An Assessment of Long-term Healthcare Expenditure Risk using a Dynamic Bayesian Model

Devin Incerti

October 22, 2015

Abstract

This study uses data from eight waves of the Health and Retirement Survey to analyze out-of-pocket expenditures over a considerably longer time period than in previous studies. It develops a dynamic Bayesian model of out-of-pocket health expenditures and mortality that is able to account for (1) the persistence of expenditures from one year to the next, (2) the large number of nonspenders, (3) the right-skewed nature of nonzero expenditures and (4) the relationship between expenditures and mortality. Although the distribution of spending is less skewed over longer periods than in a single year, there is still substantial inequality in the long-term distribution of spending due to transitory shocks, state dependence and unobserved heterogeneity. Out-of-pocket expenditures are determined more by transitory shocks and unobserved heterogeneity than observed data so individuals face substantial uncertainty in future health costs. Implications for cost containment efforts such as consumer directed health plans and disease management programs are discussed.

1 Introduction

Healthcare spending in the United States accounted for 17.4% of gross domestic product (GDP) in 2013 and is projected to reach 19.6% of GDP by 2024 (Centers for Medicare & Medicaid Services). These costs have placed increasing strains on U.S. firms–the primary sponsors of health insurance in the United States—as well as on government budgets. At the same time, individuals face rising premiums, lower wages and higher tax rates. Policy reforms that slow the rising cost of healthcare are consequently a top priority.

To shed light on cost containment efforts, I use dynamic Bayesian model to study 1) the long-term concentration of out-of-pocket healthcare spending and 2) whether out-of-pocket costs are predictable in the long run. This has important implications for two particular cost containment approaches. One approach is to use disease management programs, which provide organized, proactive care to patients with specific conditions, to help control costs associated with avoidable complications. These programs are the most likely to succeed when costs are highly concentrated and when costs can be predicted accurately.[1] A second approach aims to reduce spending through consumer-directed health plans (CDHPs), which require consumers to use personal healthcare accounts to pay for non-catastrophic healthcare costs out-of-pocket. These plans have fewer tradeoffs when future costs are predictable because insuring against risk is less valuable. Furthermore, incentives for risk selection (by consumers and insurers) and the distributional implications of different cost sharing arrangements are directly related to the concentration of spending.

There is considerable evidence that spending varies significantly across individuals in a given year. Indeed, one of the most commonly cited statistics in American healthcare is that the top 5% of spenders account for approximately half of all U.S. healthcare spending (e.g. Zuvekas and Cohen 2007). There is also a large body of evidence suggesting that only a small

[1]Total spending—rather than out-of-pocket spending—is a more relevant cost measure for evaluating disease management programs. However, out-of-pocket and total costs are highly correlated so differences between the two measures should be small.

percentage—at most 15 to 20 percent—of the variation in healthcare costs in a given year is predictable (e.g. Van Vliet 1992; Newhouse 2004; Pope et al. 2004; Breyer, Bundorf and Pauly 2012). Yet much less is known about the distribution of spending or the uncertainty of future health costs over extended periods. In this paper I use 8 waves (16 years) of the Health and Retirement Survey (HRS) to estimate these quantities over a significantly longer time period than in previous studies, which provides new evidence for evaluating the potential consequences of cost containment policies.

Modeling healthcare spending is not straightforward. For one, healthcare spending data is what is typically referred to as semicontinuous because a large portion of responses are equal to 0 while the remaining values follow a continuous, heavily right skewed distribution. Secondly, spending is persistent over time. To address these concerns I use a longitudinal two-part model (Olsen and Schafer 2001; Tooze, Grunwald and Jones 2002; Neelon, O'Malley and Normand 2011) with lagged dependent variables.[2] The two-part model addresses the skewness of the data by modeling expenditures in two stages. In the first stage, a probit model is used to predict the probability of non-zero expenditures. Then, conditional on positive spending, a regression model predicts the log of expenditures. I account for the persistence in spending with past spending and unobserved individual effects. Since the probability of some use is likely correlated with the level of use, the individual effects are allowed to be correlated across the two equations.[3]

An additional complication is that long-term spending depends on life-expectancy. Since health spending tends to increase with age, mortality can have a large impact on future spending. Paradoxically, the sickest individuals can actually end up spending less over longer periods because they are more likely to die early in life. Thus, to forecast expenditures over part (or all) of an individual's lifetime it is necessary to forecast mortality as well. I consequently model expenditures and mortality jointly. The model is thus capable of

[2]Also see Zhang et al. (2006) for a Bayesian two-part hierarchical model in which patients are nested within physicians.

[3]Su, Tom and Farewell (2009) have shown that not accounting for this correlation can cause biased inferences.

simulating expenditures from any given date until death.[4]

The model highlights the importance of unobserved heterogeneity as it accounts for approximately 35 and 18 percent of the unexplained variation of the first and second stages of the two-part model. The correlation of the unobserved individual effects across equations is 0.516 which suggests that allowing the individual effects to be correlated is important. The model also exhibits strong state dependence as both lagged and initial values of the dependent variables are important predictors of future spending.

Out-of-pocket spending over longer periods is less skewed than spending in a single period, but a small fraction of the population still accounts for a large share of total health spending. For example, over a 16-year period, the top 5% of spenders still account for 38% of total spending. In simulations, the distributions become less and less skewed as the number of periods increase; however, even after 26 years the top 5% of spenders account for 28% of total spending. Most of the variation in spending is explained by random shocks and unobserved heterogeneity, which implies that individuals face substantial uncertainty in future health costs. As a result, it is difficult to predict high spenders and welfare calculations suggest large potential utility gains from using insurance to reduce financial risk.

