Economic Quarterly—Volume 104, Number 3—Third Quarter 2018—Pages 103–135

The Lifetime Medical Spending of Retirees

John Bailey Jones, Mariacristina De Nardi, Eric French, Rory McGee, and Justin Kirschner

Despite nearly universal enrollment in the Medicare program, most elderly Americans still face the risk of catastrophic health care expenses. There are many gaps in Medicare coverage: for example, Medicare does not pay for long hospital and nursing home stays and requires copayments for many medical goods and services. Medical spending is thus a major financial concern among elderly households. In a recent survey, affluent individuals were more worried about rising health care costs than about any other financial issue (Merrill Lynch Wealth Management 2012).

Several papers (De Nardi et al. 2010; Kopecky and Koreshkova 2014; Ameriks et al. 2015) show that health care costs that rise with age and income explain much of the US elderly's saving behavior.[1] Differences in medical spending risk are also important in explaining cross-country differences in the consumption (Banks et al. 2016) and

John Bailey Jones: Federal Reserve Bank of Richmond, jbjones.albany@gmail.com. Mariacristina De Nardi: Federal Reserve Bank of Chicago, UCL, CEPR, and NBER, e-mail: denardim@nber.org. Eric French: UCL, CEPR, and IFS, e-mail: eric.french.econ@gmail.com. Justin Kirschner: Federal Reserve Bank of Richmond, justin.kirschner@rich.frb.org. Rory McGee: UCL and IFS, rory.mcgee.13@ucl.ac.uk.
For helpful comments we thank Sara Ho, Marios Karabarbounis, Christian Matthes, Kerry Pechter, and John Weinberg. De Nardi and French gratefully acknowledge support from Norface Grant (TRISP 462-16-120). French gratefully acknowledges support from the Economic and Social Research Council (Centre for Microeconomic Analysis of Public Policy at the Institute for Fiscal Studies (RES-544-28-50001) and from Inequality and the Insurance Value of Transfers across the Life Cycle (ES/P001831/1)). The views expressed in this paper are those of the authors and not necessarily those of the Federal Reserve Bank of Chicago, the Federal Reserve Bank of Richmond, the IFS, the NBER, or the CEPR.

[1] Additional mechanisms proposed to explain the "elderly savings puzzle," or the slow decumulation of assets in old age, include bequests (De Nardi 2004; and Lockwood 2012) and the desire of older individuals to remain in their current homes (Nakajima and Telyukova 2012). See De Nardi et al. (2016b) for a review.

DOI: https://doi.org/10.21144/eq1040301

savings decisions (Nakajima and Telyukova 2018) of elderly households. More generally, the literature on the macroeconomic implications of health and medical spending is growing rapidly. Recent studies have considered important questions such as: bankruptcy (Livshits et al. 2007); the adequacy of savings at retirement (Skinner 2007; Scholz et al. 2006); annuitization (Pashchenko 2013; Lockwood 2012; Reichling and Smetters 2015); portfolio choice (Love 2010; Hugonnier et al. 2013); optimal taxation of health (Boerma and McGrattan 2018); and health insurance reform (Pashchenko and Porapakkarm 2013; Jung and Tran 2016; Conesa et al. 2018).

All of the aforementioned studies rely on accurate measures of medical risk and medical spending. But even though there is a large literature documenting annual medical spending at older ages, there has been relatively little work documenting the distribution of cumulative lifetime spending. Yet it is in many ways lifetime totals, rather than spending in any given year, that are most important for saving decisions and household financial well-being. The canonical permanent income hypothesis posits that forward-looking agents base their consumption not on their current income but on the average income they expect to receive over their lifetimes. The same logic applies to medical expenses. Households care not only about the risk of catastrophic expenses in a single year, but also about the risk of moderate but persistent expenses that accumulate into catastrophic lifetime costs.

In this paper, we estimate the distribution of lifetime medical spending for retired households whose heads are 70 or older. Our focus is out-of-pocket spending, the payments made by households themselves. High out-of-pocket expenses, however, can leave households financially indigent and reliant on Medicaid, the means-tested public insurance program. Medicaid eligibility depends on financial as well as health-related factors (De Nardi et al. 2012). Our benchmark spending estimates therefore include payments made by Medicaid to capture all of the medical spending risk that households potentially face. In accounting terms, our benchmark estimates measure the medical spending not covered by Medicare or supplemental private insurance, although they do include Medicare and supplemental private insurance premia. In economic terms, our estimates measure the medical spending risk that wealthier households would face and the medical spending risk that less wealthy households would face were Medicaid not available (absent any other changes in their insurance). We also consider an alternative measure of out-of-pocket spending that excludes Medicaid payments.

Our main dataset is the Health and Retirement Study (HRS), which has high-quality information on out-of-pocket medical spending over the period 1995 to 2014. Because the HRS does not have Medicaid

payment data, we impute Medicaid payments using the the Medicare Current Beneficiary Survey (MCBS). Ideally, these data would allow us to estimate medical spending directly by calculating discounted sums of household spending histories. Unfortunately, even the HRS, which has a very long panel dimension for a survey of its type, is not long enough to track all 70-year-olds through the ends of their lives. We thus resort to models.[2] Our data allow us to estimate dynamic models of health, mortality, and out-of-pocket medical spending. Medical spending depends on age, household composition, health, and idiosyncratic shocks. Simulating our estimated models allows us to construct household histories, calculate discounted sums, and ultimately compute the distribution of lifetime medical spending.

This paper uses the estimated model of De Nardi et al. (2018), which builds on earlier analyses of the HRS data by French and Jones (2004) and De Nardi et al. (2010, 2016a). French and Jones (2004) show that medical spending shocks are well described by the sum of a persistent AR(1) process and a white noise shock.[3] They also find that the innovations to this process can be modeled with a normal distribution that has been adjusted to capture the risk of catastrophic health care costs. Simulating this model, they find that in any given year 0.1 percent of households receive a health cost shock with a present value of at least $125,000 (in 1998 dollars). That paper abstracts away from much of the variability in costs coming from demographics or observable measures of health. De Nardi et al. (2010, 2016a) extend the spending model to account for health and lifetime earnings, but they consider only singles and do not control for end-of-life events (see French et al. [2006] and Poterba et al. [2017] on the importance of these events). The model used here addresses both shortcomings by including couples and singles and accounting for the additional medical expenditures incurred at the end of life.

Closely related papers include Fahle et al. (2016), who document the HRS medical spending data in some detail, and Hurd et al. (2017), who use the HRS to calculate the lifetime incidence and costs of nursing home services. Alemayehu and Warner (2004) construct a measure of lifetime spending by combining data from the MCBS and the Medical Expenditure Panel Survey with detailed data for Blue Cross Blue Shield members in Michigan. However, their estimates only distinguish gender, current age, and age of death—abstracting from health and marital status, among other factors—and for each of these groups only mean expenditures are estimated.

[2] Using a parametric model also improves our ability to measure tail risks.

[3] Feenberg and Skinner (1994) find a similar result. See also, Hirth et al. (2015).

Of particular note is Webb and Zhivan (2010), who use the HRS to estimate the distribution of lifetime expenses at ages 65 and above. Our paper is complementary to theirs but differs along two dimensions. The first is methodology. While both papers rely on simulation, our approach is to combine a three-state model of health (good, bad, or nursing home) with a two-component idiosyncratic shock and to control for socioeconomic status with a measure of permanent income (PI). In contrast, Webb and Zhivan (2010) estimate a rich model of stochastic morbidity and mortality with multiple health indicators and assume that medical expenditures are a function of these health conditions, along with a collection of socioeconomic indicators.[4] In their framework, all of the variation in medical spending is due to variation in these controls; there are no residual shocks. In contrast, in our framework, the idiosyncratic shocks capture any spending variation not attributable to age, PI, health, or household composition. The second major difference between our exercise and Webb and Zhivan's (2010) is the spending measure. As discussed above, the HRS data exclude expenses covered by Medicaid, which otherwise might have been paid out of pocket. Webb and Zhivan (2010) address this issue by excluding households that receive Medicaid. But all else equal, Medicaid beneficiaries tend to have higher medical expenses, in part because households that face overwhelming medical expenses are more likely to qualify for Medicaid (De Nardi et al. 2016a,c). Our approach is to impute the missing Medicaid expenditures, using data from the MCBS, and work with the sum of out-of-pocket and Medicaid expenditures. To put our results in context, we also analyze the HRS out-of-pocket spending measure. Comparing the two measures reveals the extent to which Medicaid reduces out-of-pocket expenditures.

We find that lifetime medical spending during retirement is high and uncertain. Households who turned 70 in 1992 will on average incur $122,000 in medical spending, including Medicaid payments, over their remaining lives. At the top tail, 5 percent of households will incur more than $300,000 and 1 percent of households will incur over $600,000 in medical spending inclusive of Medicaid. The level and the dispersion of remaining lifetime spending diminishes only slowly with age. The reason for this is that as they age, surviving individuals on average have fewer remaining years of life, but they are also more likely to live to extremely old age when medical spending is very high. Although PI, initial health, and initial marital status have large effects on this

[4] Using data from Catalonia, Carreras et al. (2013) perform a similar analysis.

spending, much of the dispersion in lifetime spending is due to events realized in later years.

We find that Medicaid lowers average lifetime expenditures by 20 percent. It covers the majority of the medical costs of the poorest households and significantly reduces their risk. Medicaid also reduces the level and volatility of medical spending for high-income households, but to a much smaller extent.

The rest of the paper is organized as follows. In Section 1, we discuss some key features of the datasets that we use in our analysis, the HRS and the MCBS, and describe how we construct our measure of medical spending. In Section 2, we introduce our model and describe our simulation methodology. We discuss our results in Section 3 and conclude in Section 4.

1. DATA

The medical spending models used here were developed and estimated as inputs for the structural savings model used in De Nardi et al. (2018). Our description of these models and the underlying data thus borrows heavily from the text of that project.

The HRS

We use data from the Asset and Health Dynamics Among the Oldest Old (AHEAD) cohorts of the HRS. The AHEAD is a sample of noninstitutionalized individuals aged 70 or older in 1993. These individuals were interviewed in late 1993/early 1994 and again in 1995/96, 1998, 2000, 2002, 2004, 2006, 2008, 2010, 2012, 2014, and 2016. We use data for ten waves, from 1995/96 to 2014. We exclude data from the 1994 wave because medical expenses are underreported (Rohwedder et al. 2006), and we exclude data from the 2016 wave because they are preliminary.

We only consider retired households, defined as those earning less than $3,000 in every wave. Because our demographic model allows for household composition changes only through death, we drop households that get married or divorced or report other marital transitions not consistent with the model. Consistency with the demographic model also leads us to drop households that: have large differences in ages; are same-sex couples; or have no information on the spouse. This leaves us with 4,324 households, of whom 1,249 are initially couples and 3,075 are singles.

Households are followed until both members die; attrition for other reasons is low. When the respondent for a household dies, in the next

wave an "exit" interview with a knowledgeable party—usually another family member—is conducted. This allows the HRS to collect data on end-of-life medical conditions and expenditures (including burial costs). Fahle et al. (2016) compare the medical spending data from the "core" and exit interviews in some detail.

The HRS has a variety of health indicators. We assign individuals to the nursing home state if they were in a nursing home at least 120 days since the last interview or if they spent at least sixty days in a nursing home before the next scheduled interview and died before that scheduled interview. We assign the remaining individuals a health status of "good" if their self-reported health is excellent, very good, or good and a health status of "bad" if their self-reported health is fair or poor.

The HRS collects data on all out-of-pocket medical expenses, including private insurance premia and nursing home care. The HRS medical spending measure is backward-looking: medical spending in any wave is measured as total out-of-pocket expenditures over the preceding two years. It is thus not immediately obvious whether medical spending reported in any given wave should be expressed as a function of medical conditions reported in that wave or those reported in the prior wave. Our empirical spending model includes indicators from both sets of dates. French et al. (2017) compare out-of-pocket medical spending data from the HRS, MCBS, and MEPS. They find that the HRS data match up well with data from the MCBS. They also find that the HRS matches up well with the MEPS for items that MEPS covers but that the HRS is more comprehensive than the MEPS in terms of the items covered.

To control for socioeconomic status, we construct a measure of lifetime earnings or "permanent income" (PI). We first find each household's "nonasset" income, a pension measure that includes Social Security benefits, defined benefit pension benefits, veterans benefits, and annuities. Because there is a roughly monotonic relationship between lifetime earnings and these pension variables, postretirement nonasset income is a good measure of lifetime permanent income. We then use fixed effects regression to convert nonasset income, which depends on age and household composition as well as lifetime earnings, to a scalar measure comparable across all households. In particular, we assume that the log of nonasset income for household i at age t follows

$$\ln y_{it} = \alpha_i + \kappa(t, f_{it}) + \omega_{it}, \tag{1}$$

where: α_i is a household-specific effect; $\kappa(t, f_{it})$ is a flexible function of age and family structure f_{it} (i.e., couple, single man, or single woman); and ω_{it} represents measurement error. The percentile ranks of the estimated fixed effects, $\widehat{\alpha}_i$, form our measure of permanent income,

$\widehat{I}_i$. Because we study retirees, in our simulations we treat $\widehat{I}_i$ as time-invariant.

The MCBS

While the HRS contains reasonably accurate measures of out-of-pocket medical spending, it does not contain Medicaid payments. To circumvent this issue, we use data from the 1996-2010 waves of the MCBS. The MCBS is a nationally representative survey of Medicare beneficiaries. Survey responses are matched to Medicare records, and medical expenditure data are created through a reconciliation process that combines survey information with Medicare administrative files. MCBS respondents are interviewed up to twelve times over a four-year period, resulting in medical spending panels that last up to three years. We use the same sample selection rules for the MCBS that we use for the HRS data.

The MCBS data include information on marital status, health, health care spending, and household income. One drawback of the MCBS is that it does not have information on the medical spending or health of the spouse.

Our Medical Spending Measure

Because the HRS medical spending data exclude expenses covered by Medicaid, which otherwise might have been paid out of pocket, they are censored. If the incidence of Medicaid were random, we could simply drop Medicaid recipients from our sample. However, this is not the case, because Medicaid beneficiaries tend to have higher medical expenses, in part because households that face overwhelming medical expenses are more likely to qualify for Medicaid (De Nardi et al. 2016a,c). Our approach is to use MCBS data to impute the missing Medicaid expenditures in the HRS and to then sum observed out-of-pocket and imputed Medicaid expenditures into a single cost measure. In addition to removing the censoring, our measure allows us to assess the spending risk that older households would face in the absence of Medicaid. Knowing this risk is key to assessing the effects of Medicaid itself.

