{c}$	\$2,000	% employment among public insurance
Medicaid			
- Income test	y^{cat}	$0.94FPL$	publicly insured profile
- Asset test	k^{cat}	\$17,000	"
Fixed costs of work	ϕ_w	0.315	employment profiles (healthy)
Time loss due to bad health			
- age 25-40	ϕ^{UH}_t	0.001	employment profiles (unhealthy)
- age 64	ϕ^{UH}_t	0.003	"
Saturation point	Δ	0.329	difference in medical spending ESHI/uninsured

Table 1: Parameters in baseline model

6 Model performance

6.1 Employment, labor income and health insurance

Figure (2) compares the average labor income of workers and employment profiles by health in the data and in the model. The model closely tracks the data. The average labor income profiles and employment profiles by health were targeted in our calibration by adjusting the exogenous productivity and the disutility from work parameters.

Table 2 shows how the aggregate health insurance statistics in the model compare to the corresponding ones in our sample. We are able to replicate the empirical distribution of people by health insurance status. Figure (3) shows that we can also capture the life-cycle insurance profiles for individuals in different health statuses.

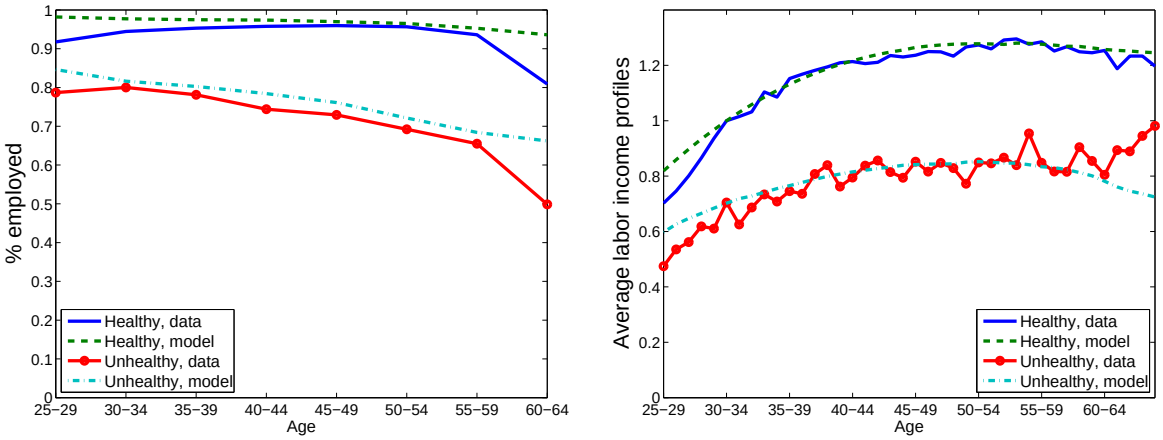

Figure 2: Left panel: employment by health. Right panel: average labor income among workers by health (normalized by average income). Solid lines (lines with round markers): data for the healthy (unhealthy) from the MEPS. Dashed (dash-dotted) lines: model for the healthy (unhealthy).

Finally, our model can replicate the selection of unhealthy people into different health insurance groups (the last row of Table 2). Specifically, the fraction of unhealthy individuals among the uninsured is higher than the fraction among those with ESHI and our model captures this observation, which is important in our estimation of the saturation point. The highest fraction of unhealthy individuals is observed among the publicly insured: 54.4% of Medicaid beneficiaries are unhealthy, which is more than twice the corresponding number for other insurance categories. In our model, the fraction of unhealthy Medicaid beneficiaries is 52.7%.

	Data			Baseline model		
	ESHI	uninsured	public	ESHI	uninsured	public
all	68.3	22.3	9.3	71.6	19.5	8.9
healthy	73.7	21.0	5.3	75.9	19.1	5.0
unhealthy	47.8	27.0	25.2	48.9	21.3	29.7
% unhealthy by insurance	14.0	24.4	54.4	10.8	17.3	52.7

Table 2: Insurance coverage among individuals aged 25 to 64: data (MEPS) versus baseline model

6.2 Medical expenses

Figure (4) shows that our calibration strategy allows the model to replicate the life-cycle profiles of the mean, median and standard deviation of medical expenses, as well as the fraction of people with zero medical expenses for each health group. The mean, standard deviation, and fraction of people with zero expenses were explicitly targeted by our calibration, whereas the median was not.

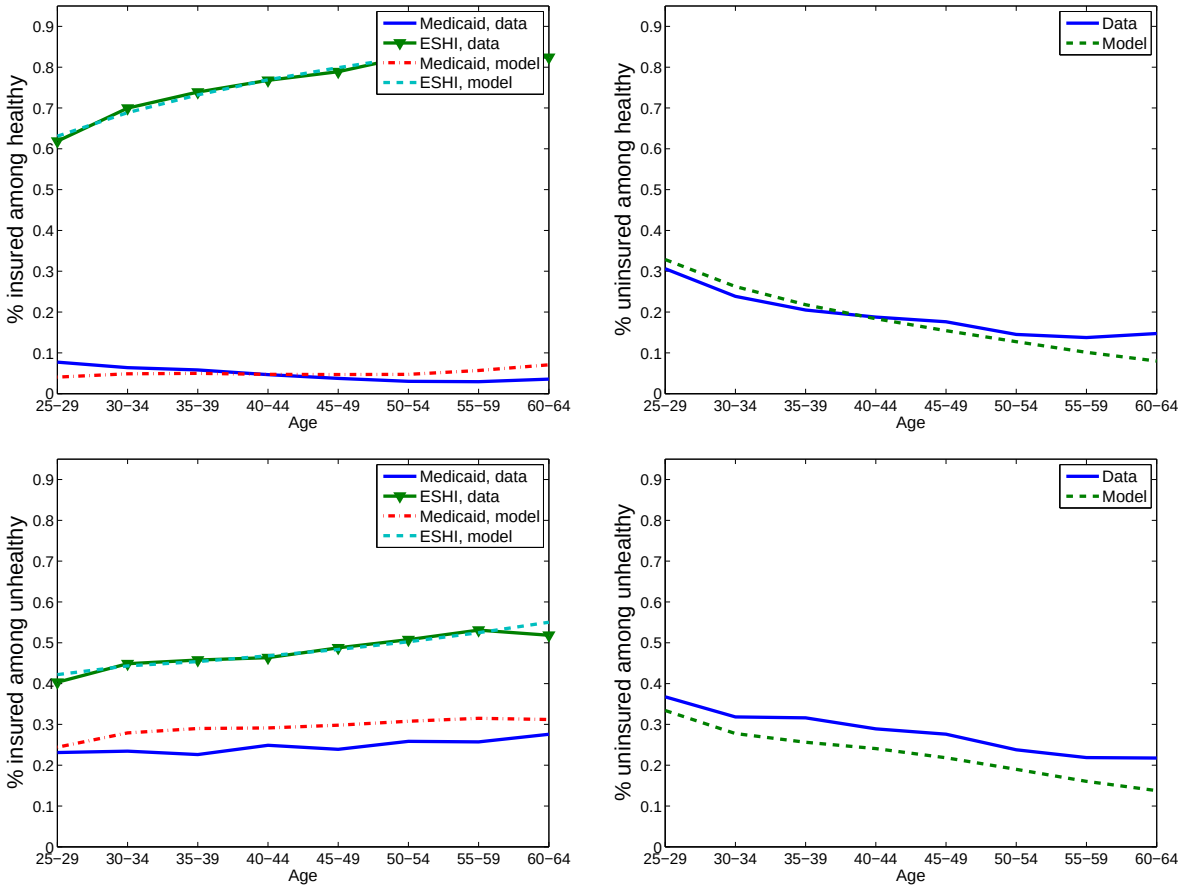

Figure 3: Insurance profiles for the healthy (top panel) and the unhealthy (bottom panel). Solid lines and lines with triangle markers are from the MEPS data. Dotted and dash-dotted lines are from the baseline model.

Figure (5) shows that our model captures the large disparities in medical expenses observed for people with different health insurance statuses. The difference in medical spending between the privately insured and the uninsured was targeted in our calibration. Although we did not explicitly target the medical spending profile of Medicaid beneficiaries, our model captures the fact that the publicly insured spend considerably more compared with the other groups.

It is also important to evaluate how well the model matches the price elasticity of the demand for medical care. The "gold standard" in empirical studies of price elasticity is the RAND health insurance experiment (Keeler and Rolph, 1988, Manning et al., 1987). This experiment was a randomized trial conducted in the 1970s. Individuals were enrolled in health insurance plans that differed in generosity, and the elasticity was computed by comparing medical care utilizations across plans. The results of this experiment have recently been reevaluated and adjusted by Aron-Dine et al. (2013).