This paper is related to a small collection of studies that have analyzed the dynamics of healthcare expenditures. Eichner, McClellan and Wise (1997) estimated a simple two-part model using 3 years of data and lagged dependent variables in order to gauge the feasibility of medical savings accounts. Feenberg, Skinner et al. (1994) and French and Jones (2004) estimate models with complex auto-correlated error structures on fairly long panels (6 and 8 years respectively) using GMM but do not account for the large point mass at zero in medical care data.[5]

[4]This makes the model suitable for state transition modeling (see Siebert et al. (2012) for an overview of state transition modeling and its use in cost-effectiveness analysis). The model can also be used to estimate lifetime out-of-pocket medical costs from a given age until death for households managing finances for retirement (see Webb and Zhivan (2010) for a related model used for this purpose).

[5]Feenberg, Skinner et al. (1994) only observe individuals whose medical expenses are larger than 3% of income and attempt to correct this using a Tobit model; French and Jones (2004) recode all expenditures below $250 to $250.

My study differs from these articles in a number of important respects. First, I use a considerably longer panel than previous studies. Second, I use a model that captures both the cross-sectional (two-part model) and dynamic aspects (unobserved heterogeneity and lagged dependent variables) of medical care data. Third, I model both mortality and spending. Lastly, the model is estimated using Bayesian methods which allows me to form a complete probability distribution for out-of-pocket expenditures given any initial state or spending history.

The remainder of this article proceeds as follows. Section 2 describes the HRS data. Section 3 introduces the model. In Section 4, I outline the Bayesian algorithm used to estimate the model, summarize results, and check the fit of the model to the data. Section 5 analyzes the distribution and uncertainty of long-term spending, and then examines implications for cost-containment reforms. Finally, Section 6 concludes.

2 Data

I use data from the 4th through 11th waves (1998 - 2012) of the Health and Retirement Survey (HRS). The HRS is a nationally representative longitudinal survey dataset of individuals over age 50. The main goal of the survey is to provide information for the study of health and retirement. It is the only nationally representative panel dataset containing information on health expenditures in the United States. I use the RAND HRS data files, which contain cleaned versions of the original variables that are renamed to allow for easier comparability across waves.[6] The main variable is total out-of-pocket expenditures which is only available from wave 4 onward.

To facilitate longitudinal analysis, the sample used in this study consists of all survey respondents who participated in the 4th wave, but not those who joined the survey during later waves. I also drop individuals from the sample if they have missing data due to non-response or dropout, which reduces the number of individuals in the sample from $20,569$ to

[6] See Appendix A for more details on RAND's cleaning of the data.

16,591. An additional 8 individuals are dropped because information on race or ethnicity is unavailable. In Appendix B, I show that restricting the data in this manner does not significantly alter the characteristics of the sample. Overall, this suggests that dropping observations has a minimal impact on the coefficient estimates and that the final sample is reasonably representative of the original sample.

The HRS data is self-reported so the reported expenditures are almost certainly measured with some error. Hurd and Rohwedder (2009) and Goldman, Zissimopoulos and Lu (2011) compare the HRS out-of-pocket spending estimates to those in the Medical Expenditure Panel Survey (MEPS) and the Medicare Current Beneficiary Survey (MCBS). Each survey spends considerable resources in order to collect high quality data: the MCBS combines survey data with administrative Medicare files while the MEPS combines survey data with data from the providers who provided care for the survey respondents. Both studies find that out-of-pocket spending estimates from the HRS are comparable (albeit slightly overestimated) to those in the MCBS but significantly higher than those in the MEPS, with discrepancies largest at the mean and the upper end of the distribution.

However, as noted by Marshall, McGarry and Skinner (2011), the HRS may capture expenses from sources—such as in-home care—that are not covered in other surveys. Furthermore, the MEPS is known to underreport expenditures relative to both the MCBS (Zuvekas and Olin 2009) and the National Health Expenditure Accounts (NHEA) (Keehan 2006; Bernard et al. 2012). In fact, as Marshall, McGarry and Skinner (2011) have pointed out, the per-year sample average of out-of-pocket spending in the 2004 HRS for the over age 65 population, or approximately $2100, is nearly equivalent to the 2002 and 2004 estimates of over age 65 out-of-pocket spending from the NHEA as reported in Hartman et al. (2008). In a similar fashion, French and Jones (2004) has shown that average expenditures from the HRS tend to match US averages closely.

Table 1 reports summary statistics on the distribution of out-of pocket expenditures in the HRS. Expenditures reported in the table and the rest of the paper are in $2012. The

first row reports summary statistics from the HRS pooled across the 4th through 11th waves. Mean out-of-pocket expenditures are around $2100 per year, which is consistent with the NHEA estimates reported in Hartman et al. (2008) and discussed above. The distribution is heavily right skewed with a mean considerably larger than the median and a very large maximum value of over $1 million. The percentage of non-spenders is also quite high, which suggests that modeling out-of-pocket expenditures using a two-part model is warranted.

Table 1: The Distribution of Out-of-Pocket Health Expenditures in the HRS, Waves 4 - 11

Data	Sample size	% nonspenders	Mean	Quantile 25%	Quantile Median	Quantile 75%	Max
Pooled cross-section	99,564	10.93	4,261	456	1,518	3,840	1,539,869
Individual means	16,583	3.49	3,198	658	1,784	3,723	213,088
Individual means per wave alive	16,583	3.49	4,890	1,041	2,395	4,968	292,616

Notes: Estimates are two-year (or single wave) averages. Individual means are expenditures summed across all 8 waves divided by 8 (individuals are coded as having $0 expenditures in all waves after death). Individual means per wave alive are expenditures averaged across all waves in which an individual is alive.