We proceed in two steps. First, we use the MCBS data to regress Medicaid payments for Medicaid recipients on a set of observable variables found in both datasets. This regression has an R^2 statistic of 0.67, suggesting that our predictions are fairly accurate. Second, we impute Medicaid payments in the AHEAD data using a conditional mean-matching procedure, a procedure very similar to hot-decking. We combine the regression coefficients with the observables in the HRS to

predict Medicaid payments, then add to each predicted value a residual drawn from an MCBS household with a similar value of predicted medical spending. We describe our approach in more detail in the Appendix.

Although our principal spending measure is the sum of out-of-pocket and Medicaid payments, we also analyze out-of-pocket spending by itself. The extent to which out-of-pocket spending is lower and/or less volatile than combined spending directly reflects the extent to which Medicaid shields households from medical expenses.

2. THE MODEL

Our model of lifetime medical spending consists of two parts. The first is a Markov Chain model of health and mortality. The second part is the model of medical expenditure flows, where medical spending over any given interval depends on health, family structure, and the realizations of two idiosyncratic shocks.

Health and Mortality

Let hs_{it}^{h} and hs_{it}^{w} denote the health of, respectively, the husband h and the wife w in household i at age t. Each person's health status, hs^{g}, has four possible values: dead; in a nursing home; in bad health; or in good health. We assume that the transition probabilities for an individual's health depend on his or her current health, age, household composition, permanent income I, and gender $g \in \{h, w\}$.[5] It follows that the elements of the health transition matrix are given by

$$\pi_{i,j,k}(t, f_{it}, I_i, g) = \Pr\left(hs_{i,t+2}^{g} = k \,\middle|\, hs_{i,t}^{g} = j; t, f_{i,t}, I_i, g\right), \qquad (2)$$

with the transitions covering a two-year interval, as the HRS interviews every other year.[6] We estimate health/mortality transition probabilities by fitting the transitions observed in the HRS to a multinomial logit model.[7]

[5] We do not allow health transitions to depend on medical spending. The empirical evidence on whether medical spending improves health, especially at older ages, is surprisingly mixed (De Nardi et al. 2016a). Likely culprits include reverse causality—sick people have higher expenditures—and a lack of insurance variation—almost every retiree gets Medicare.

[6] As discussed in De Nardi et al. (2016a), one can fit annual models of health and medical spending to the HRS data. The process becomes significantly more involved, however, especially when accounting for the dynamics of two-person households.

[7] We do not control for cohort effects. Instead, our estimates are a combination of period (cross-sectional) and cohort probabilities. While our HRS sample covers eighteen years, it is still too short to track a single cohort over its entire postretirement lifespan. This may lead us to underestimate the lifespans expected by younger cohorts as

Table 1 Life Expectency in Years, Conditional on Reaching Age 70

	Men			Women		
Income Percentile	Nursing Home	Bad Health	Good Health	Nursing Home	Bad Health	Good Health
Single Individuals						
10	3.03	6.92	8.68	4.07	11.29	13.18
50	3.02	7.78	10.29	4.05	12.29	14.86
90	2.91	8.11	10.94	3.80	12.51	15.37
Married Individuals						
10	2.73	7.83	9.82	3.95	12.10	14.05
50	2.77	9.39	12.18	3.99	13.74	16.27
90	2.74	10.39	13.50	3.88	14.59	17.28

Table 1 shows the life expectancies implied by our demographic model for those still alive at age 70. The first panel of Table 1 shows the life expectancies for singles under different configurations of gender, PI percentile, and age-70 health. The healthy live longer than the sick, the rich (higher PI) live longer than the poor, and women live longer than men. For example, a single man at the 10th PI percentile in a nursing home expects to live only 3.0 more years, while a single woman at the 90th percentile in good health expects to live 15.4 more years. The second panel of Table 1 shows the same results for married men and women. Married people live longer than singles of the same health and PI unless they are in a nursing home, in which case the differences are small.[8]

Table 2 shows life expectancies for married *households*, that is, the average length of time that at least one member of the household is still alive or, equivalently, the life expectancies for the oldest survivors. While wives generally outlive husbands, a nontrivial fraction of the oldest survivors are men, and the life expectancy for a married household is roughly two years longer than that of a married woman.

Another key statistic for our analysis is the probability that a 70-year-old will spend significant time (more than 120 days) in a nursing home before he or she dies. Nursing home incidence differs relatively

they age. Nevertheless, lifespans have increased only modestly over the sample period. Accounting for cohort effects would have at most a modest effect on our estimates.

[8] The results for couples reported in Tables 1 and 2 are based on the assumption that the two spouses have the same health at age 70. While our model allows an individual's health transition probabilities to depend on his or her marital status, they do not depend on the spouse's health. Spousal health affects the life expectancy calculations only in that healthy spouses live longer.

Table 2 Life Expectancy of a Couple (Oldest Survivor) in Years, Conditional on Reaching Age 70

Income Percentile	Nursing Home	Bad Health	Good Health
10	4.51	13.94	15.93
50	4.61	15.91	18.40
90	4.50	16.85	19.41

modestly across the PI distribution. Although high-income people are less likely to be in a nursing home at any given age, they live longer, and older individuals are much more likely to be in a nursing home. In contrast, the effects of gender are pronounced, as are the effects of marital status for men. While 37 percent of single women and 36 percent of married women alive at age 70 will enter a nursing home before they die, the corresponding quantities for single and married men are 26 percent and 19 percent, respectively.[9] The differences between men and women are largely driven by the differences in life expectancy and marital status. Because women tend to live longer than men, they are more likely to live long enough to enter a nursing home. Moreover, although being married reduces the probability of entering a nursing home, wives tend to outlive their husbands. Women who are married at age 70 tend to be widows for several years, at which point they face the higher probability of entering a nursing home faced by women who are single at age 70. It is not surprising that the two groups face similar nursing home risk. In contrast, husbands usually die before their wives, so that men married at age 70 rarely face the high risk of transitioning into a nursing home faced by their single counterparts. Individuals initially in good health are 2 to 3 percentage points more likely to spend time in a nursing home than those initially in bad health, as nursing home risk is higher at older ages, and those initially in good health live longer.

Because all households in the HRS are initially noninstitutionalized, our estimates understate the fraction of individuals in nursing homes at any age. Our simulations begin with the second wave of the AHEAD cohort, at which point roughly 3 percent of men and 1 percent of women in the simulations had entered nursing homes. However, the HRS does a good job of tracking individuals as they enter in nursing homes. French

[9] These figures depend on the distribution of PI and initial health across men and women. We construct these distributions with bootstrap draws from the second wave of the AHEAD, using households whose heads were between 70 and 72 in the first wave.

and Jones (2004) show that by 2000 the HRS sample matches very well the aggregate statistics on the share of the elderly population in a nursing home. We also understate the number of nursing home visits because we exclude short-term visits: as Friedberg et al. (2014) and Hurd et al. (2017) document, many nursing home stays last only a few weeks and are associated with lower expenses. We focus only on the longer and more expensive stays faced by households.

Medical Spending

Our preferred medical spending measure is the sum of expenditures paid out-of-pocket plus those paid by Medicaid. Let m_{it} denote the expenses incurred between ages t and $t+2$. We observe the household' shealth at the beginning and the end of this interval, that is, at the time of the interview conducted at age t and at the time of the interview conducted at age $t+2$. Accordingly, we assume that medical expenses depend upon a household's PI, its family structure at both t and $t+2$, the health of its members at both dates, and the idiosyncratic component$\psi_{i,t+2}$:

$$\ln m_{it} = m(I_i, t+2, hs^h_{i,t}, hs^w_{i,t}, hs^h_{i,t+2}, hs^w_{i,t+2}, f_{i,t}, f_{i,t+2}) + \aleph_{i,t+2}, \quad (3)$$

$$\aleph_{i,t+2} = \sigma(I_i, t+2, hs^h_{i,t}, hs^w_{i,t}, hs^h_{i,t+2}, hs^w_{i,t+2}, f_{i,t}, f_{i,t+2}) \times \psi_{i,t+2}. \quad (4)$$

The variance of ψ is normalized to 1, so that $\sigma^2(\cdot)$ gives the conditional variance of $\aleph$. Including both family structure indicators allows us to account for the jump in medical spending that occurs in the period when a family member dies. Likewise, including health indicators from both periods allows us to distinguish persistent health episodes from transitory ones. Finally, we include cohort dummies in the regression.[10]

While we allow medical spending to depend on PI, we otherwise treat it as exogenous. This is a common assumption, and it allows us to calculate lifetime expenditures without solving a formal behavioral model. It bears noting that if households expect to make large medical expenditures near the ends of their lives, they will save for these expenses even when they are discretionary. On the other hand, the effects of policy reforms depend on the extent to which households can

[10] In particular, we regress log medical spending on a fourth-order age polynomial, indicators for single man (interacted with an age quadratic), single woman (interacted with an age polynomial), both the contemporaneous and lagged values of indicators for {man in bad health, man in a nursing home, woman in bad health, woman in a nursing home}, whether the man died (interacted with age and permanent income), whether the woman died (interacted with age, and permanent income), a quadratic in permanent income, and cohort dummies.

control their medical spending. De Nardi et al. (2010, 2016b) discuss this issue in some detail.

We estimate $m(\cdot)$ in two steps. In the first step, we regress log medical spending on the time-varying factors in equation (3), namely age, household structure, and health, using a fixed-effects estimator. Fixed effects regression cannot identify the effects of time-invariant factors, however, as they are not identified separately from the estimated fixed effects. To address this problem, we take the residuals from the first regression, inclusive of the estimated fixed effects, and regress them on the time-invariant factors, namely permanent income and a set of cohort dummies. In the simulations, we use the dummy coefficient for the cohort aged 72-76 in 1996. The level of $m(\cdot)$ is thus set to be consistent with the outcomes of this youngest cohort.

With the coefficients for $m(\cdot)$ in hand, we can back out the residual $\aleph$. A key feature of our spending model is that the conditional variance as well as the conditional mean of medical spending depends on demographic and socioeconomic factors, through the function $\sigma(\cdot)$ shown in equation (4). To find $\widehat{\sigma^2(\cdot)}$, we square the residuals (the $\aleph$s) from equation (3) and regress $\aleph^2$ on the demographic and socioeconomic variables in equation (4).

An accurate estimate of the lifetime medical expenditure distribution requires an accurate model of the intertemporal correlation of the idiosyncratic shock $\psi_{i,t}$. Following Feenberg and Skinner (1994) and French and Jones (2004), we assume that $\psi_{i,t}$ can be decomposed as

$$\psi_{i,t} = \zeta_{i,t} + \xi_{i,t}, \quad \xi_{i,t} \sim N(0, \sigma_\xi^2), \tag{5}$$

$$\zeta_{i,t} = \rho_m \zeta_{i,t-2} + \epsilon_{i,t}, \quad \epsilon_{i,t} \sim N(0, \sigma_\epsilon^2), \tag{6}$$

where $\xi_{i,t}$ and $\epsilon_{i,t}$ are serially and mutually independent. With the variance of $\psi_{i,t}$ normalized to 1, σ_ξ^2 can be interpreted as the fraction of idiosyncratic variance due to transitory shocks. We estimate the parameters of equations (5) and (6) using a standard error components method. Although the estimation procedure makes no assumptions on the the distribution of the error terms $\psi_{i,t}$, we assume normality in the simulations. French and Jones (2004) show that if the data are carefully constructed, normality captures well the far right tail of the medical spending distribution.[11]

[11] To help us match the distribution of medical spending, we bottom code medical spending at 10 percent of average medical spending. French and Jones (2004) also bottom code the data to match the far right tail of medical spending. Because we include Medicare B payments in our medical spending measure, which most elderly households pay, for the vast majority of households these bottom coding decisions are not important.

Approximately 40 percent of the cross-sectional variation in log medical spending is explained by the observables, which are quite persistent. Of the remaining cross-sectional variation, 40 percent comes from the persistent shock ζ and 60 percent from the transitory shock ξ. In keeping with the results in Feenberg and Skinner (1994), French and Jones (2004), and De Nardi et al. (2010), we estimate substantial persistence in the persistent component, with $\rho_m = 0.85$.

Quantitative Approach

After estimating our model, we assess its implications through a series of Monte Carlo exercises. The simulations begin at age 72 (reflecting medical spending between 70 and 72) and end at age 102.[12] Each simulated household receives a bootstrap draw of PI, initial health, and initial marital status from the HRS data used to estimate the health and spending models.[13] The household also receives initial values of ζ and ξ drawn from their unconditional distributions. We then use our Markov Chain model of health and mortality (equation (2)) to simulate demographic histories for each household and give each household a sequence of idiosyncratic shocks consistent with equations (5) and (6). Combining these inputs through equation (3) yields medical spending histories. We generate 1 million such histories and calculate summary statistics at each age.