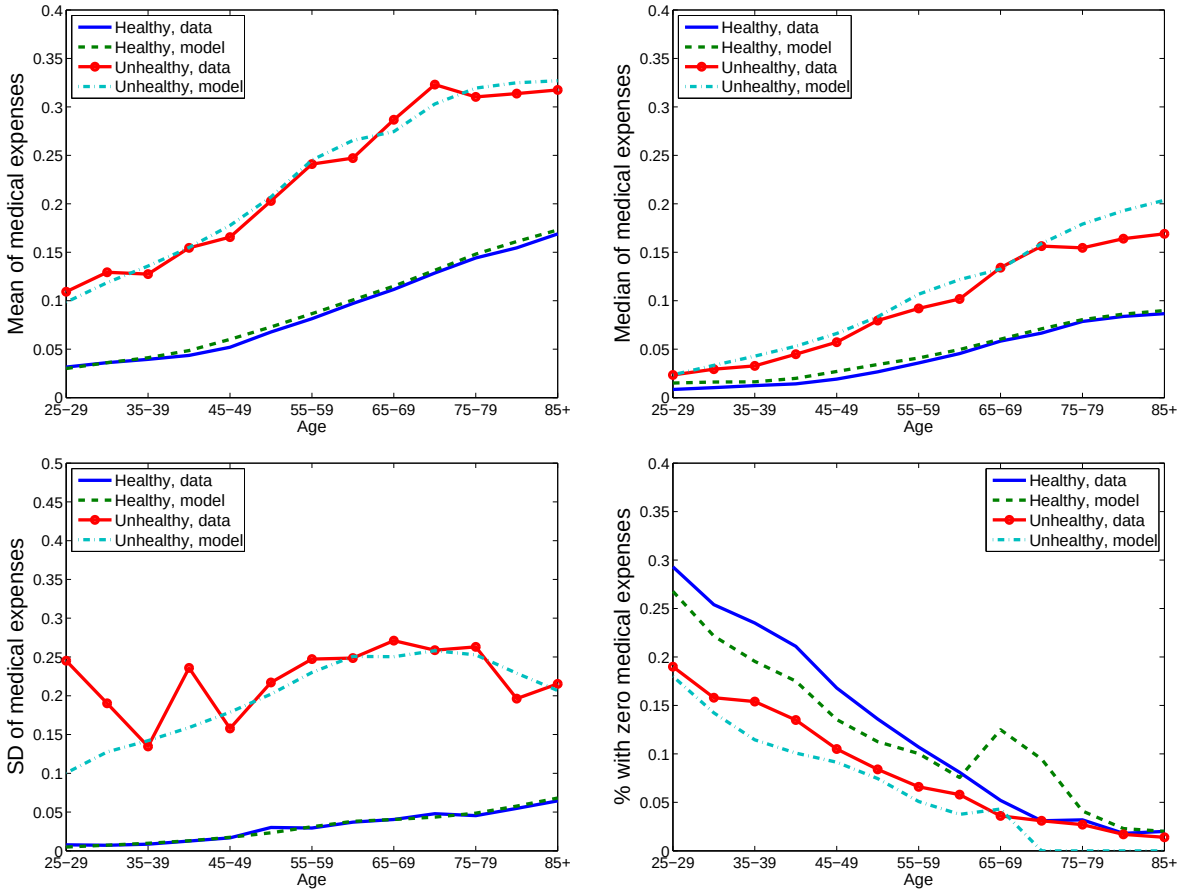

Figure 4: Top left panel: average medical expenses by health. Top right panel: median of medical expenses by health. Bottom left panel: standard deviation of medical expenses by health. Bottom right panel: fraction of people with zero medical expenses by health. Solid lines (lines with round markers) are from the MEPS data for the healthy (unhealthy). Dashed (dash-dotted) lines are from the model for the healthy (unhealthy). All level variables are normalized by average income.

To construct an elasticity measurement comparable to the RAND experiment, we consider the following two experiments. In the first one, we introduce universal health insurance with no coinsurance and no deductibles (analogous to control group A in the RAND experiment). In the second experiment, we introduce universal health insurance with no deductibles but a 25% copay rate (analogous to group B). The aggregate medical spending of working-age individuals in the second experiment constitutes 82% of the corresponding spending in the first experiment. This is very close to the result of 84% from the RAND experiment based on the adjusted estimates from Aron-Dine et al. (2013), Table 3.[26]

Another important dimension to consider is the income elasticity of medical spending. To measure this elasticity, we consider an experiment where we introduce a permanent increase in individual productivity equal to 1% and then compute the resulting change in

[26]The original number from the RAND experiment is 71%.

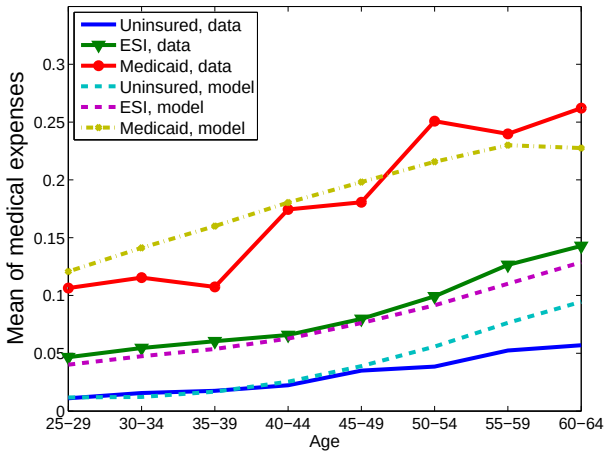

Figure 5: Average medical expenses by insurance status (normalized by average income). Solid lines (with or without markers) are from the MEPS data. Dashed and dash-dotted lines are from the model. ESI stands for employer-sponsored insurance.

the aggregate medical spending of working-age individuals. The results suggest that medical spending is income inelastic: the elasticity is equal to 0.13. This is consistent with empirical evidence. Ringel et al. (2002) review a large number of studies and conclude that the estimates of income elasticity are in the range of 0 to 0.2. In another large review, OECD (2006) also concludes that income elasticities are usually found to be small or negative. The income elasticity in the RAND health insurance experiment was found to be 0.22 (Manning and Marquis, 1996).

6.3 The role of the saturation point

Because the saturation point (controlled by the parameter Δ) plays an important role in determining the relative importance of discretionary versus non-discretionary medical spending, in this section we provide additional discussion on the identification of this parameter. To illustrate the role of the saturation point in our estimation strategy, we consider two experiments. In the first experiment, we decrease the parameter Δ by 50%, and in the second, we increase it by 20%.[27] In both experiments, we re-estimate the medical need shock process so that total medical spending profiles by health are the same as in the data.

Figure (6) shows how the difference in medical expenses between the privately insured and the uninsured in the two experiments differ from the baseline. When the saturation point is lower, a larger share of medical spending is non-discretionary, and therefore, the difference in spending between the privately insured and the uninsured decreases. In contrast, when the

[27]Increasing Δ by more than 20% does not allow us to match the total medical expense profiles in the data.

saturation point increases, the gap between the two profiles increases because discretionary spending now makes up a larger part of the total spending.

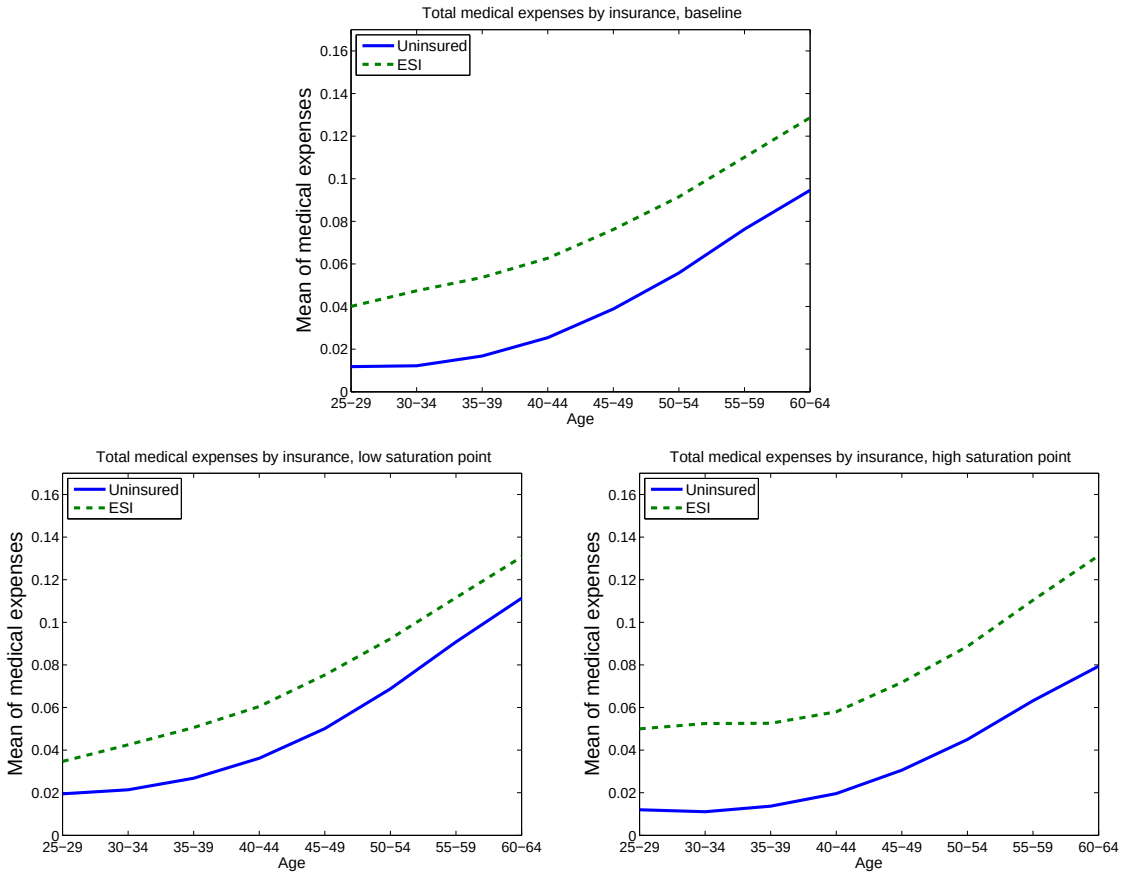

Figure 6: Average medical spending of the uninsured and individuals with employer-based insurance (normalized by average income). Top panel: baseline. Bottom left panel: Δ decreased by 50% of the baseline value. Bottom right panel: Δ increased by 20% of the baseline value. ESI stands for "employer-sponsored insurance".