The second and third rows each look at the distribution of spending averaged (by individual) across the 4th through 11th waves. The second row is the sum of expenditures across all waves divided by the number of waves in the data, or 8. Out-of-pocket expenditures are coded as $0 in all years after death. This quantity is the quantity that a public payer such as Medicare or Medicaid (or perhaps a family saving for retirement) would be concerned with. The distribution is less skewed than the distribution of the pooled cross section: the mean is lower, the median is larger, the maximum is substantially smaller, and there are very few non-spenders. Nonetheless, the distribution is still quite skewed with a mean considerably higher than the median. At 95th percentile, an individual spends $\$86,361$ over 16 years. Spending at the median is much lower, but the median individual still pays an economically significant $\$14,274$ out-of-pocket.

Individuals might be more concerned with how much they would pay when they are alive. The third row consequently provides summary statistics for mean out-of-pocket expenditures per wave alive. Mean spending in the third row is higher than mean spending in either of

the first two rows because individuals tend to spend more right before death. At the 95th percentile, individuals spend, on average, $\$130,226$, per wave alive. Spending at the 25th percentile and the median are also high, which suggests that a substantial fraction of the population experiences large out-of-pocket expenditures over their lifetimes later in life.

Table 1 showed that nonzero spending is quite common in a single wave but much less prevalent over multiple waves. Figure 1 investigates this more closely. The leftmost figure plots the cumulative share of individuals with nonzero spending by wave; that is, the fraction of (surviving) individuals with nonzero spending in either the current wave or any previous wave. In wave 4, 87% of individuals had positive expenditures. The cumulative share of individuals with nonzero expenditures then rose to 94% by wave 5 and steadily increased thereafter before surpassing 99% by wave 9. This means that nearly everyone who survived at least 6 waves had positive medical spending during at least one of those two-year periods.[7]

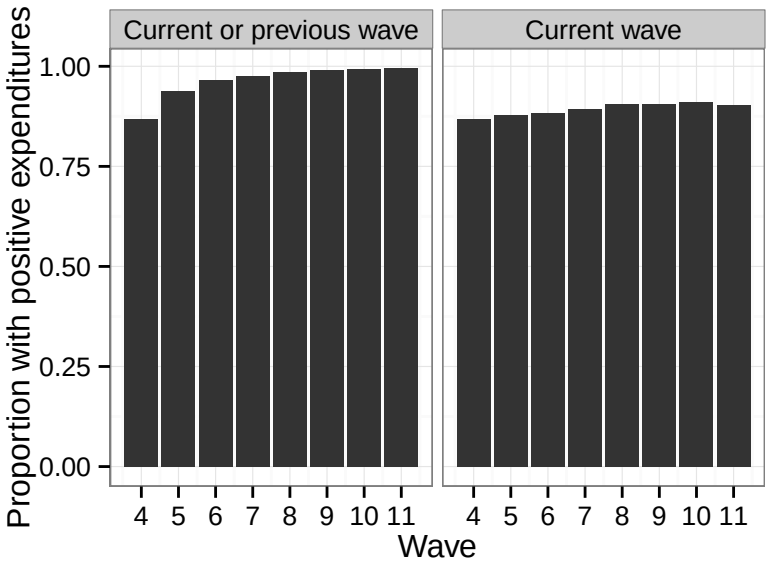

Figure 1: Nonzero Spending in the HRS

Notes: Each wave only includes survey respondents who survived to that wave.

[7]Note that the cumulative share of nonzero spending by wave 11 is higher than the fraction of individuals with nonzero spending reported in rows 2 and 3 of Table 1. This is because the statistic reported for wave 11 in Figure 1 is estimated using only those who survived from wave 4 to wave 11 while the statistic in Table 1 is based on all individuals regardless of their survival length.

The rightmost figure plots the fraction of nonzero spenders in each wave. This quantity increases slightly over time due to population aging. In wave 4, 87% of individuals had positive spending; by wave 11, 90% of survey respondents were nonzero spenders.

Table 1 showed that expenditures remain skewed even after averaging over 8 waves. This is consistent with a number of other studies that have shown that medical expenditures are highly persistent. Figure 2 examines persistence in the HRS data by looking at whether expenditures in wave 4 are correlated with expenditures in subsequent waves. More precisely, it plots mean out-of-pocket expenditures by expenditure quartile in wave 4.