3. RESULTS

Unconditional Spending Distributions

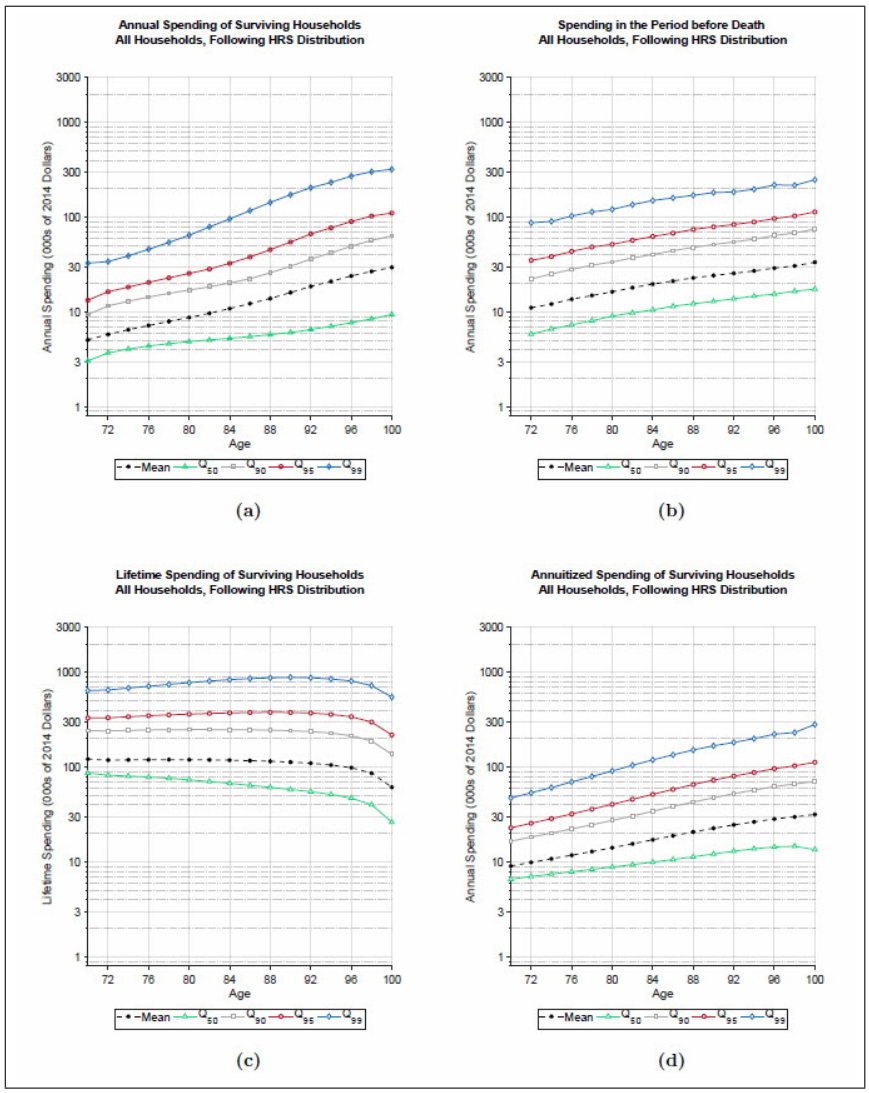
Figure 1 shows our model's implications for the cross-sectional distribution of our preferred medical spending measure, the sum of costs paid either out of pocket or by Medicaid. Costs are expressed in 2014 dollars.

Figure 1a, in the upper left corner, summarizes the health care expenditures of surviving households. Mean and median expenditures are shown, along with the 90th, 95th, and 99th percentiles. The results are dated by the beginning of the spending interval: the numbers for age 72 describe the medical expenses incurred between ages 72 and 74

[12] In couples, wives are assumed to be three years younger than husbands—the data average—and are thus initially 69. Single women are assumed to be 72.

[13] Although the simulations begin at age 72, we take bootstrap draws from the set of people aged 72 to 74 in 1996. This gives us a larger pool of households to draw from, which should in turn improve the accuracy of our exercise. We set the initial values of lagged (age 70) health and marital status equal to their age-72 values, allowing us to calculate medical expenditures made between the ages of 70 and 72.

Figure 1 Unconditional Distribution of Annual and Lifetime Medical Expenditures. Figures Show Mean, 50th, 90th, 95th, and 99th Percentiles of the Distribution



by people alive at both dates. Expenditures are expressed in annual terms. The medical expenses of surviving households rise rapidly with age. For example, mean medical spending rises from $5,100 per year at

age 70 to $29,700 at age 100. The upper tail rises even more rapidly, with the 95th percentile increasing from $13,400 to $111,200.[14]

Figure 1b shows end-of-life costs, which include burial expenses; the results for age 72 describe the expenses incurred by households who die between ages 72 and 74.[15] On average, end-of-life medical expenses exceed those of survivors. Mean end-of-life expenses range from $11,000 at age 72 to $34,000 at age 100.

Figure 1c plots our main variable of interest, lifetime expenditures. At each age, we calculate the present discounted value of remaining medical expenditures from that age forward, using an annual real discount rate of 3 percent. These lifetime totals are considerable. At age 70, households will, on average, incur over $122,000 of medical expenditures over the remainder of their lives. The top 5 and 1 percent of spenders will incur spending in excess of $330,000 and $640,000, respectively. One might expect the lifetime totals to fall rapidly as households age and near the ends of their lives. This is not the case. A household alive at age 90 will on average spend more than $113,000 before they die. The 95th percentile of remaining lifetime spending is higher at age 90 than at age 70. The slow decline of lifetime costs is due mostly to the tendency of medical costs to rise with age. Households that live to older ages have shorter remaining lives but higher annual expenditure rates.

A number of papers have considered whether medical expenses rise with age generically or mostly because older people are more likely to incur end-of-life expenses: see the discussion in De Nardi et al. (2016c). In our spending model, both forces are present. The top row of Figure 1 shows that there is considerable age growth in the medical expenses of both survivors and the newly deceased. Nonetheless, except for the 99th percentile at the oldest ages, the end-of-life expenses shown in Figure 1b are larger than the expenses faced by survivors of the same age (Figure 1a).

Figure 1d presents the annuitized spending associated with these lifetime totals. For each household, we convert lifetime expenses into the constant spending flow that would, over that household's realized lifespan, have the same present value. The average annuity payment rises from $9,100 at age 70 to $31,800 at age 100. Comparing Figures

[14] In general, our estimated model matches well the distribution of medical spending found in the raw data. However, the model overstates the 99th percentile of the medical spending distribution after age 90. Given the low probability of having medical spending in the 99th percentile, along with the low probability of living much past age 90, this discrepancy should have only a modest impact on our estimated lifetime spending distribution.

[15] We define a dead household as one that has no members. Couples who become singles are classified as survivors.

1a and 1d shows that mean annuitized spending at age 70 is almost double mean current spending. This reflects the rapid growth in medical spending that occurs as households age. The 95th and 99th percentiles of the annuitized medical spending distribution are also higher than the corresponding percentiles of the current spending distribution. One might think that those who have high medical spending in the present will usually have lower medical spending in the future, leading annuitized spending, which is essentially an average, to be less dispersed than current spending. The wide variation in annuitized spending found in Figure 1d thus shows that medical spending is persistent and that those with high spending in the present are likely to have high spending in the future.

Lifetime Medical Spending Determinants and Medical Spending Risk

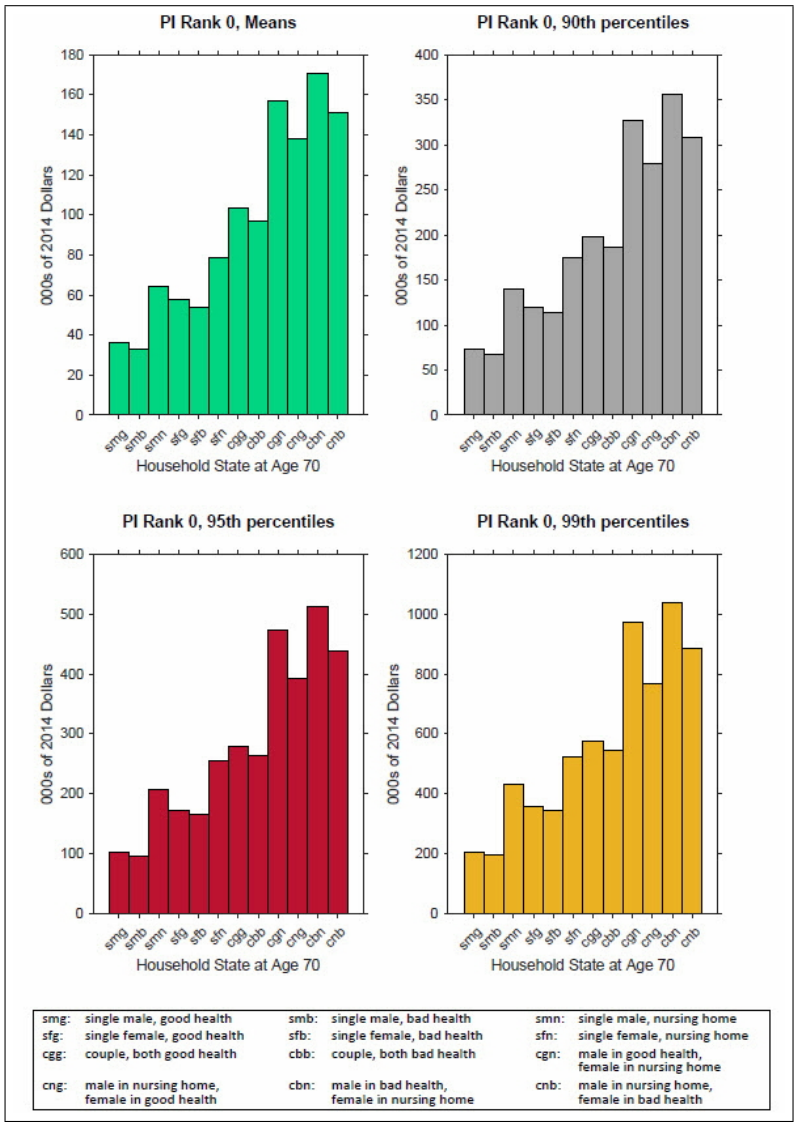
The graphs presented in Figure 1 show that the medical costs of older households are high, rising with age, and widely dispersed. A significant portion of this variation, however, is due to factors that are known to the household (PI, health, marital status, the persistent shock ζ). The spending distributions that individual households actually face, conditional on what they know at any point in time, can be quite different.

Figures 2 and 3 compare the mean, 90th, 95th, and 99th percentiles of lifetime medical spending at age 70 for different values of PI and initial health and marital status. Figure 2 shows results for households at the very bottom of the income distribution (PI = 0). Lifetime spending varies greatly across the distribution of initial health and marital status. Some trends are apparent:

1. Women have higher lifetime medical expenditures than men.
2. People initially in good health have higher lifetime expenditures than those initially in bad health. This reflects their longer life expectancies, combined with the tendency of medical costs to rise with age.
3. Households initially in nursing homes have the highest lifetime expenditures in spite of their high mortality. Nursing home care is expensive, and most people—more than 70 percent of men and 60 percent of women—outside a nursing home at age 70 never have an extended nursing home visit.

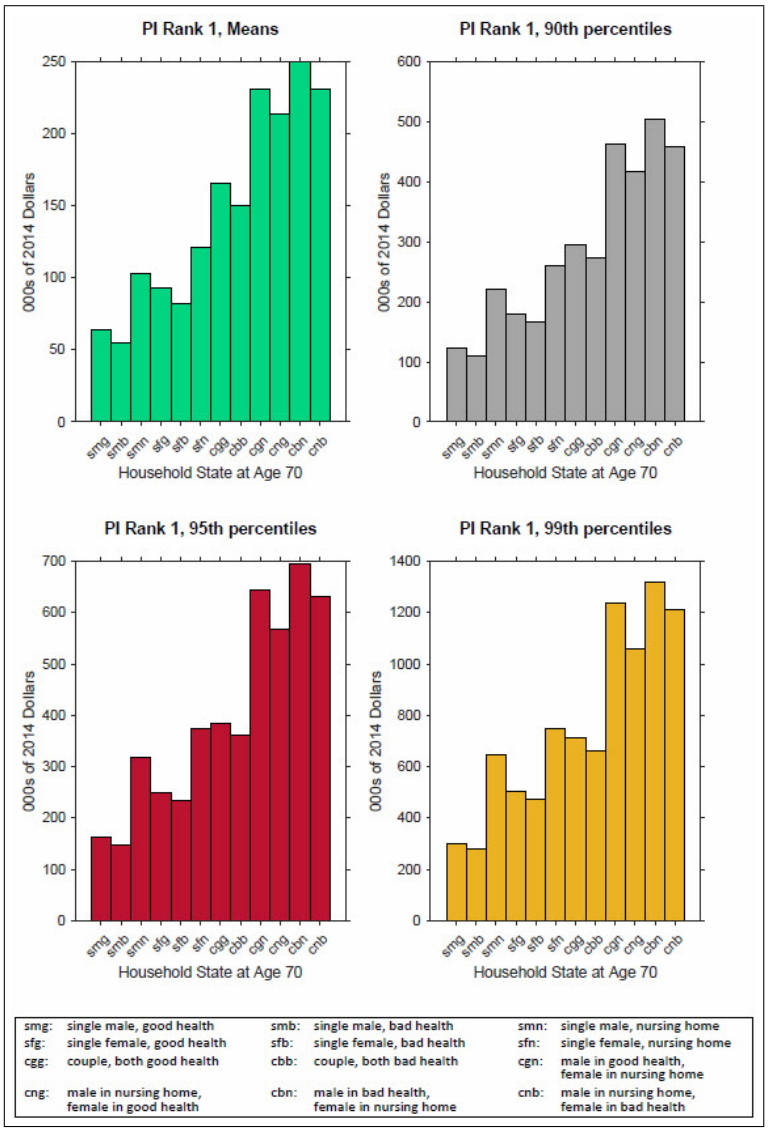
Figure 3 shows results for households at the very top of the income distribution (PI = 1). Households at the top of the income

Figure 2 Distributions of Lifetime Medical Expenditures by Initial Health and Household Structure, PI = 0