7 Policy Simulations

This section is organized as follows. We start by constructing the full information benchmark where the government can observe medical needs of Medicaid beneficiaries. We show that in this case, the government can substantially reduce their medical spending by fully covering their medical needs and allocating the rest of the Medicaid budget as lump-sum transfers.

Next, we construct a number of policy experiments in an environment where the division of medical spending into discretionary and non-discretionary parts is unobservable. To abstract from any change in the composition of Medicaid beneficiaries and to better illustrate

the mechanism, we first conduct the analysis as a *one-time policy change*. We show that to move close to the full information benchmark, it is important to introduce a trade-off between non-medical and medical consumption by allowing Medicaid beneficiaries to substitute health insurance coverage with cash transfers.

Then we remove the one-time policy change assumption and allow for a *full adjustment* to policy changes. We show that although the cash-out option remains an effective tool to reduce medical spending, there is a change in the composition of Medicaid beneficiaries because the cash-out option lowers the target efficiency of Medicaid: health insurance is most valued by the unhealthy, whereas cash transfers are valued by everyone. In the last part of this section, we show that this problem can be addressed by implementing work-dependent cash transfers.

7.1 Observable medical need

In this section, we construct an experiment where the government can observe the medical need (η_t^h) of Medicaid beneficiaries. In this case, the government can fully insure individuals against their medical need shock by covering 100% of their non-discretionary spending and making them fully responsible for their discretionary spending.[28] In other words, individuals face the full price of their discretionary medical spending. Since we preserve revenue-neutrality, we allocate the Medicaid budget that remains after covering non-discretionary spending to all beneficiaries as lump-sum transfers. The size of these transfers is adjusted until the total welfare budget is the same as in the baseline economy.[29]

In this full information economy, the aggregate medical spending of all individuals younger than 65 years old represents 92.5% of the baseline level. Thus, if Medicaid beneficiaries face the full price of their discretionary medical consumption, they substantially reduce their medical spending.

We also compute the welfare effects of this experiment. Our welfare measure is the compensation equivalent variation (CEV), which is equal to the lump-sum compensation per year needed to make an individual indifferent between being born in the baseline economy and in the economy with observable medical need. We report this number as a percentage of average consumption in the baseline economy. A positive number means welfare gains.[30]

[28] More specifically, since the CRRA part of medical spending goes to negative infinity when medical spending exactly equals medical need, we assume that the government covers medical need plus \$900. The results are robust to changes in this parameter.

[29] We consider the total welfare budget as opposed to the Medicaid-only budget because many low-income people are eligible for both Medicaid and cash assistance, and major changes in the former program affect the budget of the latter program.

[30] We compute the CEV as follows. First, we compute the ex-ante welfare at age 25 in the baseline and the experimental cases, denoting them V^B and V^E, respectively. Then we give a lump-sum transfer x to

We find that the CEV is equal to 1.48%, meaning that individuals are substantially better off in the full information economy.

Individuals in our model are ex-ante heterogeneous by fixed productivity type. The ex-ante welfare computed by type shows that the welfare gains are not uniform: the CEV for individuals with low fixed productivity is 3.15% while for those with high fixed productivity it is -0.75%. The elimination of distortions and the shift from medical to non-medical consumption benefit low fixed productivity individuals because they have low income and a relatively high marginal utility of non-medical consumption. In contrast, the welfare loss incurred by individuals with high fixed productivity happens because given their high resources (i) they are prone to over-consume medical care once on Medicaid; (ii) they value cash transfers and the resulting increase in non-medical consumption less.

Another important result from the full information economy is that the lump-sum transfers used to balance the welfare budget ($3,267 per beneficiary) draw more people into Medicaid, with the percentage of beneficiaries increasing from 8.9% (baseline) to 14.4%. This happens because in-kind transfers (medical consumption) are primarily attractive to unhealthy people, whereas cash transfers are equally attractive to all. Because Medicaid eligibility depends on labor income, some individuals now choose to stop working so that they can enroll in the program and receive cash transfers.[31] We discuss how to address this problem in Section 7.4.

To abstract from the increase in the number of Medicaid beneficiaries, we also construct the full information experiment as a *one-time policy change*. More specifically, we assume that the government can only observe each individual's medical need for one period and that new enrollees into Medicaid are not allowed. The one-time policy change assumption allows us to fix the distribution of agents across the state variables, and thus identify a person who was enrolled in Medicaid in the baseline economy. In contrast, when we allow for the full adjustment, individuals change their behavior over the life-cycle and this changes the distribution of agents.

The result from the full information economy under a one-time policy change is reported in Row 1 of Table 3. Note that the reduction in aggregate medical spending (93.2% of the baseline level) is slightly less compared with the full adjustment case (92.5%), but the size of the lump-sum cash transfers is significantly larger ($5,254 versus $3,267). This happens because in the one-time policy change case new Medicaid enrollees are not allowed, so lump-

everyone in the baseline economy in each period, resolve the model, and recompute welfare V^B. We adjust x until $V^B = V^E$. The interpretation of x is how much a newborn in the baseline case has to be compensated to be indifferent between the baseline and experimental economies.

[31] See Nichols and Zeckhauser (1982) for a more detailed discussion of the effect of in-kind versus cash transfers on target efficiency. See Pashchenko and Porapakkarm (2017) for further discussion of the effect of Medicaid on labor supply decisions.

sum transfers are allocated to a smaller group of people.

7.2 Asymmetric information environment: one-time policy change

In the previous subsection, we showed that the discretionary medical consumption of Medicaid beneficiaries can be considerably reduced if they face the full price of their care. Our goal in this section is to explore how discretionary medical consumption can be reduced in asymmetric information settings, i.e., when medical need is unobservable. We first restrict our analysis to one-time policy changes.

We start by considering increases in the coinsurance rate: we consider experiments where the Medicaid program covers less of total medical spending and allocates the "saved" money as lump-sum transfers to all beneficiaries. We show that this policy has a limited effect on medical spending. Next, we use the insights from our theoretical model and show that much better outcomes can be achieved if Medicaid enrollees are allowed to choose between cash transfers and in-kind Medicaid benefits.

	Lump-sum transfers ($000)	Medical spending (% BS)
Baseline (BS)	-	100.0
1. Observable medical need	5.3	93.2
Increasing Medicaid coinsurance		
2. Medicaid covers 90%	2.0	97.8
3. Medicaid covers 80%	3.0	96.3
4. Medicaid covers 70%	3.9	95.3
5. Medicaid covers 60%	4.5	94.7
6. Medicaid covers 50%	5.1	94.3
7. Medicaid covers 40%	5.5	93.9

Table 3: The effects of increasing Medicaid coinsurance, *one-time policy change*.

Rows 2 to 7 of Table 3 show the aggregate medical spending of individuals younger than 65 years old when Medicaid coinsurance is increased. In each experiment, the increase in Medicaid coinsurance is compensated by lump-sum cash transfers equally distributed among all beneficiaries. The size of the transfers is adjusted so that the total welfare budget in each experiment is the same as in the baseline economy.

The results in Table 3 show that to achieve a reduction in the aggregate medical spending similar to the full information case, the Medicaid coinsurance rate must be drastically increased: the percentage of medical spending covered by Medicaid should decrease from 97% (baseline) to 40%. This is because the policy is applied uniformly to all beneficiaries

regardless of their medical need. Individuals with low medical need still pay less than full price for their medical care, consequently they have higher medical consumption than in the full information case. At the same time, individuals with high medical need have little room to decrease their medical spending.[32]

Next, we consider a more flexible policy. Two important theoretical implications from Section 3 are that in the second-best economy, (i) individuals with low medical consumption should be rewarded with high non-medical consumption, (ii) individuals with low medical need should not face distortions on their medical versus non-medical consumption choices. Following this intuition, we introduce an option for Medicaid beneficiaries to substitute their medical coverage with cash transfers. In other words, each beneficiary is given a choice to either enroll in the traditional Medicaid program (with the same coverage as in the baseline economy) or the cash subprogram. The latter provides beneficiaries lump-sum transfers but leaves them fully responsible for paying their medical consumption.[33]

This policy introduces a trade-off between obtaining higher non-medical consumption (by receiving cash transfers) at the cost of reducing medical consumption (forgoing insurance coverage). Obviously, the cash program is only attractive to people with low medical need. Importantly, once an individual chooses the cash option, he faces the full price of his medical care, i.e., his choice between non-medical and medical consumption is undistorted. As before, we adjust the size of the cash transfers so that the total welfare budget remains the same as in the baseline economy. Thus, the higher the reduction in medical spending by Medicaid beneficiaries, the larger the transfers will be to people who choose the cash-out option. As with all experiments in this subsection, we only consider one-time policy changes.