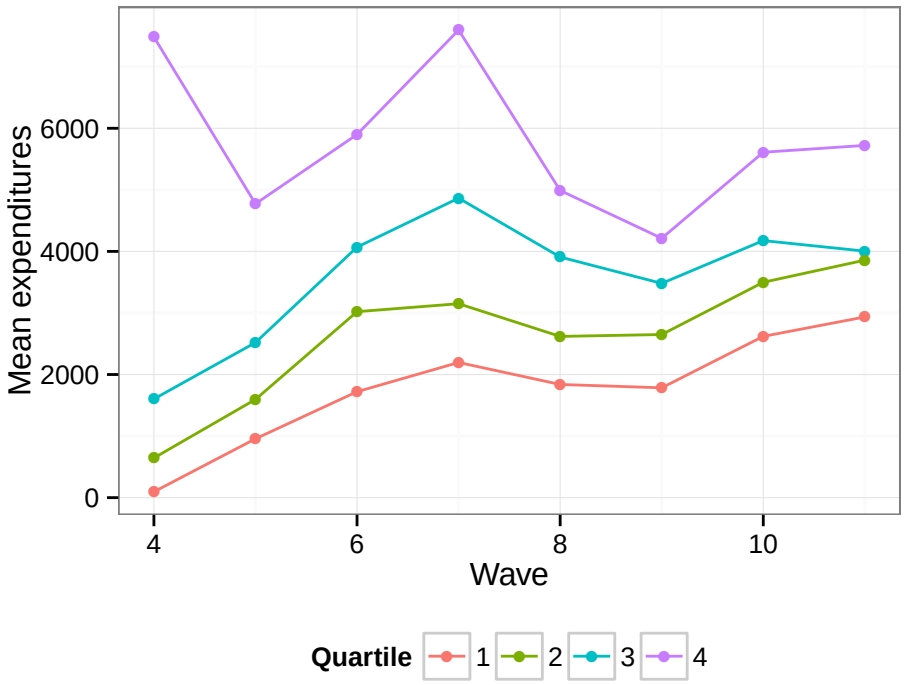

Figure 2: Mean Out-of-Pocket Expenditures by Wave 4 Quartile

During the initial wave an individual in the highest quartile spent, on average, almost $7483 while those in the lowest quartile spent, on average, only $96. Even mean spending in the third quartile was less than one-third mean spending in the highest quartile. But expenditures tend to gravitate toward the mean in all quartiles. Spending in the bottom

three quartiles rose over time while spending in the highest quartile declined. By the 11th wave mean spending in the lowest quartile rose to nearly one-half mean spending in the highest quartile. Nonetheless, the lines never cross which implies that, on average, those that had higher expenditures in wave 4 continued to have higher expenditures in subsequent waves. There is thus both significant mean reversion and persistence in the data.

3 The Model

This paper models out-of-pocket expenditures for individuals over age 50. Expenditures are modeled in two-parts with separate equations for nonzero expenditures and log expenditures conditional on positive expenditures. Spending is modeled as a function of exogenous personal characteristic, lagged spending, unobserved heterogeneity and transitory errors. Since the model includes lagged variables, it is estimated on the sample of $15,273$ individuals who survived to at least wave 5.

In order to forecast expenditures over an individual's lifetime it is necessary to forecast their mortality as well. Mortality is modeled using a probit equation as a function of out-of-pocket spending in the previous period and exogenous personal characteristics. The probit model is estimated using the sample of $15,273$ individuals based on their survival from wave 6 until wave 11.

3.1 Expenditure Model

Let y_{it} denote out-of-pocket expenditures where the indices i and t refer to individual i and period t respectively—for clarity, the initial wave (wave 4) is referred to as period 0. Furthermore, let $d_{it} \equiv I(d^*_{it} > 0) = I(y_{it} > 0)$ so that the latent variable d^*_{it} describes whether expenditures are positive or zero. The dynamic longitudinal two-part model can then be written as,

$$d^*_{it} = \phi_1 d_{it-1} + z^T_{1it}\gamma_1 + c_{1i} + \epsilon_{1it}, \tag{1}$$

$$\ln y_{it} | d^*_{it} > 0 = g(y_{it-1})^T \phi_2 + z^T_{2it}\gamma_2 + c_{2i} + \epsilon_{2it}, \tag{2}$$

where γ_1 and γ_2 are the vectors of coefficients on the explanatory variables; c_{1i} and c_{2i} are individual random effects; and ϕ_1 is the coefficient on the lagged dependent variable in the first part of the model. The function $g(\cdot)$ is a column vector that allows the observed response in the second part of the model to depend on lagged values in a number of ways. ϕ_2 is an $m \times 1$ vector where m is the number of variables contained in $g(y_{it-1})$. As is typical of two-part models, it is assumed that $\epsilon_{1it} \sim N(0,1)$ so that the first equation is a probit model and that $\epsilon_{2it} \sim N(0, \sigma_\epsilon^2)$. As will be shown later, the data generating process is well approximated by a lognormal distribution for nonzero y_{it} so assuming that ϵ_{2it} is normally distributed is not problematic. If this were not the case, the model could be extended so that the distribution of ϵ_{2it} were a student-t distribution, a skew normal distribution or some mixture of normals.[8]

A well documented problem with dynamic non-linear models with individual random effects is that the estimates depend on assumptions about the initial conditions [see Hsiao (2014, section 7.5)]. Standard assumptions are that the initial conditions are either fixed or random, but neither is without its limitations. Assuming that an initial condition is fixed implicitly assumes that it is uncorrelated with the individual effects. On the other hand, it is difficult to properly specify the density of the initial conditions given the individual effects. In this paper I use an alternative approach suggested by Wooldridge (2005) in which the distribution of the individual effects is modeled as a function of initial conditions and exogenous variables.[9] Specifically,

$$c_{1i} = \delta_{10} + \delta_{11} d_{i0} + z_{1i}^T \delta_{12} + b_{1i} \tag{3}$$

$$c_{2i} = \delta_{20} + h(y_{i0})^T \delta_{21} + z_{2i}^T \delta_{22} + b_{2i}, \tag{4}$$

[8] See chapter 15 in Koop, Poirier and Tobias (2007) for details on estimating these types of non-normal models.