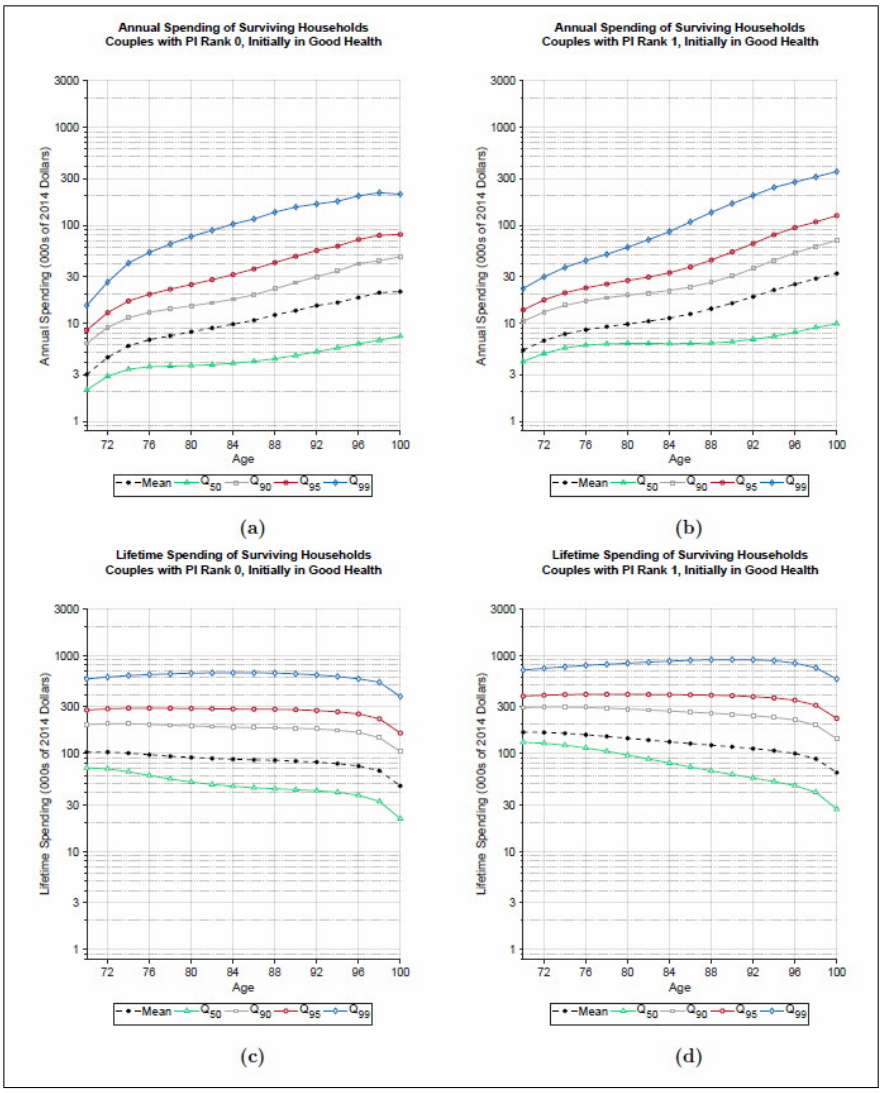
distribution spend considerably more than those at the bottom, often well in excess of 50 percent more. By way of example, consider 70-year-old couples where both members are initially in good health. With a PI rank of 0, these couples would, on average, spend $104,000 over their

Figure 3 Distributions of Lifetime Medical Expenditures by Initial Health and Household Structure, PI = 1



remaining lives. With a PI rank of 1, they would spend over $165,000. Households with higher income may have higher lifetime expenditures because they live longer or because they have higher expenses at any given age. Figure 4, which compares the same two groups in more

Figure 4 Annual and Lifetime Medical Expenses of Couples in Initial Good Health, with PI Ranks of 0 (Left Column) and 1 (Right Column)



detail, shows that both effects are present. The top two panels of this figure compare annual expenditures for surviving individuals. While high-income households typically spend more each year, at earlier ages and higher percentiles the opposite is often true.

Figures 2-3 show that a significant part of the dispersion in retiree medical spending can be attributed to health and demographic factors known at the very beginning of retirement. On the other hand, Figure 4 shows that spending remains dispersed even after conditioning on these factors. For example, the gaps between the conditional means and 99th percentiles of lifetime spending shown in Figures 4c and 4d are of roughly the same size as the unconditional gap shown in Figure 1c.

Another potentially predictable source of spending variation is the persistent idiosyncratic component of medical spending, ζ. The importance of the initial idiosyncratic shocks can be seen in Figure 5. The two panels in the left-hand column of this figure are directly comparable to the corresponding columns in Figure 4; the only difference is that the results in the new graphs are generated using a permanent income rank of 0.5. The two panels in the right-hand column differ from those on the left in that all the simulated histories begin with $\zeta = \xi = 0$. This can be seen in Figure 5b, where the distribution of annual expenses is initially degenerate. Comparing the panels in the top row shows that the effects of the shocks last for several years. On the other hand, the bottom panels show that eliminating the initial spending shocks has a relatively small effect on the dispersion of lifetime expenditures. Knowing the initial idiosyncratic shocks removes little risk.

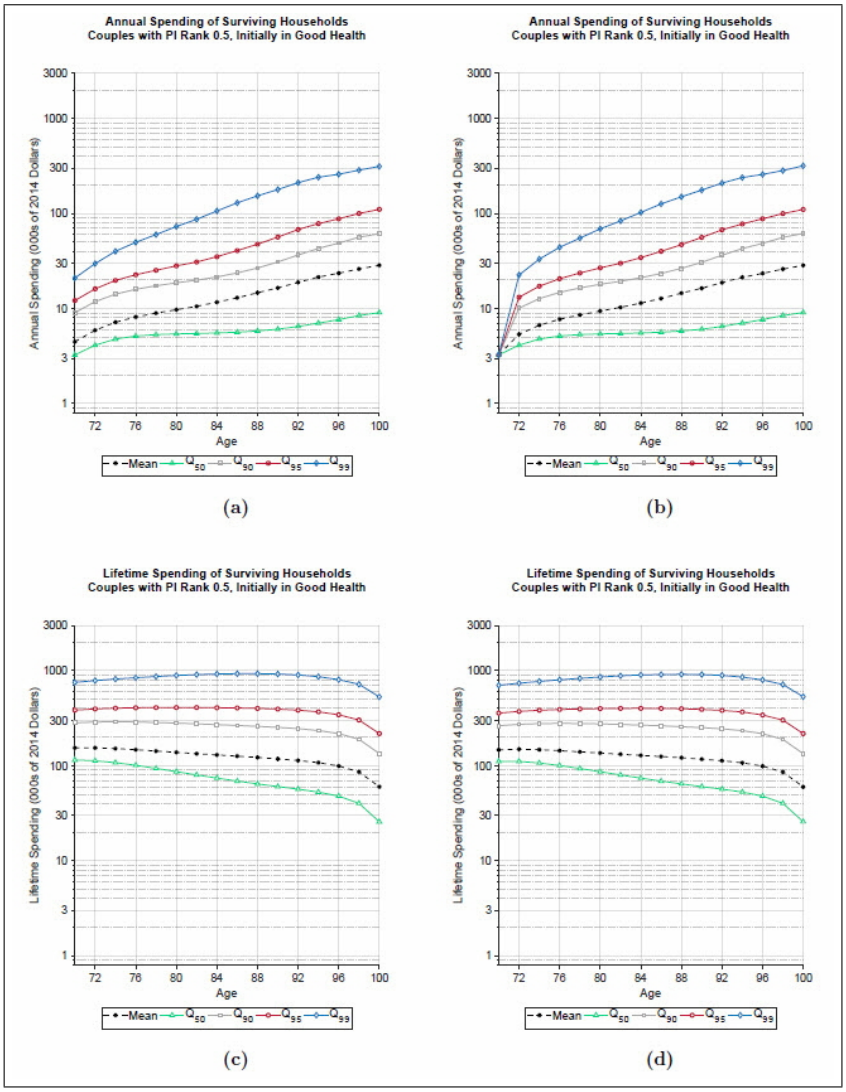
There are three reasons why the effects of the initial shocks wear off. First, as we document above, a significant portion of the variation in annual medical spending is driven by the health status of the household. Second, the transitory component ξ accounts for a large fraction of the residual variation, and imposing an initial condition has no effect on future transitory shocks. Finally, the effect of the initial realization of the persistent component ζ declines with age, as the persistence parameter ρ_m is less than 1.

The previous results notwithstanding, a large part of the elderly's medical spending uncertainty is due to the idiosyncratic shocks. Recall that only about 40 percent of the cross-sectional variation in log medical spending is explained by the observables. Likewise, if we remove all of the idiosyncratic shocks in our simulations, so that the only uncertainty is health and household structure, the unconditional variation of lifetime medical spending is only a fraction of its original value.

Out-of-Pocket Medical Spending

Our baseline measure of medical spending is the sum of payments made out of pocket and Medicaid. A number of recent papers have argued that Medicaid significantly reduces the out-of-pocket spending risk faced by older households. Brown and Finkelstein (2008) con-

Figure 5 Annual and Lifetime Medical Expenses of Couples in Initial Good Health, with (Left Column) and without (Right Column) Initial Idiosyncratic Spending Shocks



clude that Medicaid crowds out private long-term care insurance for about two-thirds of the wealth distribution. De Nardi et al. (2016a) find that most single retirees, including those at the top of the income

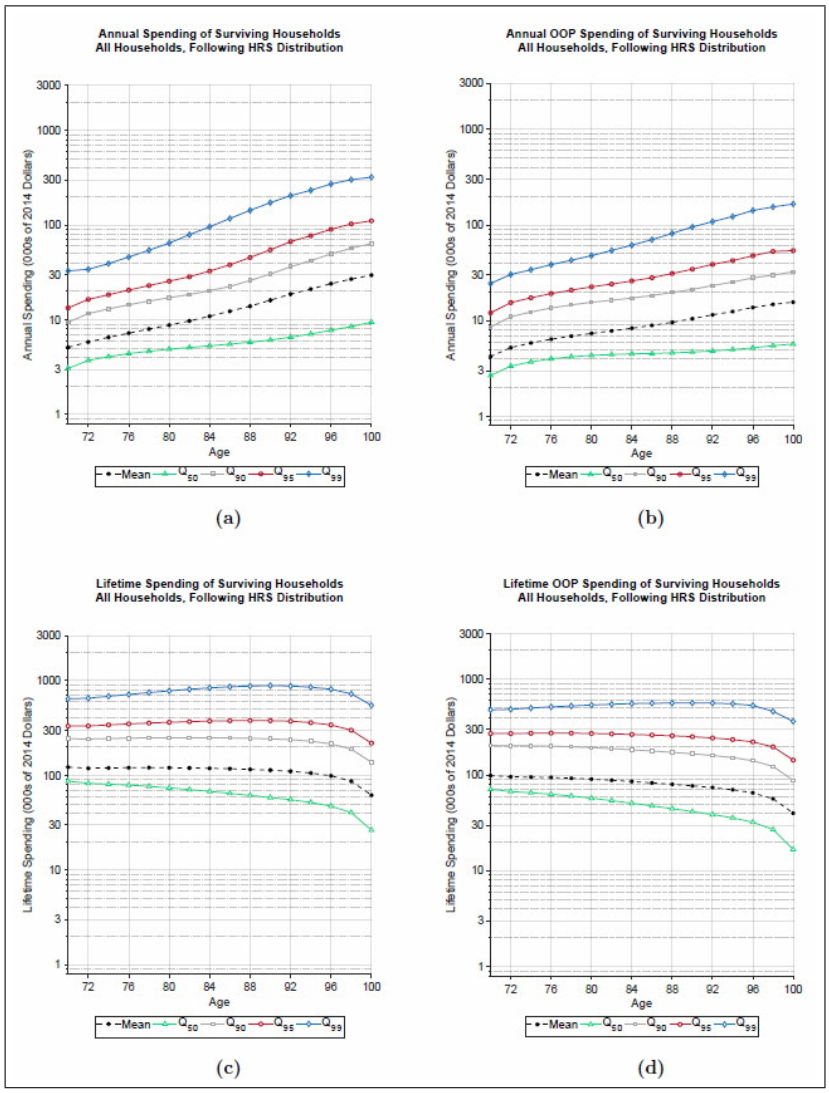
distribution, value Medicaid at more than its actuarial cost. While both of these papers model Medicaid formally, as part of the budget set in a dynamic structural model, it is also useful to assess the program in a less structured way. In particular, repeating our Monte Carlo exercises with the HRS out-of-pocket measure, which excludes Medicaid, allows us to compare the pre- and post-Medicaid distribution of medical spending.

Figure 6 compares unconditional distributions. The two panels in the left-hand column of this figure show results for our baseline spending measure; the panels in the right-hand column show results for out-of-pocket spending alone. The first row of Figure 6 compares annual expenditures for survivors. At age 70, mean out-of-pocket expenditures ($4,200) are about 18 percent less than mean combined expenditures ($5,100). In other words, Medicaid covers about 18 percent of the total for 70-year-olds. However, at older ages and higher spending percentiles, out-of-pocket expenditures are considerably lower. The second row of Figure 6 shows lifetime expenditures. At age 70, mean lifetime out-of-pocket expenses are about 20 percent lower than mean combined expenditures. This difference may seem small given the differences in the first row, but end-of-life expenditures (not shown) are fairly similar across the two spending measures.

Because Medicaid is means-tested, it is most prevalent at the bottom of the income distribution. To show this more clearly, Figure 7 compares the annual spending of surviving households at different points of the PI distribution. Consistent with Figures 4 and 5, we look at couples where both spouses were initially in good health. The top row of Figure 7, which compares the two spending measures for households at the bottom of the PI distribution, shows that Medicaid picks up a large share of these households' medical expenditures. At age 70, mean out-of-pocket expenditures are about 45 percent lower than mean combined expenditures, meaning that Medicaid constitutes about 45 percent of the total. The share of costs covered by Medicaid rises rapidly with age, however, to around 85 percent. The bottom row of Figure 7 repeats the comparison for the top of the PI distribution. Not surprisingly, Medicaid covers a much smaller fraction of these households' expenditures.