	Lump-sum transfers ($000)	Medical spending (% BS)
Baseline (BS)	-	100.0
1. Observable medical need	5.3	93.2
Increasing Medicaid coinsurance		
2. Medicaid covers 97%	2.8	96.9
3. Medicaid covers 90%	3.9	95.2
4. Medicaid covers 80%	5.2	93.9
5. Medicaid covers 70%	5.9	93.4

Table 4: The effects of introducing a cash-out option into Medicaid, *one-time policy change*.

[32]Table 7 in Appendix C reports the change in the mean and standard deviation of out-of-pocket medical spending as a result of these experiments.

[33]Note that individuals who choose the cash subprogram of Medicaid can still rely on other means-tested programs that guarantee a minimum subsistence level, so that they are not entirely uninsured against medical shocks.

Row 2 of Table 4 displays the results of this experiment. As a reference, we report the corresponding results from the full information case in Row 1. The aggregate medical spending of all non-elderly decreases by around 3%. These savings in medical costs are reallocated as a transfer of $2,800 per beneficiary to those who choose the cash-out option.

The introduction of the cash-out option creates a trade-off between non-medical and medical consumption, however, the experiment above shows that this trade-off is not strong enough to achieve the reduction in medical spending similar to the full information case. This is because the size of the cash transfers is too small. To reinforce this trade-off, in the next set of experiments we also increase the coinsurance of regular Medicaid coverage.

Rows 3 to 5 of Table 4 show that the cash-out option combined with an increase in Medicaid coinsurance is an effective tool to decrease aggregate medical spending: to reduce the spending to 93.4% of the baseline case, the percentage of medical expenses covered by Medicaid should be reduced to 70%. In comparison, to achieve a similar result without the cash-out option, we would need to decrease Medicaid coverage to 40% (Table 3). This is because decreased traditional (in-kind) Medicaid generosity in the presence of a cash-out option triggers a "virtuous spiral". As more people switch to the cash plan, thereby lowering the aggregate medical spending, more resources can be allocated as cash transfers. At the same time, higher cash transfers induce more people to move from the traditional (in-kind) Medicaid plan to the cash option, further reducing the aggregate medical spending. Note that transfers received by people who opt for the cash plan increase from $2,800 in the case when traditional Medicaid covers 97% of medical costs to $5,900 when this number is reduced to 70%.[34]

7.3 Asymmetric information environment: full policy adjustments

Table 5 shows the results of introducing the cash-out option when we allow for a full adjustment to this policy change. Row 1 of this table includes the full information case discussed in Section 7.1 as a benchmark for comparison.

As reported in the fourth column of Table 5, there are now more people enrolling in the Medicaid program. For example, Medicaid enrollment constitutes 9.4% and 14.1% for cases when traditional Medicaid covers 97% and 70% of medical spending, respectively (compared with the baseline enrollment of 8.9%). This is due to an inflow of healthy enrollees attracted by cash transfers.[35]

[34]Table 8 in Appendix C reports the change in the mean and standard deviation of out-of-pocket medical spending for these experiments. Table 9 reports additional statistics to illustrate sorting between the two Medicaid subprograms.

[35]The fifth and sixth columns of Table 5 report Medicaid enrollment by fixed productivity type. In all experiments, the majority of Medicaid beneficiaries are of low productivity type. Only around 1-1.5% of

	Medical spending (% BS)	Lump-sum transfers ($000)	Medicaid(%)[a]			Welfare (%)[b]		
			all	ξ^1	ξ^2	all	ξ^1	ξ^2
Baseline (BS)	100	-	8.9	16.4	1.3	-	-	-
1. Observable medical need	92.5	3.3	14.4	27.6	1.2	1.48	3.15	-0.75
Increasing Medicaid coinsurance								
2. Medicaid covers 97%	98.6	1.1	9.4	17.3	1.5	0.62	1.22	0.03
3. Medicaid covers 90%	95.5	2.5	11.9	22.2	1.5	1.32	2.68	-0.34
4. Medicaid covers 80%	94.2	3.0	13.4	25.5	1.3	1.10	2.51	-0.94
5. Medicaid covers 70%	93.5	3.3	14.1	27.0	1.2	0.62	1.86	-1.59

[a] Percentage of working- age people insured through Medicaid.

[b] Welfare is measured as dollars compensation and reported as a percentage of average consumption.

Table 5: The effects of introducing a cash-out option for Medicaid beneficiaries, *full policy adjustment*. ξ^1 denotes individuals with low fixed productivity and ξ^2 denotes individuals with high fixed productivity.

The reduction in aggregate medical spending with the full policy adjustment is similar to the one-time policy change.[36] However, due to the increase in the number of Medicaid enrollees, individuals in the cash subprogram receive significantly lower transfers. For example, when Medicaid covers 70% of medical spending, cash transfers are equal to $3,300 while this amount was $5,900 in the case of a one-time policy change.

The last three columns of Table 5 display the welfare effects of these experiments. The highest welfare gains (1.32%) are achieved when traditional Medicaid covers 90% of medical costs. In this case, every Medicaid beneficiary who chooses the cash option receives $2,500, and aggregate medical spending of the non-elderly constitutes 95.5% of the baseline level. Similar to the results in the full information case, the welfare gains are not uniform: individuals with low fixed productivity gain while those with high fixed productivity lose.

7.4 Medical consumption distortions and target efficiency

As shown in the previous subsection, the cash option is an important mechanism to mitigate distortions in the medical consumption of Medicaid beneficiaries. However, this policy has a drawback: it induces some people to stop working to gain Medicaid eligibility and receive cash transfers. The left panel of Figure (7) shows the fraction of people enrolled in Medicaid by age for the baseline case, and for the case when there is a cash-out option and traditional (in-kind) Medicaid covers 90% of medical spending (the case with the highest welfare).

individuals with high productivity type are Medicaid enrollees.

[36]The larger number of Medicaid beneficiaries does not lead to an increase in aggregate medical spending because most new enrollees are healthy and choose the cash option; thus, they face the full price of their medical care.

In the latter case, the number of people enrolled in Medicaid is higher than in the baseline at every age, but the largest difference is observed among the young. This group has relatively low productivity, so their opportunity costs of not working are not as high as middle-aged people. Moreover, the fact that they have not accumulated much assets makes it easy for them to meet the asset testing requirement of the Medicaid program. As a result, the employment rate among relatively young Medicaid beneficiaries decreases considerably. Among beneficiaries aged 25 to 29 years old, 49.1% work in the baseline economy but only 33.7% work in the economy where there is a cash-out option and traditional Medicaid covers 90% of medical expenses.

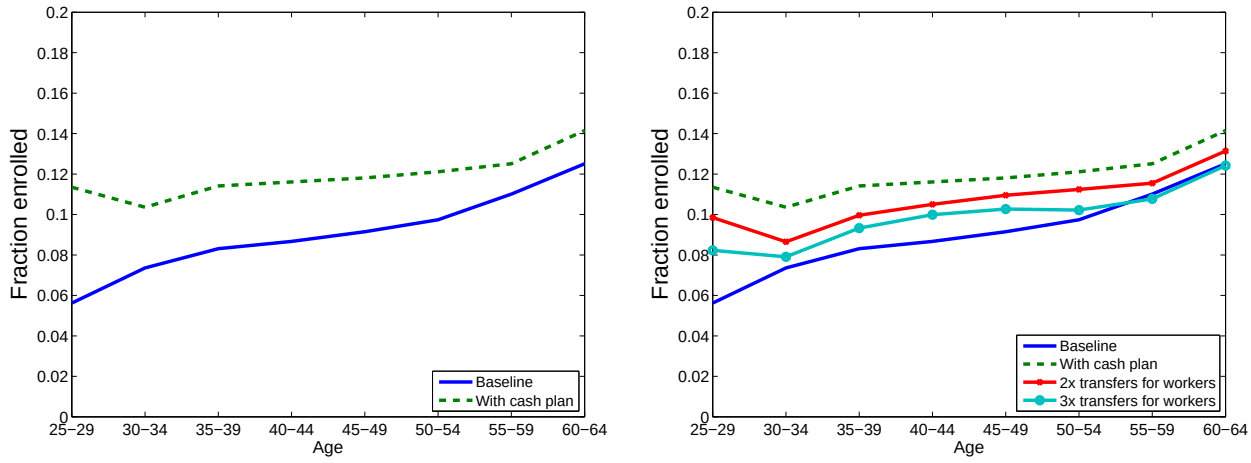

Figure 7: Fraction of people enrolled in Medicaid. Left panel: baseline economy and economy with two Medicaid subprograms: cash plan + traditional insurance that covers 90% of medical spending. Right panel: baseline economy and economies with two Medicaid subprograms: cash plan + traditional insurance that covers 90% of medical spending with three variations: i) cash plan offers *work-independent* transfers, ii) cash plan offers transfers to workers that are twice as high as those for non-workers, iii) cash plan offers transfers to workers that are three times as high as those for non-workers.