[9] My model is therefore similar to Li and Zheng's 2008 Bayesian semi-parametric dynamic type I Tobit panel data model which also follows Wooldridge's 2005 suggestion (their model is semiparametric because it models the distribution of unobserved individual effects as a mixture of normals.)

where d_{i0} and y_{i0} are initial conditions and z_{1i} and z_{2i} are column vectors of time invariant explanatory variables for individual i. The role of the function $h(\cdot)$ is the same as the function $g(\cdot)$ above in that it allows the initial conditions to appear in a number of ways. Variables that can be included in z_{1i} and z_{2i} include time-constant variables such as race and sex and averages (over time) of the time-varying variables. Based on evidence from previous two-part longitudinal models, the individual specific error terms, b_{1i} and b_{2i}, are allowed to be correlated across the two equations. That is, they are given a bivariate normal distribution,

$$b_i = \begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix} \sim N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \Sigma_b = \begin{pmatrix} \sigma_1^2 & \rho\sigma_1\sigma_2 \\ \rho\sigma_1\sigma_2 & \sigma_2^2 \end{pmatrix} \right], \tag{5}$$

where ρ is the correlation between b_{1i} and b_{2i}. In this paper, I refer to b_{1i} and b_{2i} as unobserved heterogeneity or unobserved individual effects.

The model is assumed to be dynamically complete which implies that there is no serial correlation in the error terms after controlling for the lagged dependent variables. The observed persistence in out-of-pocket spending is consequently modeled through the lagged dependent variables and unobserved heterogeneity. The error terms are what is often referred to as equicorrelated because the co-variance in the errors between one wave and the next are equal to the unobserved individual effects and do not vary with the time difference between waves.

Variables included in the model are those that would be known to an individual at the initial wave and could be used to forecast spending over long periods of time given this initial information. The time-varying vectors z_{1it} and z_{2it} contain a quadratic function of age.[10] The time-constant vectors z_{1i} and z_{2i} consist of standard variables for demographics and socioeconomic status. Both age and a variable for years of education are centered to

[10] Additional time-varying variables could be included but these would not be known to an individual at the initial wave and would consequently need to be forecasted.

make the estimate of the intercept more meaningful. The functions $g(\cdot)$ and $h(\cdot)$ were chosen to best approximate the relationship between lagged (and initial) log spending and current log spending. As illustrated in figures E.1 and E.2 in the appendix, the effect of lagged (initial) expenditures should differ depending on whether previous (initial) expenditures were zero or strictly positive. Accordingly, I set $g(y_{it-1}) = (d_{it-1}, d_{it-1} \times \ln y_{it-1})^T$ and $h(y_{i0}) = (d_{i0}, d_{i0} \times \ln y_{i0})^T$ so that current log expenditures are predicted by an indicator variable if spending was zero in the previous (initial) period and lagged (initial) log expenditures if lagged (initial) spending was positive. Sample quantiles for the explanatory variables used in the model are provided in Table 2.

Table 2: Explanatory Variables Included in Model

Variable	Variable included in?			Sample Quantiles		
	Expenditure model		Mortality model			
	Binary component	Continuous component		Min	Median	Max
Intercept	Yes	Yes	Yes	1.0	1.0	1.0
$D_{it-1} \times \ln y_{it-1}$	No	Yes	Yes	0.0	7.3	14.2
D_{it-1}	Yes	Yes	Yes	0.0	1.0	1.0
$D_{i0} \times \ln y_{i0}$	No	Yes	No	0.0	7.0	12.6
D_{i0}	Yes	Yes	No	0.0	1.0	1.0
$(Age - 65)/10$	Yes	Yes	Yes	−1.5	0.5	4.4
$[(Age - 65)/10]^2$	Yes	Yes	Yes	0.0	0.5	19.4
Female	Yes	Yes	Yes	0.0	1.0	1.0
Black	Yes	Yes	Yes	0.0	0.0	1.0
Other race	Yes	Yes	Yes	0.0	0.0	1.0
Hispanic	Yes	Yes	Yes	0.0	0.0	1.0
Years of education - 12	Yes	Yes	Yes	−12.0	0.0	5.0

It is worth mentioning that the two-part model assumes that ϵ_{1it} and ϵ_{2it} are uncorrelated. An alternative approach is to use a type 2 Tobit or Heckman selection model to jointly model the errors. I believe that this would likely only add to the complexity of the current model with little added benefit, but future work might want to consider extending the model to allow for dependence in the error terms.

3.2 Mortality Model

One approach to modeling mortality is to generate survival curves using a hazard/survival modeling approach. Parametric proportional hazards models based on the Weibull or Gompertz distributions are natural candidates. However, in this paper, I use a model appropriate for discrete time since survival times are aggregated into two-year periods in the HRS.

Binary choice models are a simple and common choice for grouped survival data (Cox 1972; Lynch and Brown 2005; Cameron and Trivedi 2005). Here, I use a probit model but a logistic model could be used instead. It is convenient for Bayesian analysis to write the model in terms of a latent variable, say m_{it}^*, which can thought of as representing the risk of death. Then, letting $m_{it} = I(m_{it}^*) > 0$, the model can be written as,

$$m_{it}^* = x_{Mit}^T \kappa + \epsilon_{Mit}, \tag{6}$$

where x_{Mit} is the vector of explanatory variables for individual i, κ is the corresponding coefficient vector and $\epsilon_{Mit} \sim N(0, 1)$. The observed death indicator, m_{it}, is equal equal to 0 in all periods except the period of death in which case it is equal to 1. Individuals are dropped from the estimation sample following the period in which they died. The data vector x_{Mit} contains the same function of lagged expenditures from equation 2 and the same time constant variables from the expenditure model. All variables are summarized in Table 2.