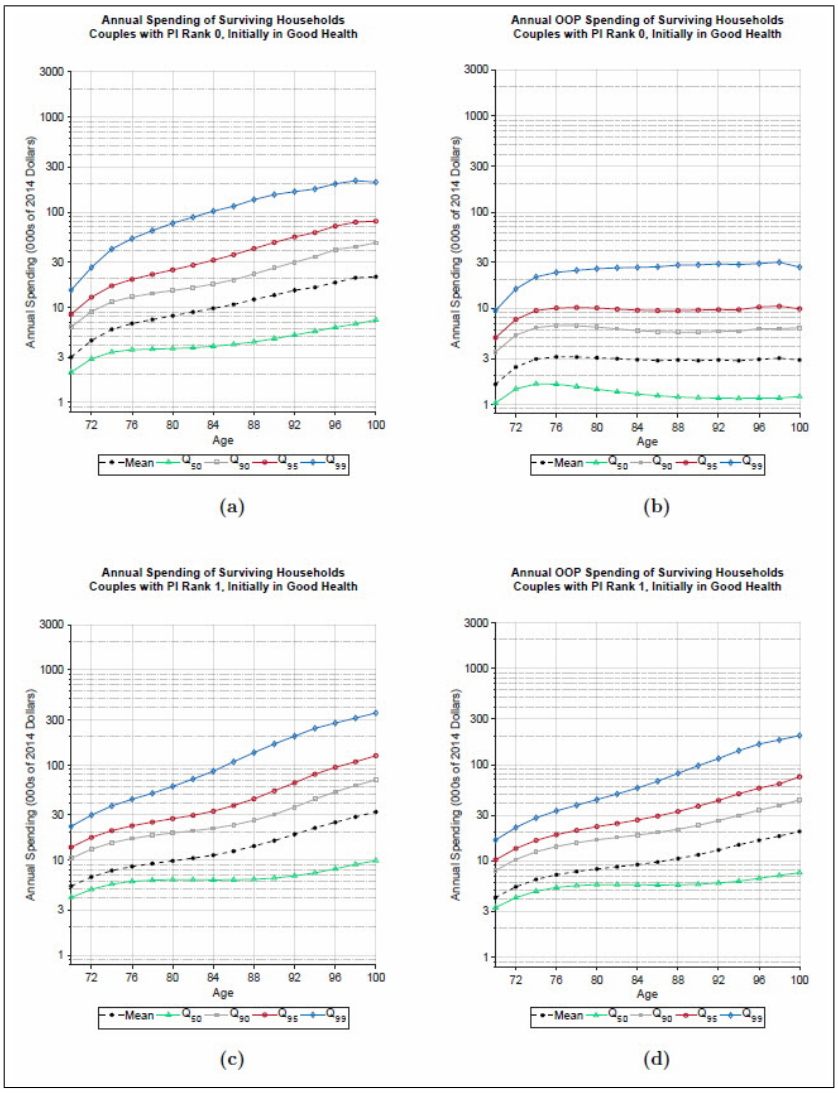
Figure 8 compares lifetime spending totals. The top row of this figure shows that at the bottom of the income distribution, Medicaid covers 57 percent of lifetime costs as of age 70. At older ages and higher percentiles, it covers even more. The bottom row shows results for households at the top of the income distribution. Medicaid covers 21 percent of lifetime costs at age 70, with the fraction rising to nearly 30 percent at age 100. While most high-income households do not receive Medicaid, those that do receive it qualify under the Medically

Figure 6 Unconditional Distribution of Annual and Lifetime Medical Expenditures, with (Left Panels) and without (Right Panels) Medicaid Payments



Needy provision, which assists households whose financial resources have been exhausted by medical expenses. Such households tend to have high medical expenses and tend to receive large Medicaid benefits (De Nardi et al. 2016a).

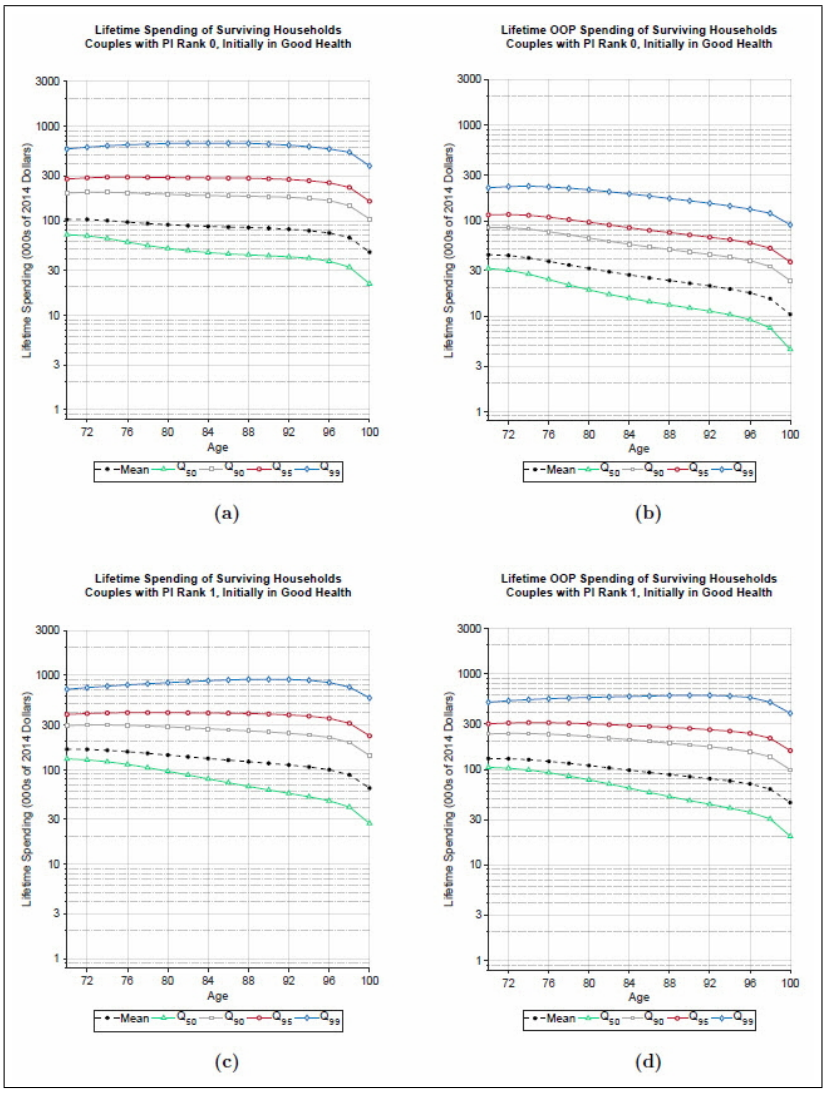
Figure 7 Annual Medical Expenses of Couples in Initial Good Health, with (Left Panels) and without (Right Panels) Medicaid Payments



4. DISCUSSION AND CONCLUSIONS

In this paper, we use the health and spending models developed in De Nardi et al. (2018) to simulate the distribution of lifetime medical expenditures as of age 70, adding to the handful of studies on this topic.

Figure 8 Lifetime Medical Expenses of Couples in Initial Good Health, with and without Medicaid Payments



We also assess the importance of Medicaid in reducing the lifetime medical spending risk. The simulations show that lifetime medical spending is high and uncertain and that the level and the dispersion of this spending diminish only slowly with age. Although PI, initial health, and initial marital status have large and predictable effects,

much of the dispersion in lifetime spending is due to events realized at older ages. The poorest households have the majority of their medical costs covered by Medicaid, which significantly reduces their spending volatility as well. Medicaid also reduces the level and volatility of medical spending for high-income households, albeit to a much smaller degree.

The paper closest to ours is Webb and Zhivan (2010), which we discussed in our introduction. Webb and Zhivan (2010) also find that lifetime out-of-pocket medical spending is high and widely dispersed and that the level and conditional dispersion of this spending diminish only slowly as households age. The levels of their estimated expenses, however, are even higher than ours, even though their spending measure excludes Medicaid. For instance, they find that 65-year-old couples with high school degrees and no chronic diseases will on average spend about $300,000 over their remaining lives, and 5 percent of these households will spend well over $600,000.[16] We find that 70-year-old couples with a PI rank of 0.5 and good initial health will on average spend about $150,000 over their remaining lives and that 5 percent of these households will spend in excess of $380,000.

One likely reason why Webb and Zhivan (2010) find higher medical spending is that they estimate their model using a pooled cross-section regression. They then correct for cohort bias—the fact that, for instance, the medical spending of a 90-year-old observed in 1996 is likely to be lower than the medical spending a 70-year-old observed in 1996 would face when she turned 90 in 2016—by allowing their simulated medical expenses to grow over time, independent of age, at long-term historical rates. In contrast, we estimate our spending model using a fixed effects regression with no time controls, so that our age effects measure the year-to-year spending growth that households realized over the sample period. Because medical spending growth has been fairly slow in recent years—and out-of-pocket spending was reduced by the introduction of Medicare Part D in 2006—Webb and Zhivan's (2010) assumed growth rates likely exceed recent experience.[17] A second, related, reason is that Webb and Zhivan's (2010) estimates are for the cohort turning 65 in 2009, while our results are for the cohort that

[16] We inflate their results from 2009 to 2014 dollars—roughly 10 percent—using the CPI.

[17] Webb and Zhivan (2010) assume that the real growth rate of per capita health costs exclusive of long-term care follows a stochastic process with a mean of 4.2 percent per year, consistent with the data for 1960-2007. They assume that long-term care costs grow by 1.1 percent per year. According to the National Health Expenditure Accounts (Center for Medicare and Medicaid Services 2018), between 2002 and 2012, per capita personal health care spending for those 65 and older grew at a real (CPI-deflated) rate of 0.94 percent per year.

turned 70 in 1992. This represents a more than twenty-year gap in birth dates, during which time medical spending rose at every age.

We conclude by pointing out some caveats to our analysis. We assume, as do many other empirical papers, that medical spending is exogenous, while in reality it is a choice variable. Although the demand for some medical goods and services is extremely inelastic, the demand for others might be elastic. Nursing home care, for example, is a bundle of medical and nonmedical commodities, and the latter can vary greatly in quality, with the choice between a single and a shared room being just one example. It is also worth noting that our analysis excludes payments made by Medicare and private insurers. Medicare substantially reduces out-of-pocket medical expenses throughout the retiree population (Barcellos and Jacobson 2015). While the combination of out-of-pocket and Medicaid expenditures considered here may be sufficient for some analyses, such as studies of household saving, other analyses require that all health costs be accounted for. Extending our exercise to include all medical expenditures would be useful, and we leave it to future research.

APPENDIX: IMPUTING MEDICAID EXPENDITURES

Let i index individuals in the HRS. Define oop_{it} as out-of-pocket medical expenses, Med_{it} as Medicaid payments, and m_{it} as the sum of out-of-pocket and Medicaid payments that we wish to plug in the model. To impute Med_{it}, which is missing in the HRS, we follow David et al. (1986) and French and Jones (2011) and use a predictive mean-matching regression approach. There are two steps to our procedure. First, we use the MCBS data to regress Medicaid payments (for Medicaid recipients) on observable variables that exist in both datasets. Second, we impute Medicaid payments in the HRS data using a conditional mean-matching procedure, a procedure very similar to hot-decking.

First Step Estimation Procedure

Let j index individuals in the MCBS. For the subsample of the MCBS with a positive Medicaid indicator (i.e., a Medicaid recipient), we regress the variable of interest, Med_{jt}, on the vector of observable variables z_{jt}, yielding $Med_{jt} = z_{jt}\beta + \varepsilon_{jt}$. We include in z_{jt} nursing home status, the number of nights spent in a nursing home, a fourth-order age polynomial, total household income, marital status, self-reported health, race, visiting a medical practitioner (doctor, hospital, or dentist), out-of-pocket medical spending, education, and death of an individual. Because the measure of medical spending in the HRS is medical spending over two years, we take two-year averages of the MCBS data to be consistent with the structure of the HRS. The regression of Med_{jt} on z_{jt} yields a R^2 statistic of 0.67, suggesting that our predictions are accurate.

Next, for every observation in the MCBS subsample we calculate the predicted value $\widehat{Med}_{jt} = z_{jt}\widehat{\beta}$ and the residual $\hat{\varepsilon}_{jt} = Med_{jt} - \widehat{Med}_{jt}$. We then sort the observations into deciles by predicted values, $\{\widehat{Med}_{jt}\}_{j,t}$, keeping track of the residuals, $\{\hat{\varepsilon}_{jt}\}_{j,t}$, as well.

Second Step Estimation Procedure

For every observation in the HRS sample with a positive Medicaid indicator, we impute $\widehat{Med}_{it} = z_{it}\widehat{\beta}$, using the values of $\widehat{\beta}$ estimated from the MCBS. Then, we impute ε_{it} for each observation of this subsample by finding a random observation in the MCBS with a value of $\widehat{Med}_{jt}$

in the same decile as $\widehat{Med}_{it}$, and setting $\hat{\varepsilon}_{it} = \hat{\varepsilon}_{jt}$. The imputed value of Med_{it} is $\widehat{Med}_{it} + \hat{\varepsilon}_{it}$.

As David et al. (1986) point out, our imputation approach is equivalent to hot-decking when the "z" variables are discretized and include a full set of interactions. The advantages of our approach over hot-decking are twofold. First, many of the "z" variables are continuous. Second, to improve goodness of fit we use a large number of "z" variables. Because hot-decking uses a full set of interactions, this would result in a large number of hot-decking cells relative to our sample size.

REFERENCES

Alemayehu, Berhanu, and Kenneth E. Warner. 2004. "The Lifetime Distribution of Health Care Costs." *Health Services Research* 39 (June), 627–42.

Ameriks, John, Joseph S. Briggs, Andrew Caplin, Matthew D. Shapiro, and Christopher Tonetti. 2015. "Long-Term-Care Utility and Late-in-Life Saving." Working Paper No. 20973. Cambridge, Mass.: National Bureau of Economic Research. (February).

Banks, James, Richard Blundell, Peter Levell, and James Smith. 2016. "Life-Cycle Consumption Patterns at Older Ages in the US and the UK: Can Medical Expenditures Explain the Difference?" Working Paper No. 22513. Cambridge, Mass.: National Bureau of Economic Research. (August).

Barcellos, Silvia Helena, and Mireille Jacobson. 2015. "The Effects of Medicare on Medical Expenditure Risk and Financial Strain." *American Economic Journal: Economic Policy* 7 (November): 41–70.

Boerma, Job, and Ellen McGrattan. 2018. "Health Capital Taxation." Mimeo.

Brown, Jeffrey, and Amy Finkelstein. 2008. "The Interaction of Public and Private Insurance: Medicaid and the Long-Term Care Insurance Market." *American Economic Review* 98 (June): 1083–1102.

Carreras, Marc, Pere Ibern, Jordi Coderch, Inma Sanchez, and Jose M. Inoriza. 2013. "Estimating Lifetime Healthcare Costs with Morbidity Data." *BMC Health Services Research* 13 (October): 1–11.

Center for Medicare and Medicaid Services. 2018. "National Health Expenditure Data." http://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/NationalHealthAccountsHistorical.html.

Conesa, Juan Carlos, Daniela Costa, Parisa Kamali, Timothy J. Kehoe, Vegard M. Nygard, Gajendran Raveendranathan, and Akshar Saxena. 2018. "Macroeconomic Effects of Medicare." *Journal of the Economics of Ageing* 11 (May): 27–40.

David, Martin, Roderick Little, Michael Samuhel, and Robert Triest. 1986. "Alternative Methods for CPS Income Imputation." *Journal of the American Statistical Association* 81 (March): 29–41.

De Nardi, Mariacristina. 2004. "Wealth Inequality and Intergenerational Links." *Review of Economic Studies* 71 (July): 743–68.