To alleviate labor supply distortions created by the cash-out option, we introduce work-dependent cash transfers: those beneficiaries who choose the cash option receive larger transfers if they are working. We increase cash transfers for workers by a factor of 2 and 3. The cash transfers are adjusted until the total welfare budget in each experiment is the same as in the baseline economy. We allow for a full adjustment to these policies. The results of these experiments are reported in Rows 6 and 7 of Table 6. To ease the comparison, Row 5 shows the corresponding results when the transfers are independent of working status.

Increasing the gap between cash transfers for workers and non-workers is effective in decreasing the number of Medicaid beneficiaries: when working Medicaid beneficiaries who choose the cash option receive three times more in transfers than non-working beneficiaries, the percentage of people enrolled in Medicaid decreases to 9.8%, which is closer to the baseline

	Medical spending (% BS)	Lump-sum transfers ($000)	Medicaid (%)[a]			Welfare (%)[b]		
			all	ξ^1	ξ^2	all	ξ^1	ξ^2
1. Baseline (BS)	100	-	8.9	16.4	1.3	-	-	-
Observable medical need								
2. Uniform transfers	92.5	3.3	14.4	27.6	1.2	1.48	3.15	-0.75
3. Workers get 2 times more	93.3	2.8	12.3	23.7	1.0	2.33	4.84	-0.86
4. Workers get 3 times more	93.8	2.4	10.6	20.4	0.7	2.70	5.62	-0.97
Unobservable medical need								
Cash option+Medicaid (90%)								
5. Uniform transfers	95.5	2.5	11.9	22.2	1.5	1.32	2.68	-0.34
6. Workers get 2 times more	95.9	2.0	10.7	20.0	1.3	1.93	3.87	-0.36
7. Workers get 3 times more	96.1	1.5	9.8	18.4	1.2	2.10	4.20	-0.38

[a] Percentage of working- age people insured through Medicaid.

[b] Welfare is measured as dollars compensation and reported as a percentage of average consumption.

Table 6: The effects of introducing work-dependent transfers into cash plans, *full policy adjustment*. ξ^1 denotes individuals with low fixed effects and ξ^2 denotes individuals with high fixed effects.

economy (8.9%). The right panel of Figure (7) shows that the age profile of the percentage of Medicaid enrollees is close to this profile in the baseline economy. The size of the resulting cash transfers is around $1,500 for non-workers and $4,500 for workers compared with around $2,500 in the case of uniform transfers.

Note that when cash transfers are work-dependent more people choose traditional Medicaid over the cash option or choose to work and disenroll from Medicaid. As a result, the aggregate medical expenses of working-age adults are slightly higher when cash transfers are work-dependent compared to the case of uniform transfers (96.1% of the baseline level in the former case versus 95.5% in the latter).

The policy with work-dependent transfers produces higher welfare gains: the CEV is 1.93% when workers receive transfers that are two times greater than those of non-workers, and 2.09% when workers receive transfers that are three times greater. In other words, reducing distortions on work incentives substantially improves the welfare effects of the cash-out policy.

As a reference, we also construct a full information benchmark with work-dependent cash transfers. Specifically, we construct an experiment similar to the one outlined in Section 7.1: medical need is observable and fully insured, but the remaining welfare budget is allocated as work-dependent transfers. Rows 3 and 4 of Table 6 show the results of this experiment when workers receive transfers that are two and three times greater than transfers to non-workers, respectively. Note that in the asymmetric information environment, a cash-out policy with work-dependent cash transfers captures a substantial portion of the gains in the

full information case with the same work-dependent transfers scheme.

Overall, these experiments illustrate that even when medical need is unobservable, it is possible to substantially decrease distortions in medical spending of the publicly insured; however, it is important to balance the trade-off between these distortions and distortions on labor supply.

8 Conclusion

In this paper, we study how to improve upon existing public health insurance policies in an asymmetric information environment, i.e., when the division of medical spending into discretionary and non-discretionary parts is unobservable. We construct a simple theoretical framework in the spirit of Mirrlees (1971), which allows us to derive two important properties of the optimal insurance contract. First, individuals who consume less medical care should be rewarded with more non-medical consumption. Second, the non-medical/medical consumption choice of individuals with the lowest medical need should not be distorted.

We evaluate the quantitative impact of this type of policy with application to the Medicaid program for non-elderly adults. More specifically, we construct and calibrate/estimate a rich structural life-cycle model that reflects important institutional features of the US health insurance system and captures key features of the data. We model medical spending as a combination of discretionary and non-discretionary components, and then identify the quantitative importance of each component by matching the difference in medical spending profiles between the privately insured and the uninsured, while controlling for the composition of these two groups.

We use this model, first, to construct the full information economy where the government can observe each individual's medical need, and therefore can fully insure non-discretionary medical spending and make individuals face the full price of their discretionary medical spending. We show that in this case, the aggregate medical spending of non-elderly adults substantially decreases.

Next, we show that to decrease distortions of medical consumption in the asymmetric information case, we need to introduce a cash-out option for Medicaid beneficiaries and slightly reduce the generosity of the traditional Medicaid benefits. One key feature of this policy is that it creates a trade-off between medical and non-medical consumption. Individuals with relatively low medical need prefer to increase their non-medical consumption by choosing the cash option and decreasing their discretionary medical spending.

Our analysis also highlights the important interaction between distortions of different policies. Specifically, the cash-out option of Medicaid alleviates the distortions on medical

consumption among the publicly insured but reduces the target efficiency of this program by increasing distortions on labor supply. This happens because while in-kind transfers (health issuance) are mostly attractive for the unhealthy, cash transfers are attractive to everyone. Some individuals choose to reduce their labor supply (and thus income) to meet the eligibility requirements for Medicaid. We show that this issue can be addressed by making transfers in the cash subprogram of Medicaid work-dependent.

References

[1] Ameriks, J., Briggs, J., Caplin, A., Shapiro, M., Tonetti, C., 2017. Long-term Care Utility and Late in Life Saving. NBER Working Paper 20973

[2] Aron-Dine, A., Einav, L., Finkelstein, A., 2013. The RAND Health Insurance Experiment, Three Decades Later. NBER Working Paper 18642.

[3] Arrow, K., 1963. Uncertainty and the Welfare Economics of Medical Care. American Economic Review, 53, pages 941-973.

[4] Attanasio, O., Kitao, S., Violante, G. L., 2011. Financing Medicare: A General Equilibrium Analysis, In Demography and the Economy, edited by J. Shoven, University of Chicago Press

[5] Baicker, K., Taubman, S., Allen, H., Bernstein, M., Gruber, J., Newhouse, J., Schneider, E., Wright, B., Zaslavsky, A., Finkelstein, A., 2013. The Oregon Experiment: Effects of Medicaid on Clinical Outcomes, New England Journal of Medicine, 368(18), pages 1713-1722.

[6] Besley, T., 1988. Optimal Reimbursement Health Insurance and the Theory of Ramsey Taxation. Journal of Health Economics, 7, pages 321-336.

[7] Blomqvist, A., 1997. Optimal Non-linear Health Insurance. Journal of Health Economics 16, pages 303-321.

[8] Buchanan, J., Keeler, E. B., Rolph, J. E., Holmer, M. R., 1991. Simulating Health Expenditures Under Alternative Insurance Plans. Management Science, 37, pages 1067-1089.

[9] Capatina, E., 2015. Life-cycle Effects of Health Risk. Journal of Monetary Economics, 74, pages 67-88.

[10] De Nardi, M., French, E., Jones, J., 2010. Why Do the Elderly Save? Journal of Political Economy, 118(1), pages 39-75.

[11] De Nardi, M., French, E., Jones, J., 2016. Medicaid Insurance in Old Age. American Economic Review, 106(11), pages 3480-3520.

[12] Feldman, R., Dowd, B., 1991. A New Estimate of the Welfare Loss of Excess Health Insurance. American Economic Review, 81(1), pages 297-301.

[13] Floden, M., 2008. A Note on the Accuracy of Markov-chain Approximations to Highly Persistent AR(1) Processes. Economic Letters 99(3), pages 516-520.

[14] Fonseca, R., Michaud, P., Galama, T., Kapteyn, A., 2009. On the Rise of Health Spending and Longevity. Working Papers 722, RAND Corporation

[15] French, E., 2005. The Effects of Health, Wealth, and Wages on Labor Supply and Retirement Behaviour. Review of Economic Studies, 72(2), pages 395-427.

[16] French, E., Jones, J. 2011. The Effects of Health Insurance and Self-Insurance on Retirement Behavior. Econometrica, 79(3), pages 693-732.

[17] Gouveia and Strauss, 1994. Effective Federal Individual Tax Functions: An Exploratory Empirical Analysis, National Tax Journal, 47(2), pages 317-339.

[18] Grossman, M., 1972. On the Concept of Health Capital and the Demand for Health. Journal of Political Economy, 80, pages 223-255.

[19] Heathcote, J., Storesletten, K., Violante, G., 2010 The Macroeconomic Implications of Rising Wage Inequality in the United States. Journal of Political Economy, 118(4), pages 681-722.

[20] Hubbard, R., Skinner, J., Zeldes, S., 1995. Precautionary Saving and Social Insurance. Journal of Political Economy, 103(2), pages 360-399.

[21] Jung, J., Tran, C., Chambers, M., 2017. Aging and Health Financing in the US: A General Equilibrium Analysis. European Economic Review, 100, pages 428-462.