4 Bayesian Estimation

4.1 Priors

In the Bayesian approach, prior distributions are assigned to all of the model parameters. To ensure a proper posterior that is determined almost entirely by the data, weakly informative prior distributions were chosen for the hyperparameters. The regression coefficients (γ_1, γ_2, δ_{10}, δ_{11}, δ_{12}, δ_{20}, δ_{21}, δ_{22}, κ) and the coefficients on the lagged dependent variables (ϕ_1, ϕ_2)

are given diffuse normal priors. The precision of the second equation of the two-part model, σ_ϵ^{-2} is assumed to have a $\text{Ga}(\text{a}_0, \text{b}_0)$ conjugate prior. As suggested by Neelon, O'Malley and Normand (2010) and Neelon, O'Malley and Normand (2011) the covariance matrix of the unobserved individual effects, Σ_b, is assumed to have a conjugate inverse-Wishart $IW(S_0, v_0)$ distribution. To make the prior uninformative, I set the degrees of freedoms, v_0, equal to 3 and S_0 equal to the identity matrix.

4.2 Posterior Computation

In the expenditure model, substitute equations 3 and 4 into equations 1 and 2 respectively so that the ith individual during period t has the vectors of explanatory variables $x_{1it} = (d_{it-1}, z_{1it}, d_{i0})$ and $x_{2it} = (g(y_{it-1}), z_{2it}, h(y_{i0}))$.[11] Denote the corresponding coefficient vectors as α and β respectively. In addition let expenditure data be available through time T_i and T_i' be either the the period of death or the final wave of the survey. That is, $T_i' = T_i$ for those who survived all 8 waves and $T_i' = T_i + 1$ for those who died during the survey period. Letting $\theta = (\alpha^T, \beta^T, \sigma_\epsilon^2)^T$, the conditional density, $f(y_{i1}, y_{i2}, \ldots, y_{iT}|y_{i0}, b_i, \theta)$ for individual i can then be written as,

$$\begin{aligned}
\prod_{t=1}^{T_i} f(y_{it}|y_{it-1}, b_i, \theta) &= \prod_{t=1}^{T_i} (1 - \Phi(\mu_{1it})^{1-d_{it}} \left[\Phi(\mu_{1it}) \times \text{LN}(\text{y}_{\text{it}}; \mu_{2\text{it}}, \sigma_\epsilon^2)\right]^{d_{it}}, \\
\mu_{1it} &= x_{1it}^T \alpha + b_{1i}, \\
\mu_{2it} &= x_{2it}^T \beta + b_{2i}, \qquad (7)
\end{aligned}$$

where $\Phi(\cdot)$ is the CDF of the standard normal distribution and $LN(\cdot)$ is the lognormal distribution. For individual i, the conditional density for mortality is just the standard likelihood for a binary choice model,

$$\prod_{t=2}^{T_i'} p(m_{it}|\kappa, y_{it-1}) = \prod_{t=2}^{T_i'} \Phi(x_{Mit}^T \kappa)^{m_{it}} \left[1 - \Phi(x_{Mit}^T \kappa)\right]^{1-m_{it}}, \qquad (8)$$

[11] The vectors z_{1it} and z_{2it} therefore include both time-varying and time-invariant variables.

except that m_{it} can only equal 1 if $t = T_i'$.

Given prior distributions $p(\theta)$, $p(\kappa)$ and $p(\Sigma_b)$, it follows that the full joint posterior is,

$$p(\theta, \kappa, b_i, \Sigma_b | y, m) \propto p(\theta)p(\Sigma_b) \prod_{i=1}^{n} \prod_{t=1}^{T_i} f(y_{it}|y_{it-1}, b_i, \theta)p(b_i|\Sigma_b)$$
$$\times p(\kappa) \prod_{i=1}^{n} \prod_{t=2}^{T_i'} p(m_{it}|\kappa, y_{it-1}), \tag{9}$$

where y is the the stacked vector of y_{it}, m is the stacked vector of m_{it}, and there are n individuals. A Gibbs sampling algorithm is used to estimate the parameters by partitioning the joint posterior into conditional distributions. The parameters of the expenditure and mortality models are simulated separately. κ can be sampled using the standard normal regression model by using the data augmentation technique described in Albert and Chib (1993).

Convenient priors were chosen for the expenditure model so sampling from β, σ_ϵ^2 and Σ_b is straightforward. α can be sampled using the same data augmentation technique used to sample κ. However, the conditional distribution for b_i is nonstandard so it is sampled using a random-walk Metropolis step. Full details of the Markov Chain Monte Carlo (MCMC) algorithm used to sample the parameters in the expenditure model are provided in Appendix C.

The algorithms were implemented using R version 3.1.2. Inferences for the mortality equation were based on $2,100$ draws of κ after a burn-in period of $1,000$ iterations simulated using the *MCMCprobit* command in the R package *MCMCpack*.[12] Estimation of the expenditure equation was more time consuming. 7 initially dispersed chains of $65,000$ iterations were run in parallel. The first $50,000$ iterations were discarded and the sequence was thinned by only keeping every 50th draw after burn in, yielding $2,100$ random draws from the posterior distribution. Run-time on a ThinkPad W530 Mobile Workstation with 8 cores is approximately 7 hours. Convergence was assessed using the R package coda by visually

[12]Estimates were nearly identical to those from maximum likelihood estimation. Run time was less than 1 minute and traceplots suggested that the parameters successfully converged.

inspecting traceplots of all the parameters (which are available upon request) and using the Gelman-Rubin diagnostic. The Gelman-Rubin potential scale reduction factor $\hat{R}$ was less than 1.1 for all parameters suggesting approximate convergence.