De Nardi, Mariacristina, Eric French, and John Bailey Jones. 2010. "Why Do the Elderly Save? The Role of Medical Expenses." *Journal of Political Economy* 118 (February): 39–75.

De Nardi, Mariacristina, Eric French, and John Bailey Jones. 2016a. "Medicaid Insurance in Old Age." *American Economic Review* 106 (November): 3480–520.

De Nardi, Mariacristina, Eric French, and John Bailey Jones. 2016b. "Savings After Retirement: A Survey." *Annual Review of Economics* 8 (October): 177–204.

De Nardi, Mariacristina, Eric French, John Bailey Jones, and Angshuman Gooptu. 2012. "Medicaid and the Elderly." Federal Reserve Bank of Chicago *Economic Perspectives* 36 (First Quarter), 17–34.

De Nardi, Mariacristina, Eric French, John Bailey Jones, and Jeremy McCauley. 2016c. "Medical Spending of the US Elderly." *Fiscal Studies* 37 (September-December): 717–47.

De Nardi, Mariacristina, Eric French, John Bailey Jones, and Rory McGee. 2018. "Couples' and Singles' Savings After Retirement." Work in progress.

Fahle, Sean, Kathleen McGarry, and Jonathan Skinner. 2016."Out-of-Pocket Medical Expenditures in the United States: Evidence from the Health and Retirement Study." *Fiscal Studies* 37 (September-December): 785–819.

Feenberg, Daniel, and Jonathan Skinner. 1994. "The Risk and Duration of Catastrophic Health Care Expenditures." *Review of Economics and Statistics* 76 (November): 633–47.

French, Eric, and John Bailey Jones. 2004. "On the Distribution and Dynamics of Health Care Costs." *Journal of Applied Econometrics* 19 (November): 705–21.

French, Eric, and John Bailey Jones. 2011. "The Effects of Health Insurance and Self-Insurance on Retirement Behavior." *Econometrica* 79 (May): 693–732.

French, Eric, John Bailey Jones, and Jeremy McCauley. 2017. "The Accuracy of Economic Measurement in the Health and Retirement Study." *Forum for Health Economics and Policy* 20 (December): 1–16.

French, Eric, Mariacristina De Nardi, John Bailey Jones, Olyesa Baker, and Phil Doctor. 2006. "Right Before the End: Asset Decumulation at the End of Life." Federal Reserve Bank of Chicago *Economic Perspectives* 30 (Third Quarter): 2–13.

Friedberg, Leora, Wenliang Hou, Wei Sun, Anthony Webb, and Zhenyu Li. 2014. "New Evidence on the Risk of Requiring Long-term Care." Center for Retirement Research at Boston College Working Paper 2014-12 (November).

Hirth, Richard A., Teresa B. Gibson, Helen G. Levy, Jeffrey A. Smith, Sebastian Calonico, and Anup Das. 2015. "New Evidence on the Persistence of Health Spending." *Medical Care Research and Review* 72 (June), 277–97.

Hugonnier, Julien, Florian Pelgrin, and Pascal St-Amour. 2013. "Health and (Other) Asset Holdings." *Review of Economic Studies* 80 (April): 663–710.

Hurd, Michael D., Pierre-Carl Michaud, and Susann Rohwedder. 2017. "Distribution of Lifetime Nursing Home Use and of Out-of-Pocket Spending." *Proceedings of the National Academy of Sciences* 114 (September): 9838–842.

Jung, Juergen, and Chung Tran. 2016. "Market Inefficiency, Insurance Mandate and Welfare: U.S. Health Care Reform 2010." *Review of Economic Dynamics* 20 (April): 132–59.

Kopecky, Karen, and Tatyana Koreshkova. 2014. "The Impact of Medical and Nursing Home Expenses on Savings." *American Economic Journal: Macroeconomics* 6 (July): 29–72.

Livshits, Igor, James MacGee, and Michele Tertilt. 2007. "Consumer Bankruptcy: A Fresh Start." *American Economic Review* 97 (March): 402–18.

Lockwood, Lee. 2012. "Incidental Bequests: Bequest Motives and the Choice to Self-Insure Late-Life Risks." Mimeo.

Love, David A. 2010. "The Effects of Marital Status and Children on Savings and Portfolio Choice." *Review of Financial Studies* 23 (January), 385–432.

Merrill Lynch Wealth Management. 2012. "Merrill Lynch Affluent Insights Survey National Fact Sheet." (February).

Nakajima, Makoto, and Irina A. Telyukova. 2018. "Medical Expenses and Saving in Retirement: The Case of U.S. and Sweden." Federal Reserve Bank of Minneapolis Opportunity and Inclusive Growth Institute Working Paper 8 (April).

Nakajima, Makoto, and Irina A. Telyukova. 2012. "Home Equity in Retirement." Mimeo.

Pashchenko, Svetlana. 2013. "Accounting for Non-annuitization." *Journal of Public Economics* 98 (February): 53–67.

Pashchenko, Svetlana, and Ponpoje Porapakkarm. 2013. "Quantitative Analysis of Health Insurance Reform: Separating Regulation from Redistribution." *Review of Economic Dynamics* 16 (July): 383–404.

Poterba, James M., Steven F. Venti, and David A. Wise. 2017. "The Asset Cost of Poor Health." *Journal of the Economics of Ageing* 9 (June): 172–84.

Reichling, Felix, and Kent Smetters. 2015. "Optimal Annuitization with Stochastic Mortality and Correlated Medical Costs." *American Economic Review* 105 (November): 3273–320.

Rohwedder, Susann, Steven J. Haider, and Michael D. Hurd. 2006. "Increases in Wealth Among the Elderly in the Early 1990s: How Much is Due to Survey Design?" *Review of Income and Wealth* 52 (December): 509–24.

Scholz, John Karl, Ananth Seshadri, and Surachai Khitatrakun. 2006. "Are Americans Saving 'Optimally' for Retirement?" *Journal of Political Economy* 114 (August): 607–43.

Skinner, Jonathan. 2007. "Are You Sure You're Saving Enough for Retirement?" *Journal of Economic Perspectives* 21 (Summer): 59–80.

Webb, Anthony, and Natalia A. Zhivan. 2010. "How Much is Enough?: The Distribution of Lifetime Health Care Costs." Center for Retirement Research at Boston College Working Paper 2010-1 (February).

Economic Quarterly—Volume 104, Number 3—Third Quarter 2018—Pages 137–151

Inefficiency in a Simple Model of Production and Bilateral Trade

Zachary Bethune, Bruno Sultanum, and Nicholas Trachter

Many assets, both real and financial, are traded in secondary over-the-counter (OTC) markets after their initial issuance (e.g., real estate, municipal bonds, Treasuries, asset-backed securities, etc.). The apparent importance of OTC markets has prompted an explosion of research on the topic. There is a large literature studying OTC markets with no ex-ante investment (see, Duffie et al. 2005, Duffie et al. 2007, Lagos and Rocheteau 2009, and Hugonnier et al. 2014, among others). A new and growing literature studies how the market structure affects ex-ante decisions. In Farboodi et al. (2017), the market structure interacts with the investors' incentives to invest in negotiation skills (which affects bargaining power). In Farboodi et al. (2016), the market structure interacts with the incentives to invest on the speed at which an investor will meet others in the secondary market. In Bethune et al. (2017), the market structure affects the incentives to produce new assets and to introduce them into the market. In these last three papers, the equilibrium is constrained-inefficient. Our goal in this paper is to study a source of inefficiency that stems from these models in a much simpler setup. To that end, we will work with a simplified version of Bethune et al. (2017).

We study a simple model of production and trade in OTC markets. The model is meant to capture some important features of these markets. First, as in the seminal paper Duffie et al. (2005), trade is resolved in decentralized markets, where the buyer and seller in a trade

The views expressed herein are those of the authors and do not necessarily reflect those of the Federal Reserve Bank of Richmond or the Federal Reserve System. We thank the editor and the three referees for useful comments. Email: nicholas.trachter@rich.frb.org.

meeting bargain over the terms of trade. Second, agents face uncertainty over whether their potential trade counterparty wants to or can trade with her. Third, as in Bethune et al. (2017), we endogenize the asset level by allowing agents to produce assets at some cost. Unlike these papers, and a large part of the literature on the topic, we do not include search frictions in the model. Instead, as in Babus and Kondor (2013) or Gofman (2014), we model the market with no search frictions.

The model is static. The economy is populated by two agents who produce and trade assets. The two agents differ in the level at which they value assets: one agent's asset valuation is ν_l, while the other agent's asset valuation is ν_h, with $\nu_h > \nu_l$. At the beginning of the only period in the model, agents are faced with an opportunity to produce an asset, but the production cost is random. Then, agents meet and trade. When trade happens, the trade surplus is divided among the trade participants using Nash bargaining, where θ denotes the bargaining power of the buyer. In the economy, the low-valuation agent acts as an intermediary: she produces the asset with the intention of offloading it during the meeting with the high-valuation agent. Thus, from now on, we refer to this agent as the intermediary. The high-valuation agent intends to obtain an asset. She can do so by either producing it herself or by enjoying the intermediation services provided by the low-valuation agent. We refer to this agent as the investor. We solve for the unique equilibrium in this economy, and we study its efficiency properties. The model is described and analyzed in Section 1.

We find that the equilibrium is never constrained-efficient in the sense that there is no way to split the surplus between the intermediary and the investor, i.e., no choice of $\theta \in [0, 1]$, that makes the equilibrium efficient. This result is consistent with Bethune et al. (2017), but it is different from what is found in many search models (for example, in the seminal work of Lagos and Wright [2005]) where there is always a θ such that the equilibrium is constrained-efficient. The fundamental difference lies in the fact that in these other environments there is a one-sided hold-up problem, while in ours—as in most of the recent literature on OTC markets—there is a double-sided hold-up problem.[1] The double-sided hold-up problem arises because both parties in a trade meeting have to make a costly investment. In order to have trade between the intermediary and investor, the intermediary has to produce the asset—pay a cost and gain the option of selling the asset, and thus she will be held up by the investor in the trade meeting. Likewise, for trade to happen, the investor has to refrain herself from

[1] Recent examples are Farboodi et al. (2016) and Farboodi et al. (2017).

producing the asset—to arrive at the trade meeting empty handed, she must give up the option of producing an asset, and thus she will also be held up by the intermediary in the trade meeting. The first hold-up problem could be solved by assigning all the bargaining power to the intermediary, i.e., $\theta = 0$ (given that the intermediary acts as the seller in the trade meeting). The second hold-up problem could be solved by assigning all the bargaining power to the investor, i.e., $\theta = 1$ (given that the investor acts as the buyer in the trade meeting). Both cannot be handled together. Thus, the equilibrium is inefficient. We study efficiency in Section 2.

Interestingly, the inefficiency implies that there can be over- or underinvestment in asset production, which translates into an over- or undersupply of assets in the economy. When θ is low, the inefficiency implies that there is overinvestment and thus an excess supply of assets. When θ is high, the inefficiency implies that there is underinvestment and thus an inefficiently low asset supply. We study the effect of the inefficiency on the asset supply in detail in Section 3.

Our result in this paper is the same result found in Bethune et al. (2017). In that paper, we study a dynamic model of OTC trade with production, where meetings among potential traders are subject to search frictions, as in Duffie et al. (2005) or Hugonnier et al. (2014). In this paper, we show that the inefficiency result does not stem from search frictions and its potential inefficiency related to the Hosios condition (Hosios 1990). In fact, there is no search in our model here. We also show that the inefficiency result vanishes when both intermediary and investor can perfectly forecast the actions of their trade counterparties through the equilibrium. We show this by studying the case where the upper bound of the production cost equals ν_h, which can be interpreted as the limit as the upper bound of the cost distribution approaches ν_h. It happens that, in this limit, the high-valuation agent invests for any θ and for any draw of production cost. This is perfectly forecastable by the intermediary, who therefore does not produce. We study this in Section 4. Finally, Section 5 concludes.

1. MODEL

We study a simple two-agent, one-period model of asset trading. Agents differ in their valuation of the asset. One agent values the asset at level ν_l, while the other agent values the asset at level ν_h, where $\nu_h > \nu_l$. Agents can hold either zero or one unit of the asset. At the beginning of the period, both agents are confronted with an opportunity to invest in producing an asset. The investment cost c is random, drawn from the cumulative distribution $G(c)$. This distribution is assumed to be

uniform in the support $[\underline{c}, \bar{c}]$ with $g(c) \equiv \partial G(c)/\partial c$. To keep things as easy as possible, we restrict attention to the case where $0 = \underline{c} = \nu_l < \nu_h < \bar{c} = 1$. The restriction implies that (i) agents with a low valuation of the asset only produce with the intention to sell and thus play the role of intermediaries, and (ii) because $\nu_h < \bar{c}$, we can abstract from potential kinks and discontinuities when analyzing optimal behavior.[2] Given (i), we refer to agents with low valuation as intermediaries and to agents with high valuation as investors. Finally, once production decisions are made, agents meet and trade. We restrict them to using a Nash bargaining protocol to resolve trade. We let $\theta \in [0, 1]$ denote the bargaining power of the buyer. We later argue that the investor plays the role of the buyer in the trade meeting, while the intermediary plays the role of the seller in the trade meeting. Thus, we can think of θ as the bargaining power of the investor.

We can solve this model using a backward-induction argument. First, conditional on an asset distribution among agents, we solve the problem where the two agents meet and trade. Then, with this solution at hand, we solve the investment stage of the model.