[22] Jung, J., Tran, C., 2016. Market Inefficiency, Insurance Mandate and Welfare: U.S Health Care Reform 2010. Review of Economic Dynamics, 20, pages 132-159.

[23] Kaiser Family Foundation, 2002. Employer Health Benefits Survey.

[24] Keeler, E., Rolph, J., 1988. The Demand for Episodes of Treatment in the Health Insurance Experiment. Journal of Health Economics, 7(4), pages 337-367.

[25] Klein, K., Schwartz, S, 2008. State Efforts to cover Low-Income Adults Without Children. State Health Policy Monitor, National Academy for State Health Policy.

[26] Lieber, E., Lockwood, L., 2017. Targeting with In-kind Transfers: Evidence from Medicaid Home Care. Mimeo, University of Virginia.

[27] Manning, W., Marquis, S., 1996. Health insurance: The tradeoff between risk pooling and moral hazard. Journal of Health Economics, 15(5), pages 609-639.

[28] Manning, W., Newhouse, J., Duan, N. et al., 1987. Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. American Economic Review, 77(3), pages 257-277.

[29] Mirrlees J. A., 1971. An Exploration in the Theory of Optimum Income Taxation, Review of Economic Studies, 38(2), pages 175-208

[30] Nakajima, M., Telyukova, I. 2012. Home Equity in Retirement. Working paper, Federal Reserve Bank of Philladephia.

[31] Newhouse, J., and the Insurance Experiment Group, 1993. Free for All? Lessons from the RAND Health Insurance Experiment. Harvard University Press.

[32] Nichols, A., Zeckhauser, R., 1982. Targeting Transfers Through Restrictions on Recepients. American Economic Review Paper and Proceedings, 7(2), pages 372-377.

[33] OECD, 2006. Projecting OECD Health and Long-term Care Expenditures; What are the Main Drivers? Working Paper 477. Available at `https://www:oecd:org/eco/public-finance/36085940.pdf`.

[34] Ozkan, S., 2013. Preventive vs. Curative Medicine: A Macroeconomic Analysis of Health Care over the Life Cycle. Working paper.

[35] Pashchenko, S., Porapakkarm, P., 2013. Quantitative Analysis of Health Insurance Reform: Separating Regulation from Redistribution. Review of Economic Dynamics, 16, pages 383-404.

[36] Pashchenko, S., Porapakkarm, P., 2016. Meical Spending in the US: Facts from the Medical Expenditure Panel Survey Dataset. Fiscal Studies, 37(3-4), pages 689-716.

[37] Pashchenko, S., Porapakkarm, P., 2017. Work Incentives of Medicaid Beneficiaries and the Role of Asset Testing. International Economic Review, 58(4), pages 1117-1154.

[38] Ringel, J. S., Hosek, S. D., Vollaard, B. A., Mahnovski, S., 2002. The Elasticity of Demand for Health Care: A Review of the Literature and Its Application to the Military Health System. RAND Corporation. Available at `https://www.rand.org/pubs/monograph_reports/MR1355.html`.

[39] Scholz, J., Seshadri, A., 2010. Health and Wealth in a Lifecycle Model. Mimeo, University of Wisconsin Madison.

[40] Spence, M., Zeckhauser, R., 1971. Insurance, Information, and Individual Action. American Economic Review, 61(2), pages 380-387.

[41] Storesletten, K., Telmer, C., Yaron, Y., 2004. Consumption and Risk Sharing Over the Life Cycle. Journal of Monetary Economics, 51(3), pages 609-633.

[42] Vera-Hernandez, M., 2003. Structural Estimation of a Principal-Agent Model: Moral Hazard in Medical Insurance. RAND Journal of Economics, 34(4), pages 670-693.

[43] Zweifel, P., Manning, W., 2000. Moral hazard and consumer incentives in health care. In Handbook of Health Economics, edited by Culyer, A. and Newhouse, J., Elsevier.

Appendix

A Implementation of the optimal allocation

The example of the implementation of the optimal allocation suggested in Section 3 works if $T_1 = c_L^* + m_L^* \geq c_H^* + q_2(m_H^*)m_H^* = T_2$. To satisfy this condition, we need to put some parametric restrictions on the problem. We consider the following parametrization:

$$u(x) = v(x) = \begin{cases} \dfrac{x^{1-\sigma}}{1-\sigma} & ; \text{ if } \sigma > 1 \\ \log(x) & ; \text{ if } \sigma = 1 \end{cases}$$

i.e., individuals' preferences over non-medical and discretionary medical consumption can be described by the CRRA (or log) function. In addition, we assume the medical need of the L-type is zero, i.e., $\eta_L = 0$.

We introduce the following notation:

$$\gamma = \frac{c_H^*}{c_L^*}$$
$$\alpha = \frac{m_H^*}{c_L^*}$$

Note that because $c_H^* < c_L^*$ (see Section 3), we have $\gamma \leq 1$. Also, because $u(\cdot) = v(\cdot)$, from Eq.(8) we have that $c_L^* = m_L^*$. Because $m_H^* > m_L^* = c_L^*$, we have $\alpha \geq 1$. Expressing m_L^*, c_H^* and m_H^* in terms of c_L^*, we can write the ICC (Eq.(3)) as follows:

$$\left\{ \begin{array}{ll} 2\dfrac{c_L^{*1-\sigma}}{1-\sigma} = \dfrac{(\gamma c_L^*)^{1-\sigma}}{1-\sigma} + \dfrac{(\alpha c_L^*)^{1-\sigma}}{1-\sigma} & ; \text{ if } \sigma > 1 \\ 2\log(c_L^*) = \log(\gamma c_L^*) + \log(\alpha c_L^*) & ; \text{ if } \sigma = 1 \end{array} \right\}$$

or

$$\left\{ \begin{array}{ll} \alpha = (2 - \gamma^{1-\sigma})^{1/(1-\sigma)} & ; \text{ if } \sigma > 1 \\ \alpha = \gamma^{-1} & ; \text{ if } \sigma = 1 \end{array} \right\}$$

Note that in the case of $\sigma > 1$, $\gamma^{1-\sigma} < 2$ or $\gamma > 2^{1/(1-\sigma)}$. We can rewrite the expression for wedge q in terms of α and γ as follows:

$$q = \left\{ \begin{array}{ll} \dfrac{1 + \left(\frac{\alpha}{\gamma}\right)^{-\sigma} \pi(\gamma^{-\sigma} - 1)}{1 + \pi(\gamma^{-\sigma} - 1)} & ; \text{ if } \sigma > 1 \\ \dfrac{1 + \left(\frac{\alpha}{\gamma}\right)^{-1} \pi(\gamma^{-1} - 1)}{1 + \pi(\gamma^{-1} - 1)} & ; \text{ if } \sigma = 1 \end{array} \right\}$$

Next, we can rewrite the inequality of interest $c_L^* + m_L^* \geq c_H^* + q_2(m_H^*)m_H^*$ as follows:

$$\left\{ \begin{array}{ll} 2 \geq \gamma + \dfrac{1 + \left(\frac{\alpha}{\gamma}\right)^{-\sigma} \pi(\gamma^{-\sigma} - 1)}{1 + \pi(\gamma^{-\sigma} - 1)}\alpha & ; \text{ if } \sigma > 1, \\ 2 \geq \gamma + \dfrac{1 + \left(\frac{\alpha}{\gamma}\right)^{-1} \pi(\gamma^{-1} - 1)}{1 + \pi(\gamma^{-1} - 1)}\alpha & ; \text{ if } \sigma = 1, \end{array} \right\}$$

which can be rearranged as follows:

$$\left\{ \begin{array}{ll} \dfrac{2 - \gamma - (2 - \gamma^{1-\sigma})^{1/(1-\sigma)}}{2 - \gamma^{\sigma} - \gamma^{-\sigma}} \leq 2\pi & ; \text{ if } \sigma > 1 \\ \dfrac{2 - \gamma - \gamma^{-1}}{2 - \gamma - \gamma^{-1}} \leq 2\pi & ; \text{ if } \sigma = 1 \end{array} \right\}$$

Note that the inequality sign changes direction because we divide both sides by $2 - \gamma^{\sigma} - \gamma^{-\sigma}$ (or $2 - \gamma - \gamma^{-1}$ in the case of log-utility), which is negative.

For the case of log-utility ($\sigma = 1$), the condition $c_L^* + m_L^* \geq c_H^* + q_2(m_H^*)m_H^*$ is satisfied when $\pi > 1/2$, i.e., there are more healthy individuals (with low medical need) than unhealthy individuals. For a more general case ($\sigma > 1$), the expression $\dfrac{2 - \gamma - (2 - \gamma^{1-\sigma})^{1/(1-\sigma)}}{2 - \gamma^{\sigma} - \gamma^{-\sigma}}$ is less than one except for values of γ close to $2^{1/(1-\sigma)}$, which implies a very high value of α inconsistent with the resource constraint. Thus, for the CRRA function, the restriction on π that makes the condition $c_L^* + m_L^* \geq c_H^* + q_2(m_H^*)m_H^*$ hold is less strict than for the log-utility, i.e., it is true even for values of $\pi < 1/2$. Overall, when using the CRRA (or log) parametrization of the utility from non-medical and discretionary medical consumption, our implementation mechanism works provided that a large enough fraction of individuals are healthy (at least more than half).