4.3 Parameter Estimates and Model Summary

Table 3 presents posterior means, standard deviations, and quantiles for the parameters of the expenditure model.[13] Unsurprisingly, the model predicts that women spend more out-of-pocket than men. Another expected result is that out-of-pocket costs are predicted to increase with age. However, the predicted effect of age on the probability of nonzero expenditures decreases as individuals get older while the predicted effect on positive out-of-pocket expenditures increases with age. The coefficients on the racial/ethnic and education variables are consistent with research on SES differentials in the utilization of healthcare services (e.g. Blanco et al. 2007; Cook, McGuire and Miranda 2007; Gross et al. 2008; Martinez et al. 2008; Cook and Manning 2009), which typically finds that variables correlated with low socioeconomic status are associated with less medical spending.[14]

The expenditure simulations described in the next section depend crucially on the coefficients on the lagged dependent variables and on the variance parameters. The impact of lagged spending in both parts of the model is fairly strong. In the binary component, d_{it-1} is predicted to increase d_{it}^* by 0.6 at its posterior mean. In the continuous component, each percentage point increase in lagged spending is predicted to increase current spending by 0.24 percentage points at its posterior mean when lagged expenditures are nonzero. Likewise, a 1% difference in nonzero initial spending between individuals is associated with a 0.20% difference in spending during period t. This strong association suggests that treating the initial conditions as fixed is inappropriate. The large negative coefficients on d_{it-1} and d_{i0} in the continuous component reflect the non-linear pattern between $d_{it-1} \times \ln y_{it-1}$ and

[13]The posterior distributions of the regression coefficients are all approximately normal so the posterior means are nearly identical to the posterior medians.

[14]Mediating variables like insurance status and income are not included so economic factors are allowed to contribute to these differences. Health variables are not included in the reported results either but the SES differentials are even more pronounced after controlling for self-reported health status.

Table 3: Parameter Estimates for Expenditure Model

			Posterior Quantiles		
	Mean	SD	2.5%	Median	97.5%
Binary component covariates					
Intercept	0.49	0.03	0.43	0.49	0.56
D_{it-1}	0.60	0.02	0.56	0.60	0.65
D_{i0}	0.94	0.03	0.88	0.94	1.00
$(Age-65)/10$	0.08	0.02	0.05	0.08	0.11
$[(Age-65)/10]^2$	−0.07	0.01	−0.09	−0.07	−0.06
Female	0.09	0.02	0.05	0.09	0.13
Black	−0.43	0.03	−0.49	−0.43	−0.38
Other race	−0.27	0.05	−0.37	−0.27	−0.16
Hispanic	−0.44	0.04	−0.51	−0.44	−0.36
Years of education - 12	0.07	0.00	0.06	0.07	0.08
Continuous component covariates					
Intercept	6.49	0.03	6.44	6.49	6.54
$D_{it-1} \times \ln y_{it-1}$	0.24	0.00	0.23	0.24	0.25
D_{it-1}	−1.34	0.04	−1.42	−1.34	−1.26
$D_{i0} \times \ln y_{i0}$	0.20	0.00	0.19	0.20	0.21
D_{i0}	−0.94	0.04	−1.02	−0.94	−0.85
$(Age-65)/10$	0.10	0.01	0.08	0.10	0.12
$[(Age-65)/10]^2$	0.03	0.00	0.02	0.03	0.04
Female	0.04	0.01	0.01	0.04	0.06
Black	−0.15	0.02	−0.19	−0.15	−0.11
Other race	−0.13	0.04	−0.20	−0.13	−0.05
Hispanic	−0.08	0.03	−0.14	−0.08	−0.03
Years of education - 12	0.02	0.00	0.02	0.02	0.03
Variance parameters					
σ_ϵ^2 (lognormal variance)	1.30	0.01	1.29	1.30	1.32
σ_1^2 (var$[b_{1i}]$)	0.53	0.03	0.48	0.53	0.58
σ_2^2 (var$[b_{2i}]$)	0.29	0.01	0.27	0.29	0.30
ρ (corr$[b_{1i}, b_{2i}]$)	0.52	0.02	0.48	0.52	0.56

Notes: Parameter estimates are from the model described in equations 1 and 2. d_{it-1} and d_{i0} are indicator variables for whether there was positive spending in the previous and initial wave respectively. Likewise, $\ln y_{it-1}$ ($\ln y_0$) is the natural logarithm of expenditures in the previous (initial) wave.

$d_{i0} \times \ln y_{i0}$ created by the degenerate distribution at zero (see appendix figures E.1 and E.2).

The size of the variances of both b_{1i} and b_{2i} suggest that unobserved heterogeneity is important. For instance, the posterior mean of σ_2 is $\sqrt{0.29} = 0.53$, which means that conditional on positive expenditures, a person with a value of b_{2i} one standard deviation above its mean would have out-of-pocket expenditures approximately 53% above average, given personal characteristics. Similarly, the intraclass correlations in both the first and

second part of the model are large: at posterior medians $\sigma_1^2/(\sigma_1^2+1) = 0.35$ and $\sigma_2^2/(\sigma_2^2+\sigma_\epsilon^2) = 0.18$.[15] In other words, unobserved heterogeneity explains about 35 and 18 percent of the unexplained variation in the binary and continuous components respectively. Finally, the high correlation between b_{1i} and b_{2i} of 0.516 highlights the importance of modeling the unobserved individual effects jointly.

Table 4 reports results for the mortality model. As in the expenditure model, the negative coefficient on d_{it-1} implies that there is a non-linear relationship between $d_{it-1} \times \ln y_{it-1}$ and the dependent variable (mortality) caused by the point mass at zero in medical expenditure data. Lagged log (nonzero) spending, $d_{it-1} \times y_{it-1}$, is fairly large in magnitude and precisely estimated. At posterior means and conditional on positive lagged expenditures, a 1% increase in lagged (nonzero) expenditures is predicted to increase the probability of mortality by approximately 0.02% for an individual with a predicted mortality rate of 0.09 (the sample average). The other coefficients are as expected. At older ages, mortality is an upward sloping quadratic function of age. Furthermore, women, non-blacks, and those with more years of education are predicted to survive to older ages. Finally, the large negative coefficient on the Hispanic variable is consistent with the well known "Hispanic paradox" in epidemiology in which Hispanics live longer lives despite their socioeconomic disadvantages.