Trade. When agents meet and trade, the outcomes depend on who enters this stage holding an asset. When both investors are either holding or not holding an asset, there is trivially no trade. If the investor enters the trade stage holding an asset, there is no trade. This follows because the maximum willingness to pay for the asset by a intermediary is ν_l, which is lower than the cost of giving up the asset for the investor, which is ν_h. As a result, there is no positive surplus to be split by the two parties involved in trade. The only case where there will be trade is when the intermediary is holding an asset and the investor is not. Because $\nu_h > \nu_l$, the total surplus of trading is positive, and thus trade will happen. However, the way the surplus is split among the two traders depends on the way bargaining power is assigned to them. Let p denote the transfer, or price, that the seller—the intermediary—requires to transfer the asset to the buyer—the investor. The optimal price p solves

$$\max_p (p - \nu_l)^{1-\theta} (\nu_h - p)^{\theta} \ . \tag{1}$$

Notice that when agents trade: (i) the intermediary receives p in exchange for giving up the asset, which she values at ν_l, and (ii) the investor receives the asset, which she values at ν_h, and pays p. Operat-

[2] Because $\bar{c} > \nu_h$, there is always a draw for the production cost for which neither agent wants to produce. On the flip side, because $\underline{c} = 0 < \nu_l$, there is aways a draw for the production cost for which both agents want to produce. These two observations, coupled with full support for the production cost in $[\underline{c}, \bar{c}]$, guarantees that the problem is smooth.

ing with the first-order condition of the problem presented in equation (1) provides an expression for the price,

$$p = \theta\nu_l + (1-\theta)\nu_h \ , \tag{2}$$

which is simply a weighted average of the asset valuation of the seller and buyer. When the full bargaining power is assigned to the intermediary, i.e. $\theta = 0$, we obtain that $p = \nu_h$. That is, the intermediary obtains all the surplus of the trade. Likewise, when the full bargaining power is assigned to the investor, i.e. $\theta = 1$, we obtain that $p = \nu_l$. That is, the investor obtains all the surplus of the trade. In fact, given that $\partial p/\partial\theta = -(\nu_h - \nu_l) < 0$, the price decreases as the bargaining power of the investor increases, implying that a higher fraction of the total surplus of trade is captured by the investor.

Investment. Given the solution to the trade stage, we can now solve the problem of both the intermediary and investor being faced with a production opportunity. They both have to decide for which set of production costs c they are willing to produce. Notice that once investors reach the trade stage, the investment cost is a sunk cost, thus it does not directly affect the way the surplus is divided by the trade meeting participants. It is also straightforward to argue that the value of producing an asset for either the intermediary or the investor is higher the lower the cost actually paid to produce. This implies that both agents should produce the asset when they draw low costs and potentially neither produces the asset for high costs. Let c_l denote the threshold such that the intermediary produces if she draws a cost c below c_l. Likewise, let c_h denote the threshold such that the investor produces if she draws a cost c below c_h. Characterizing these two thresholds involves striking a balance between the benefits of investing and not investing.

Let $U_l(i|c_h)$ denote the utility of producing the asset for the intermediary when the investor's production threshold is given by c_h. Likewise, let $U_l(ni|c_h)$ denote the utility of not producing the asset for the intermediary when the investor's production threshold is given by c_h. A similar logic applies for describing $U_h(i|c_l)$ and $U_h(ni|c_l)$.

When the intermediary does not produce, we trivially obtain $U_l(ni|c_h) = 0$. When the intermediary produces at cost c, we get that

$$\begin{aligned} U_l(i|c_h) &= G(c_h)\nu_l + [1-G(c_h)]p - c = \nu_h - G(c_h)(\nu_h - \nu_l) - \\ &\quad [1-G(c_h)]\theta(\nu_h - \nu_l) - c \ , \end{aligned}$$

where the last expression follows by using equation (2). With probability $G(c_h)$, the investor produces an asset, and thus the intermediary keeps the asset for herself, providing her ν_l utiles. With probability $[1 - G(c_h)]$, the investor does not produce, which implies that trade will occur at the trading stage, and the intermediary obtains p.

The threshold c_l is such that if $c = c_l$, we get $U_l(ni|c_h) = U_l(i|c_h)$. Using the expressions above and solving for c_l provides

$$c_l = \nu_h - G(c_h)(\nu_h - \nu_l) - [1 - G(c_h)]\theta(\nu_h - \nu_l) \ . \tag{3}$$

Now we solve the problem of an investor. If the investor produces the asset, she gains the asset and does not trade later on. Thus, $U_h(i|c_l) = \nu_h - c$. If she does not produce the asset, her utility is given by

$$U_h(ni|c_l) = G(c_l)[\nu_h - p] = G(c_l)\theta[\nu_h - \nu_l] \ ,$$

where the last expression follows by using equation (2). The investor only gets utility from not investing when the intermediary invests, which occurs with probability $G(c_l)$. In this event, her gains from trade are $\nu_h - p$, or a fraction θ of the total trade surplus $\nu_h - \nu_l$. As before, the threshold c_h makes the investor indifferent. Then, we have that

$$c_h = \nu_h - G(c_l)\theta[\nu_h - \nu_l] \ . \tag{4}$$

From equations (3) and (4), we can compute the following expressions,

$$\frac{\partial c_l}{\partial c_h} = -g(c_h)(1-\theta)(\nu_h - \nu_l) < 0 \quad , \quad \frac{\partial c_h}{\partial c_l} = -g(c_l)\theta(\nu_h - \nu_l) < 0 \ .$$

The first expression implies that if the intermediary expects the investor to be more likely to produce, then she will be less willing to produce herself. This occurs because the only reason the intermediary produces an asset is to attempt to sell the asset to the investor. If the probability of selling the asset falls (because the investor is more likely to hold an asset), there are lower incentives for the intermediary to produce. Likewise, because of the same logic, the second expression implies that the investor is less likely to produce if the intermediary is more likely to do it. That is, for both agents, their actions are strategic substitutes.

Equilibrium. An equilibrium is a pair of thresholds c_l^* and c_h^* that solve equations (3) and (4). Solving these two thresholds provides

$$c_l^* = \frac{[1-\nu_h]\nu_h(1-\theta)}{1-\theta(1-\theta)\nu_h^2} \in [0, \nu_h) \quad , \quad c_h^* = \frac{[1-\theta(1-\theta)\nu_h]\nu_h}{1-\theta(1-\theta)\nu_h^2} \in [0, \nu_h] \ .$$

Obviously, the fact that we were able to obtain closed-form expressions for both thresholds implies that the equilibrium exists and that it is unique. Moreover, the fact that the thresholds are interior and below ν_h follows from the fact that $\nu_h < 1$ and $\theta < 1$. The thresholds are not above ν_h, as there is always a profitable deviation that increases profits. For example, if $c_h^* > \nu_h$, the investor could trivially increase her expected utility by setting $c_l^* = \nu_h$. The fact that the threshold c_l^*

is strictly below ν_h follows from the fact that $c_l^* < c_h^*$. This follows, as there is a positive probability that the intermediary cannot sell the asset to the investor; because the intermediary values the asset at $\nu_l < \nu_h$, it follows that $c_l^* < c_h^*$.

2. EFFICIENCY

The constrained-efficient allocation is a set of thresholds c_l^e and c_h^e that solves the following problem,

$$\max_{c_l^e, c_h^e} \int_{\underline{c}}^{c_l^e} \int_{\underline{c}}^{c_h^e} [\nu_l + \nu_h - \tilde{c}_l - \tilde{c}_h] dG(\tilde{c}_h) dG(\tilde{c}_l)$$
$$+ [1 - G(c_l^e)] \int_{\underline{c}}^{c_h^e} [\nu_h - \tilde{c}_h] dG(\tilde{c}_h) + [1 - G(c_h^e)] \int_{\underline{c}}^{c_l^e} [\nu_h - \tilde{c}_l] dG(\tilde{c}_l) \ ,$$

where the problem already assumes that if the intermediary produces an asset and the investor does not, the efficient outcome is for the intermediary to transfer the asset to the investor. There are three terms in the welfare expression. The first term accounts for the case where both agents draw a cost below their respective production thresholds and produce. Because both produce, both derive utility at the end of the period from holding the asset and both pay the production cost. The second term accounts for the case where the intermediary does not produce and the investor does produce. In this case, there is no trade and the investor keeps the asset she produces, obtaining utility ν_h minus the production cost incurred. The third term accounts for the case where only the intermediary produces. In this case, she pays the production cost and transfers the asset to the investor, who values the asset at ν_h.

The first-order condition with respect to c_l^e is given by

$$g(c_l^e) \left\{ \int_{\underline{c}}^{c_h^e} [\nu_l + \nu_h - c_l^e - \tilde{c}_h] dG(\tilde{c}_h) - \int_{\underline{c}}^{c_h^e} [\nu_h - \tilde{c}_h] dG(\tilde{c}_h) + [1 - G(c_h^e)](\nu_h - c_l^e) \right\} = 0 \ .$$

Operating with this expression provides

$$c_l^e = \nu_h - G(c_h^e)(\nu_h - \nu_l) \ . \tag{5}$$

Likewise, we can obtain the first-order condition with respect to c_h^e and operate to obtain

$$c_h^e = \nu_h - G(c_l^e)(\nu_h - \nu_l) \ . \tag{6}$$

Then, it is straightforward to obtain expressions for the thresholds,

$$c_l^e = c_h^e = \frac{\nu_h}{1 + \nu_h} \ . \tag{7}$$

Having solved for the constrained-efficient allocation, we are now ready to discuss the efficiency of the equilibrium. We do so in the next proposition.

Proposition 1 *For any $\theta \in [0,1]$, the equilibrium, characterized by $\{c_l^*, c_h^*\}$, is not constrained-efficient.*

This powerful result follows from the fact that there is no way of choosing the bargaining power parameter θ so as to satisfy $c_l^* = c_l^e$ and $c_h^* = c_h^e$. One way to proof this result is the following. Notice that in the constrained-efficient allocation we have that the thresholds satisfy $c_l^e = c_h^e$. Then, if the equilibrium were to be efficient, a necessary condition is that $c_l^* = c_h^*$. However,

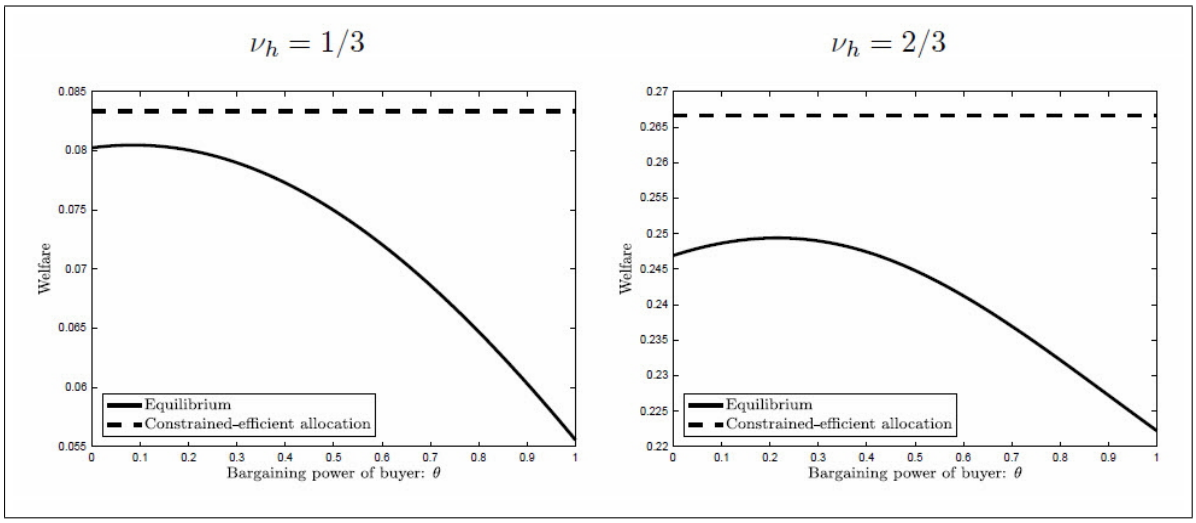
$$\frac{c_l^*}{c_h^*} = \frac{1-\theta-\nu_h(1-\theta)}{1-\theta(1-\theta)\nu_h} = 1 - \frac{\nu_h(1-\theta)+\theta(1-(1-\theta)\nu_h)}{1-\theta(1-\theta)\nu_h} < 1 \ ,$$

so that $c_l^* < c_h^*$.

There is a key difference between the social values presented in equations (5) and (6) and private values presented in equations (3) and (4). When the planner evaluates the social value of allocating an asset to the intermediary, the planner takes into account that the intermediary will, with some probability, transfer the asset to the investor. In doing so, the planner takes into account the entire surplus generated by transferring an asset from the intermediary to the investor, or the entire surplus from trade. When the intermediary evaluates her reservation value, however, she takes into account only a fraction of the surplus. The reason is because she faces a hold-up problem—her decision to invest in producing an asset occurs before meeting with the buyer for the asset, i.e., the investor. The intermediary gives up c utiles (the cost of production) and gains the option of selling the asset to the investor. As usual in hold-up problems, the only way that the equations for the social value and reservation value of the intermediary coincide is if the intermediary has all the bargaining power when selling the asset to the investor. That is, $\theta = 0$. This is clear when we compare the equations describing the social and private values for the intermediary.

In an analogous way to what happens with the intermediary, when the planner evaluates the social value of the production by the investor, the planner takes into account that the investor will, with some probability, obtain the asset from the intermediary. In doing so, the planner takes into account the entire loss of surplus generated by passing an asset from the intermediary to the investor. When the investor evaluates her reservation value, she takes into account only a fraction θ, her bargaining power, of this surplus. The investor also faces a hold-up problem—her decision not to invest in producing an asset occurs before

Figure 1 Welfare in the Equilibrium Allocation and in the Constrained-Efficient Allocation



meeting with the intermediary. That is, in order to gain the option of buying the asset from the intermediary, she gives up the option of producing the asset herself, and this option is gone once she meets with the intermediary. The only way that the equations for the social value and reservation value of an investor coincide is if the investor has all the bargaining power when buying an asset. That is, $\theta = 1$.

Fixing the two sides of the hold-up problem is not possible, as this would require that both intermediary and investor have all the surplus generated by a trade. Both agents will never value the gains from trade in the same way the planner does, and as a result, the outcome of a decentralized equilibrium cannot replicate the planner's solution—no matter how investors bargain over gains from trade. Figure 1 presents the welfare of both equilibrium and constrained-efficient allocations for two different values of ν_h as we vary θ. As discussed, because there is no θ for which the constrained-efficient allocation can be decentralized, the equilibrium welfare lies below the constrained-efficient welfare for all $\theta \in [0, 1]$.