B Alternative model of utility from medical consumption

In this section, we discuss the performance of the estimated/calibrated model where the utility over medical consumption takes the following form:

$$v(m, \eta) = \theta \frac{(m - \eta)^{1-\sigma}}{1 - \sigma},$$

where θ is greater than zero.

Apart from the functional form for the utility of medical consumption, the model is identical to the one described in the main text. We use the same approach to estimate/calibrate the parameters of the model as described in Section 5. The only exception is that now medical consumption does not have a saturation point, so instead of parameter Δ we need to estimate the multiplier θ. This parameter affects the marginal utility, and therefore determines the demand for discretionary medical consumption. We calibrate θ by targeting the difference in medical spending between the privately insured and the uninsured, the same moment we use to identify the parameter Δ in the main text.

Figures (8), (9), (10), and (11) compare the moments related to health insurance, employment, labor income and medical spending by health and insurance constructed from the data and the calibrated model. Overall, the alternative model can capture many salient features of the data, but it produces income and price elasticities that are too high.

The implied income elasticity of medical spending in the alternative model is 1.17, which is significantly higher than its empirical counterpart (0-0.2 as discussed in Section 4.2). Our baseline model produces an elasticity of 0.13. The income elasticity was computed in the same way as in Section 4.2.

As for the price elasticity, we cannot compute it in the same way as for the baseline model in Section 4.2. In that section to reproduce the setup of the RAND health insurance experiment, we compare medical spending between two experiments with universal health insurance that cover 100% and 75% of spending. In the alternative model discussed in this section, we cannot compute medical spending when health insurance provides full coverage. The consumer optimization problem does not have a solution because the marginal utility of medical consumption is always positive while the marginal costs are zero. To avoid this, we compare two health insurance schemes that cover 95% and 75% of medical spending. The resulting reduction in medical spending when moving from more to less generous coverage constitutes a 53% decrease, while the corresponding decrease is 16% in the RAND experiment and 18% in our baseline model. Because both price and income elasticities are important for our policy analysis, we chose the model with the CRRA and a saturation point in the main text.

Another possible way to model the utility from medical consumption is to use the CRRA function with a different risk aversion. Specifically, we could consider

$$v\left(m,\eta\right)=\frac{\left(m-\eta\right)^{1-\sigma^{M}}}{1-\sigma^{M}}, \tag{28}$$

where σ^M can be different from the risk aversion over non-medical consumption σ. The problem with this specification, however, is that σ^M is difficult to identify from the data.

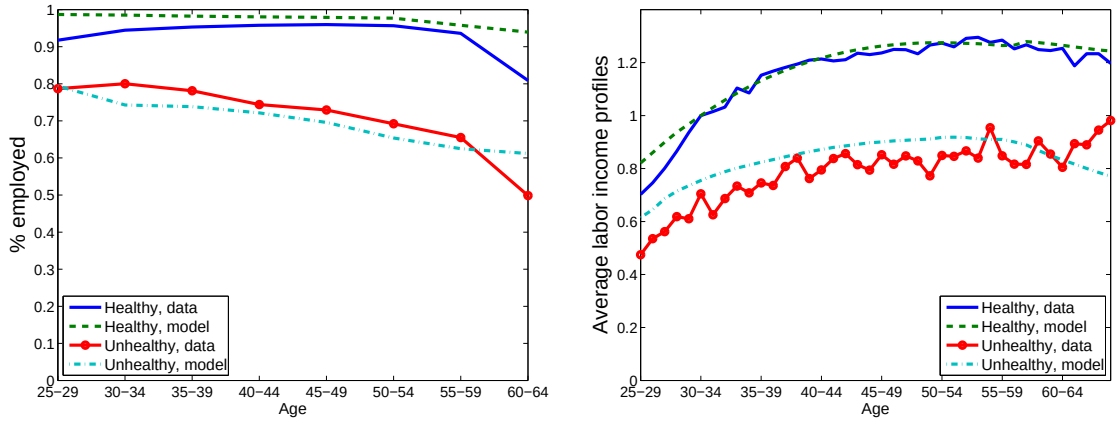

Figure 8: Left panel: employment by health. Right panel: average labor income among workers by health (normalized by average income). Solid line (line with round markers): data for the healthy (unhealthy) from the MEPS data. Dashed (dash-dotted) line: model for the healthy (unhealthy).

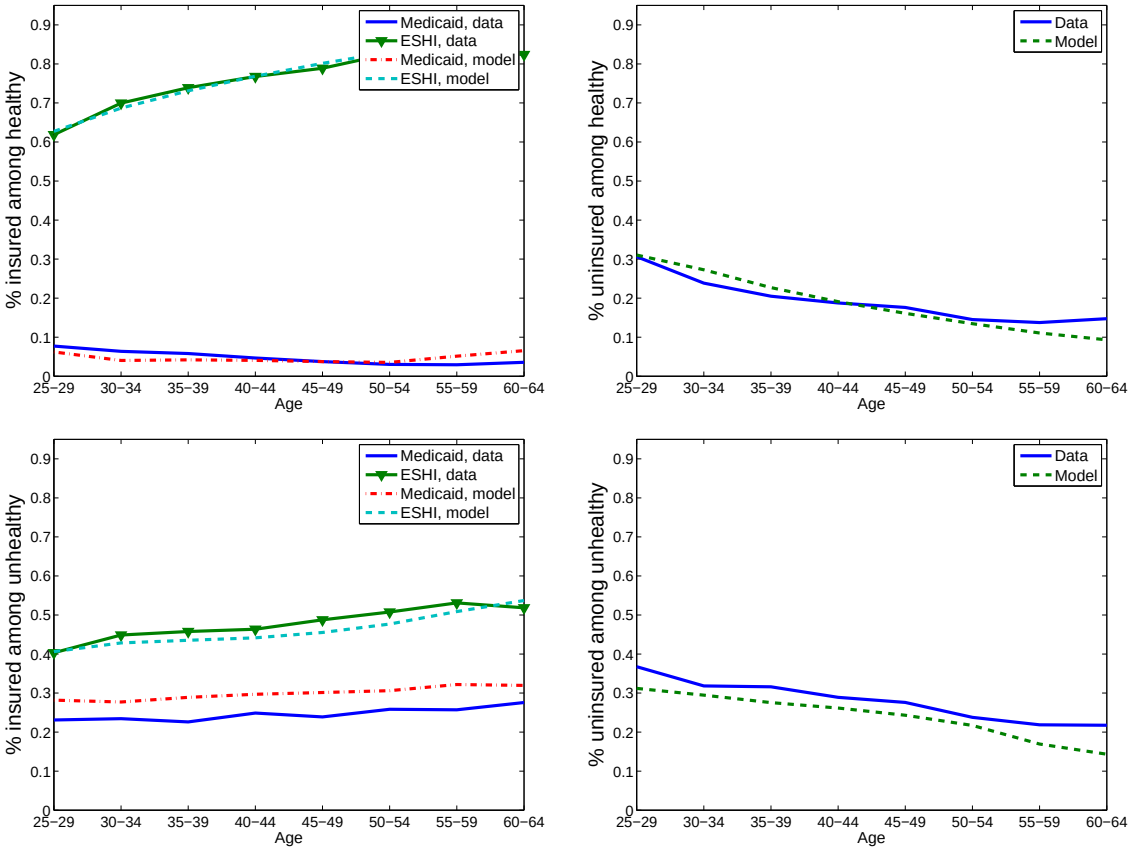

Figure 9: Insurance profiles for the healthy (top panel) and the unhealthy (bottom panel). Solid lines and lines with triangle markers are from the MEPS data. Dashed and dash-dotted lines are from the model. ESHI is "employers' sponsored insurance".

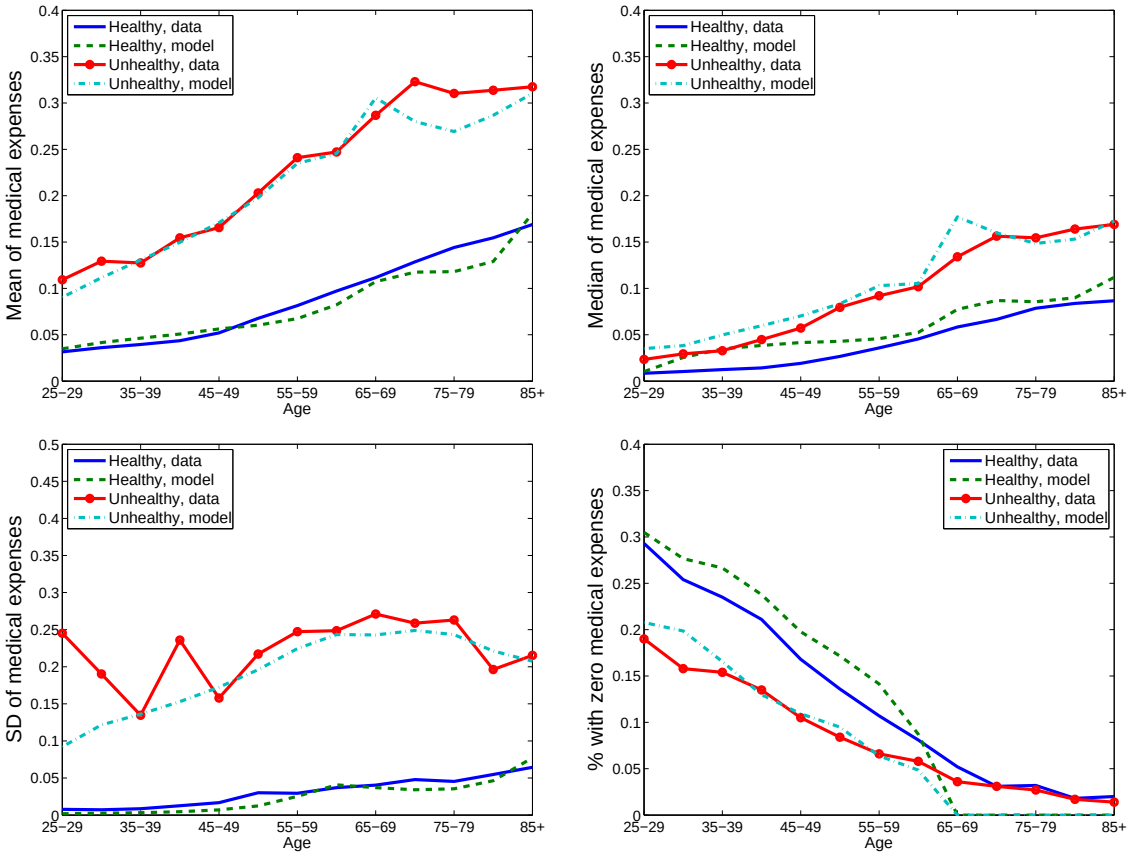

Figure 10: Top left panel: average medical expenses by health. Top right panel: median of medical expenses by health. Bottom left panel: standard deviation of medical expenses by health. Bottom right panel: fraction of people with zero medical expenses by health. Solid lines (lines with round markers) are from the MEPS data for the healthy (unhealthy). Dashed (dash-dotted) lines are from the model for the healthy (unhealthy). All level variables are normalized by average income.

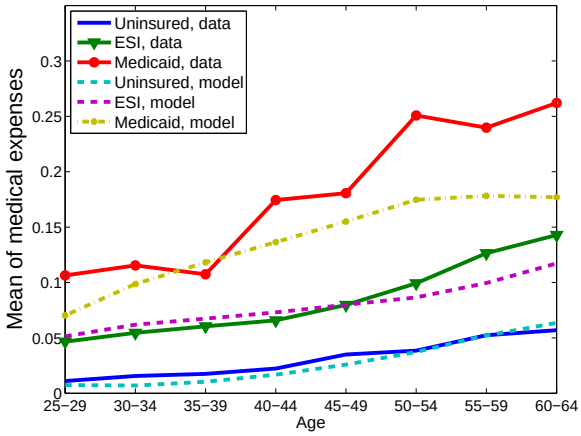

Figure 11: Average medical expenses by insurance status (normalized by average income). Solid lines (with or without markers) are from the MEPS data. Dashed and dash-dotted lines are from the model. ESI stands for employer-sponsored insurance.

We use the following example as an illustration.

Consider the following static problem of an individual with medical need η who allocates his endowment I between non-medical (c) and medical (m) consumption:

$$\underset{c,m}{max} \frac{c^{1-\sigma}}{1-\sigma} + v(m, \eta)$$

s.t.

$$c + m = I$$

The first-order condition for this problem when using the CRRA specification in Equation (28) can be written as:

$$(I - m)^{-\sigma} = (m - \eta)^{-\sigma^M}.$$

The utility from medical consumption depends on two parameters: η and σ^M. In our calibration, we want to match two moments: observed medical spending m^{obs} and the effect of insurance on medical spending. The latter depends on the fraction of non-discretionary spending in total medical spending, i.e., $\frac{\eta}{m}$. Thus, we only have one free parameter, σ^M, to match the observed spending. Note that from the first order condition it follows that an increase in σ^M can either increase or decrease total medical spending, which depends on the value of medical need η.[37] Because our model allows for heterogeneity in medical needs, the effect of changing σ^M on total medical spending is undetermined.

C Additional statistics for policy simulations (one-time policy change)

In Section 7.2 we compare the effects of two types of policies on total medical spending: a uniform reduction in Medicaid generosity versus a division of Medicaid into in-kind and in-cash subprograms. In this section, we consider the effects of these policies on out-of-pocket medical spending for all individuals younger than 65 years old. In the analysis below, we focus on the one-time policy change.

Table 7 reports the change in the mean and standard deviation of out-of-pocket medical spending when Medicaid generosity is uniformly reduced. Table 8 reports the same statistics when the cash-out option is introduced and the generosity of traditional (in-kind) Medicaid

[37] This can be shown by differentiating the first-order condition with respect to σ^M:

$$\frac{\partial m}{\partial \sigma^M} = -\frac{ln(m - \eta)}{\frac{\sigma^M}{m-\eta} + \frac{\sigma}{I-m}}.$$

The derivative can be positive or negative depending on whether $m - \eta$ is greater or less than one.

is reduced. Row 1 of each table reports the results for the full information case as a reference.

Both tables demonstrate a similar pattern: as Medicaid coinsurance increases, both the mean and standard deviation of out-of-pocket spending grow larger. As Tables 3 and 4 in the main text show, we can achieve the same reduction in total medical spending as in the full information case by either increasing the Medicaid coinsurance rate to 60% or by increasing it to 30% and introducing the cash-out option. In the former case the corresponding increase in the average out-of-pocket spending is the same as in the full information economy (20%), while in the latter case it is somewhat smaller (16%). The smaller increase in the average out-of-pocket spending in the case of the cash-out option occurs for the following reason. In the full information economy, all discretionary medical expenses are paid out-of-pocket while all necessary expenses are covered. In the case with a cash-out option, those who enroll in traditional Medicaid have only 70% of their necessary medical expenses covered (which increases their out-of-pocket spending compared with the full information case) but they are responsible for only 30% of their discretionary expenses (which decreases their out-of-pocket spending). The latter effect quantitatively exceeds the former.

As for the change in the standard deviation, the full information case and the case with a cash-out option and 30% Medicaid coinsurance rate are similar: the standard deviation increases by 7.6% and 6.7%, respectively. In the case where the Medicaid coinsurance rate is 60% and there is no cash-out option, the increase in the standard deviation is higher at 9.7%. This happens because people in this case are more exposed to the risk of high out-of-pocket medical expenses.

Table 9 provides additional statistics to better illustrate the sorting created by the division of Medicaid into the in-cash and in-kind subprograms. The second and third columns of the table report the average out-of-pocket spending separately for enrollees of each subprogram. Not surprisingly, people who opt out of traditional Medicaid have substantially lower out-of-pocket spending. For example, when traditional Medicaid covers 90% of medical costs its enrollees pay $1,177 out-of-pocket while this number is only $173 for the participants of the cash subprogram.

Column 4 of Table 9 illustrates the composition of enrollees in the traditional Medicaid subprogram. Specifically, it reports the percentage of enrollees who are in the bottom 75 percent of the medical need distribution among Medicaid beneficiaries. Note that since medical need represents unavoidable or necessary spending, the top 25% of its distribution can be roughly categorized as those with catastrophic expenses. Table 9 illustrates that as the generosity of traditional Medicaid decreases and the size of transfers in the cash subprogram increases, individuals with non-catastrophic medical spending switch to the cash plan. When traditional Medicaid requires coinsurance of 30% and the cash plan offers transfers of around

	Change in out-of-pocket spending (%BS)	
	Average	Standard deviation
Baseline (BS)	0.0	0.0
1. Observable medical need	20.2	7.6
Increasing Medicaid coinsurance		
2. Medicaid covers 90%	4.8	0.6
3. Medicaid covers 80%	9.6	2.4
4. Medicaid covers 70%	13.4	4.6
5. Medicaid covers 60%	16.3	6.4
6. Medicaid covers 50%	18.7	8.1
7. Medicaid covers 40%	20.5	9.7

Table 7: The effects of increasing Medicaid coinsurance on out-of-pocket medical spending, *one-time policy change*.

	Change in out-of-pocket spending (%BS)	
	Average	Standard deviation
Baseline (BS)	0.0	0.0
1. Observable medical need	20.2	7.6
Increasing Medicaid coinsurance		
2. Medicaid covers 97%	0.4	0.0
3. Medicaid covers 90%	4.8	1.0
4. Medicaid covers 80%	11.3	3.8
5. Medicaid covers 70%	16.2	6.7

Table 8: The effects of introducing a cash-out option on out-of-pocket medical spending, *one-time policy change*.

	Out-of-pocket spending ($)		$\eta_t^h < 75^{th}$ percentile (Traditional Medicaid)
	cash option	Traditional Medicaid	
Medicaid covers 97%	75	341	24%
Medicaid covers 90%	173	1,177	10%
Medicaid covers 80%	494	2,258	2%
Medicaid covers 70%	882	3,143	0%

Table 9: Characteristics of Medicaid beneficiaries when the cash-option is introduced, *one-time policy change*. The second and third columns show average out-of-pocket medical spending. The fourth column shows the fraction of enrollees in the in-kind Medicaid subprogram whose medical need (η_t^h) is below the 75^{th} percentile of the medical need distribution among Medicaid enrollees.

$6,000, the in-kind Medicaid subprogram is composed exclusively of people in the top 25 percent of the medical need distribution.