3. ASSET DISTRIBUTION

Through altering production incentives, the inefficiency manifests in the asset distribution of the economy. In this section, we aim to provide

some insights into how this happens. To this end, let s denote the asset level in the economy. There are three possible asset levels: 0, 1, or 2. There will be zero assets if both agents do not produce, there will be one asset if only one of them produces, and there will be two assets if both produce. To compute the expected asset level, we first need to compute the probability of each of the three potential outcomes. We have that

$$\Pr(s=1) = (1-G(c_l))G(c_h)+(1-G(c_h))G(c_l) = c_l+c_h-2c_lc_h\ ,\ \Pr(s=2) = G(c_l)G(c_h) = c_lc_h\ ,$$

and where $\Pr(s=0) = 1 - \Pr(s=1) - \Pr(s=2)$. Then, the expected asset level is given by $E(s) = \Pr(s=1) + 2\Pr(s=2) = c_l + c_h$, which provides the following expressions for the equilibrium and efficient allocation,

$$E(s^*) = \nu_h \frac{2-\theta-\nu_h(1-\theta)(1+\theta)}{1-\theta(1-\theta)v_h^2} > 0,\ E(s^e) = 2\frac{\nu_h}{1+\nu_h} \in (0,1).$$

We can combine these two expressions to obtain a relationship between the asset level in both cases:

$$E(s^*) = E(s^e)\frac{[2-\theta-\nu_h(1-\theta)(1+\theta](1+\nu_h)}{2(1-\theta(1-\theta)v_h^2)}, \tag{8}$$

so that the equilibrium has a higher (lower) asset level than the efficient allocation if the second term on the right-hand side of the expression is larger (smaller) than one. Interestingly, this term can be above or below one, depending on the value for θ. In particular, when $\theta = 0$, we get that it is larger than one, and when $\theta = 1$, we get that it is smaller than one. This implies the following result.

Proposition 2 *The equilibrium allocation exhibits a lower asset level than in the constrained-efficient allocation when $\theta = 1$, and it exhibits a higher asset level than in the constrained-efficient allocation when $\theta = 0$.*

In equilibrium, when $\theta = 1$, all the gains from trade go to the investor. Thus, the intermediary has no incentives to produce the asset. However, the planner would like the intermediary to produce the asset with some probability, as this increases the chances that the investor ends up holding the asset. As a result, the asset supply is inefficiently low when θ is high. When $\theta = 0$, the intermediary has maximum incentive to produce the asset (as she gets all the gains from trade) but so does the investor, as she would get no gains from trading with the intermediary. As a result, the asset supply is inefficiently high when θ is low. Figure 2 presents the asset level in the equilibrium

and constrained-efficient allocations for two levels of ν_h. As seen in the figure, consistent with the discussion about $\theta = 0$ and $\theta = 1$, for low values of the bargaining parameter θ, the equilibrium exhibits overproduction, while for high values of θ, it exhibits underproduction.

The figure also shows that there is a value for the bargaining power such that the equilibrium asset supply equates the asset supply in the constrained-efficient allocation. However, given our impossibility result in Proposition 1, we know that the allocation is inefficient even when the equilibrium asset supply matches the efficient asset supply. What happens is the distribution of assets across the two agents is inefficient, in the sense that even though we may find a θ such that $E(s^*) = E(s^e)$, at this θ it must be the case that the equalities required for efficiency, i.e., $c_l^* = c_l^e$ and $c_h^* = c_h^e$, are not satisfied. In fact, given that $c_l^* < c_h^*$ and $c_l^e/c_h^e = 1$, it is the case that, for a given asset level, the way assets are distributed across the two agents is never the same as in the efficient allocation. This observation allows us to provide an alternative interpretation of the inefficiency result. The double hold-up problem affects efficiency by distorting both the level of the asset supply and the way assets are distributed across agents. There is always a way of choosing θ in order to fix the first distortion: from equation (8), and by continuity argument, it is clear that there always exists a way to choose θ such that $E(s^*) = E(s^e)$. However, given $c_l^*/c_h^* \neq c_l^e/c_h^e$ for all θ, there is no way of correcting the second distortion.

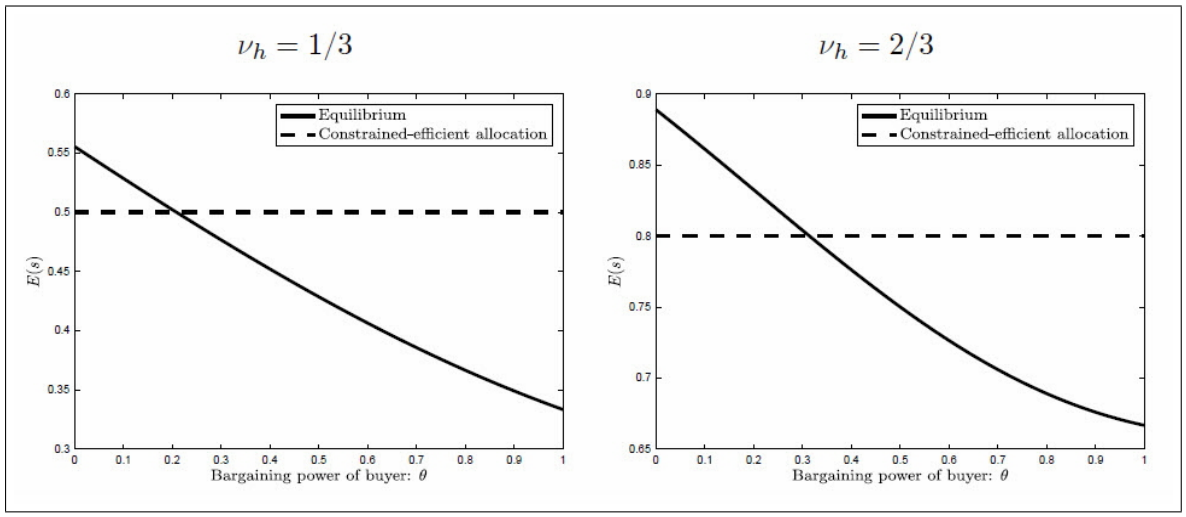
4. AN EXAMPLE WHERE THE EQUILIBRIUM IS CONSTRAINED-EFFICIENT

Although the model we developed here has no search frictions, in the sense that at the production stage agents know who they will meet at the trading stage, there is a fundamental friction in the model that stems from the fact that both parties in trade, when faced with the opportunity to produce, are unaware if their trade counterparty will be holding an asset or not when they meet at the trade stage. This friction is important for understanding the inefficiency result presented in Proposition 1. To study this in some detail, we allow the upper bound of the production cost distribution $\bar{c}$ to be in the set $[\nu_h, 1]$. Recall that, before, we had assumed $\bar{c} = 1$. Notice that the inefficiency results discussed above hold for $\bar{c} > \nu_h$. Thus, think of the case $\bar{c} = \nu_h$ as the limiting case when $\bar{c} \downarrow \nu_h$.

When $\bar{c}_h \in [\nu_h, 1]$, the equilibrium thresholds, equations(3) and (4), reduce to the following expressions,

$$c_l^* = \nu_h(1-\theta)\left(1 - \frac{c_h^*}{\bar{c}}\right) \quad , \quad c_h^* = \nu_h\left(1 - \frac{c_l^*}{\bar{c}}\theta\right) \ .$$

Figure 2 Asset Level in the Equilibrium Allocation and in the Constrained-Efficient Allocation



Likewise, the thresholds in the constrained-efficient allocation, equations (5) and (6), reduce to

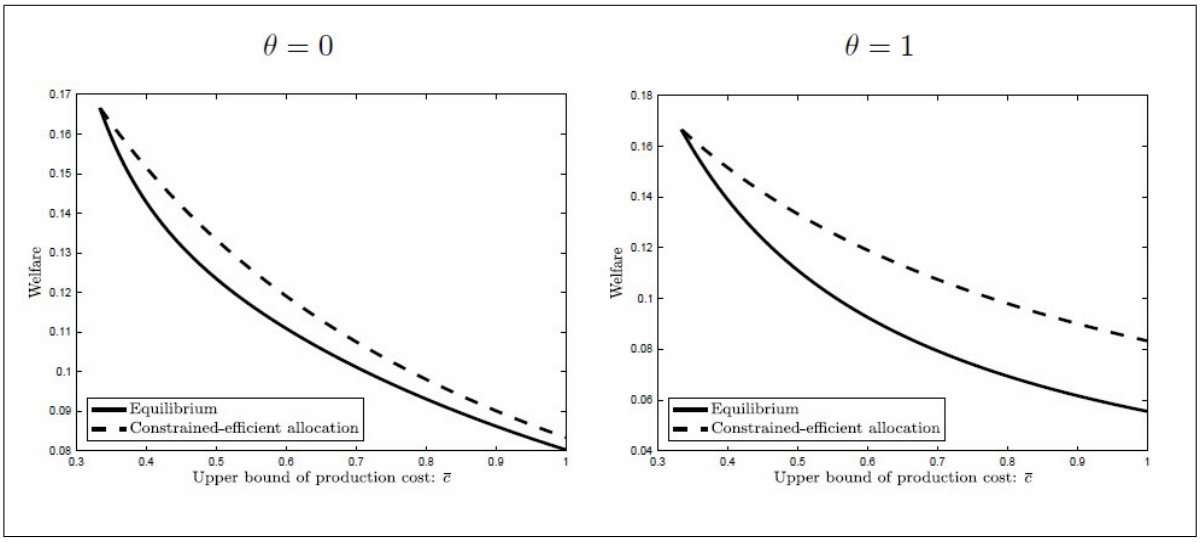
$$c_l^e = \nu_h \left(1 - \frac{c_h^e}{\bar{c}}\right) \quad , \quad c_h^e = \nu_h \left(1 - \frac{c_l^e}{\bar{c}}\right) \ .$$

Suppose $\bar{c} = \nu_h$. In the constrained-efficient allocation, we obtain $c_l^e + c_h^e = \nu_h$. Because the production cost distribution is uniform, the density is the same at every production cost. As a result, the efficient allocation cares about maximizing the production probability, which equals ν_h, but it is silent on how production is divided between the intermediary and the investor. Notice that the efficient allocation can be attained with trade, for example, if only the intermediary produces ($c_l^e = \nu_h$ and $c_h^e = 0$), with cases where trade sometimes happens (both agents having interior production thresholds), or cases with no trade (which occurs when only the investor produces).

In particular, the planner can attain the constrained-efficient allocation by setting $c_l^e = 0$ and $c_h^e = \nu_h$: the intermediary does not produce, and all production is carried over by the investor. Now we turn to the equilibrium thresholds. Combining c_l^* and c_h^* provides an expression for c_h^*,

$$c_h^* = \frac{1 - \theta(1-\theta)}{\frac{1}{\nu_h} - \frac{\theta(1-\theta)}{\nu_h}} = \nu_h \ ,$$

Figure 3 Welfare and the Upper Bound of Production Cost Distribution $\nu_h = 1/3$



where we already used that $\bar{c} = \nu_h$. Also, we immediately obtain that $c_l^* = 0$. In other words, when $\bar{c} = \nu_h$, the equilibrium is constrained-efficient for any $\theta \in [0, 1]$. This occurs because by reducing $\bar{c}$ up to ν_h we are taking to zero the probability that the investor does not produce. This implies that the intermediary can forecast perfectly that the investor will arrive to the trade meeting holding an asset. Likewise, because $c_l^* = 0$, the investor can forecast perfectly that the intermediary will arrive to the trade meeting not holding an asset. Thus, the intermediary does not produce as she knows she cannot offload the asset to the investor, and the investor produces as she knows she cannot acquire the asset from the intermediary. Figure 3 presents this result for this simple example. It shows that as $\bar{c}$ approaches $\nu_h = 1/3$, the equilibrium welfare approaches the welfare of the constrained-efficient allocation. It shows this in the left panel for the $\theta = 0$ case and in the right panel for the $\theta = 1$ case.

5. CONCLUSION

In this paper, we show that there is a generic source of inefficiency in OTC markets stemming from the fact that agents must make production choices (or, in general, trade decisions) ex ante without knowing

the choices of their trade counterparties ex post. Due to the lack of a large mass of agents, individual actions have strong effects on other agents' decisions. This implies that the equilibrium is not constrained-efficient, independent of the way agents split the surplus among them.

We then show that the inefficiency result vanishes in an example where the equilibrium provides more information to market participants at the production stage, thus improving their forecasting power regarding the actions of other players in the market. In the example we study, when the investor always invests, the intermediary never invests, so agents know with certainty if their trade counterparties will be holding an asset or not when they meet. We show that this attains efficiency.

Certainly, the model is a simple one. However, the inefficiency result seems robust. Given the growing interest in understanding trading in OTC markets, we believe that understanding the welfare properties of models of OTC trading is of paramount importance and calls for further research.

REFERENCES

Babus, Ana, and Péter Kondor. 2013. "Trading and Information Diffusion in OTC Markets." Centre for Economic Policy Research Discussion Papers 9271 (January).

Bethune, Zachary, Bruno Sultanum, and Nicholas Trachter. 2017. "Asset Issuance in Over-the-Counter Markets." Federal Reserve Bank of Richmond Working Paper 17-13 (October).

Duffie, Darrell, Nicolae Garleanu, and Lasse Heje Pedersen. 2005. "Over-the-Counter Markets." *Econometrica* 73 (November): 1815–47.

Duffie, Darrell, Nicolae Garleanu, and Lasse Heje Pedersen. 2007. "Valuation in Over-the-Counter Markets." *Review of Financial Studies* 20 (November): 1865–1900.

Farboodi, Maryam, Gregor Jarosch, and Guido Menzio. 2016. "Intermediation as Rent Extraction." Penn Institute for Economic Research Working Paper 16-026 (December).

Farboodi, Maryam, Gregor Jarosch, and Robert Shimer. 2017. "Meeting Technologies in Decentralized Asset Markets." Working Paper (February).

Gofman, Michael. 2014. "A Network-Based Analysis of Over-the-Counter Markets." Working Paper (December).

Hosios, Arthur J. 1990. "On the Efficiency of Matching and Related Models of Search Unemployment." *Review of Economic Studies* 57 (April): 279–98.

Hugonnier, Julien, Benjamin Lester, and Pierre-Olivier Weill. 2014. "Heterogeneity in Decentralized Asset Markets." Working Paper 20746. Cambridge, Mass.: National Bureau of Economic Research (December).

Lagos, Ricardo, and Guillaume Rocheteau. March 2009. "Liquidity in Asset Markets with Search Frictions." *Econometrica* 77 (March): 403–26.

Lagos, Ricardo, and Randall Wright. 2005. "A Unified Framework for Monetary Theory and Policy Analysis." *Journal of Political Economy* 113 (June): 463–84.