In order to compare the predictive ability of the expenditure model to previous studies, I calculated conditional per period mean costs for each individual using the following formula,

$$y_{it}^*|\theta, b_i = \Phi(x_{1it}^T\alpha + b_{1i}) \exp(x_{2it}^T\beta + b_{2i} + \sigma_\epsilon^2/2), \tag{10}$$

where the second term is the mean of a lognormal distribution. To make predictions, I use the mean of the posterior distribution produced by equation 10, or $\hat{y}_{it}$. Predictions were made separately for a two-part model that did not include unobserved heterogeneity and the primary model with correlated unobserved individual effects. The fit of each model was

[15] For $t \neq s$, the intraclass correlations in the binary and continuous components are $corr(\epsilon_{1it}+b_{1i}, \epsilon_{1is}+b_{1i})$ and $corr(\epsilon_{2it} + b_{2i}, \epsilon_{2is} + b_{2i})$ respectively.

Table 4: Parameter Estimates for Mortality Model

			Posterior Quantiles		
	Mean	SD	2.5%	Median	97.5%
Intercept	−1.55	0.02	−1.59	−1.55	−1.51
$D_{it-1} \times \ln y_{it-1}$	0.12	0.00	0.11	0.12	0.13
D_{it-1}	−1.11	0.04	−1.19	−1.11	−1.02
$(Age - 65)/10$	0.29	0.01	0.26	0.29	0.32
$[(Age - 65)/10]^2$	0.07	0.00	0.06	0.07	0.08
Female	−0.21	0.01	−0.24	−0.21	−0.19
Black	0.11	0.02	0.07	0.11	0.15
Other race	0.04	0.04	−0.03	0.04	0.12
Hispanic	−0.20	0.03	−0.25	−0.20	−0.14
Years of education - 12	−0.03	0.00	−0.04	−0.03	−0.03

Notes: Parameter estimates are from the model described in equation 6. d_{it-1} is an indicator variable for whether there was positive spending in the previous wave. Likewise, $\ln y_{it-1}$ is the natural logarithm of expenditures in the previous wave.

summarized using the percent of the variation explained, or, $R^2 = 1 - \text{Var}(\hat{\text{y}}_{\text{it}} - \text{y}_{\text{it}})/\text{Var}(\text{y}_{\text{it}})$.

In the model with no unobserved heterogeneity, the R^2 of the model is only 0.07, which suggests that only a very small fraction of out-of-pocket costs can be predicted with the observed data. Adding unobserved heterogeneity doubles the proportion of variation explained by the model as the R^2 increases to 0.16, which is around the maximum percentage—15 to 20 percent—of variation in health costs that previous studies have been able to predict.

4.4 Model Checking

To evaluate the fit of a Bayesian model, Gelman et al. (2013) suggest using posterior predictive checks in which observed data, y, is compared to data generated from the model, y^{rep}. One way to draw replicated data is to simply use the draws of the parameters from the simulated posterior. However, a better test of the model in this paper is whether it can simulate an entire earnings history—while simultaneously accounting for mortality—by only using information available during period 0. I therefore use a simulation procedure that first randomly draws the parameter values from the joint posterior distribution, then simulates b_i from $N_2(0, \Sigma_b)$ (rather than using actual draws from the posterior of b_i), and finally recursively simulates y_{it} and m_{it} conditional on spending in period 0. Overall, $1,000$ simulations, y^{sim}, were generated for each of the $15,273$ individuals who survived to period

1.[16] (For complete details of the simulation procedure, see Appendix D.)

Figure 3 plots the density of a single 7-period simulation, y^{sim}, against the corresponding empirical density of y from waves 5 to 11.[17] The top-left figure is a simple plot of the cross-sectional distribution of out-of-pocket spending and the top-right figure is just the $log(x+1)$ transformation. The large spike at zero in the logarithm of spending plot illustrates the large proportion of nonspenders while the plot in levels is a testament to the skewed nature of the data. The simulated data captures the number of non-spenders accurately: the mean percentage of non-spenders across simulated datasets is 10.03% (95% CI: 9.35% to 10.73%) while 10.47% of individuals are non-spenders in the observed data. The log of positive spending is also approximately normally distributed which suggests that the normality assumption on the error term is valid. Finally, the simulated data matches the observed data very closely which implies that the model is able to simulate the cross-sectional distribution of spending.

The bottom two figures are within individual spending averages across the 7 periods, with expenditures in periods after death coded as \$0. Consistent with Table 1, the distribution of the 14-year averages is somewhat less skewed than the distribution of cross-sectional spending pooled across waves. The distribution of y^{sim} is again similar to the distribution of y so it would seem that it is possible to simulate the distribution of expenditures accurately.

Figure 4 looks at whether the mortality model accurately predicts the number of survey respondents surviving from one wave to the next. The figure reports the proportion of the $15,273$ individuals alive during wave 4 still alive during each subsequent wave. The model tends to slightly underpredict survival in earlier waves but by wave 11 the simulated survival rate is nearly identical to the observed rate.

To asses whether the model adequately captures persistence in out-of-pocket expenditures, Figure 5 considers the distribution of expenditures in wave $t+4$ conditional on the

[16] Since the simulation accounts for mortality the length of the vector y^{sim} is random.

[17] There is very little variation in the distribution of spending from one simulation to the next.

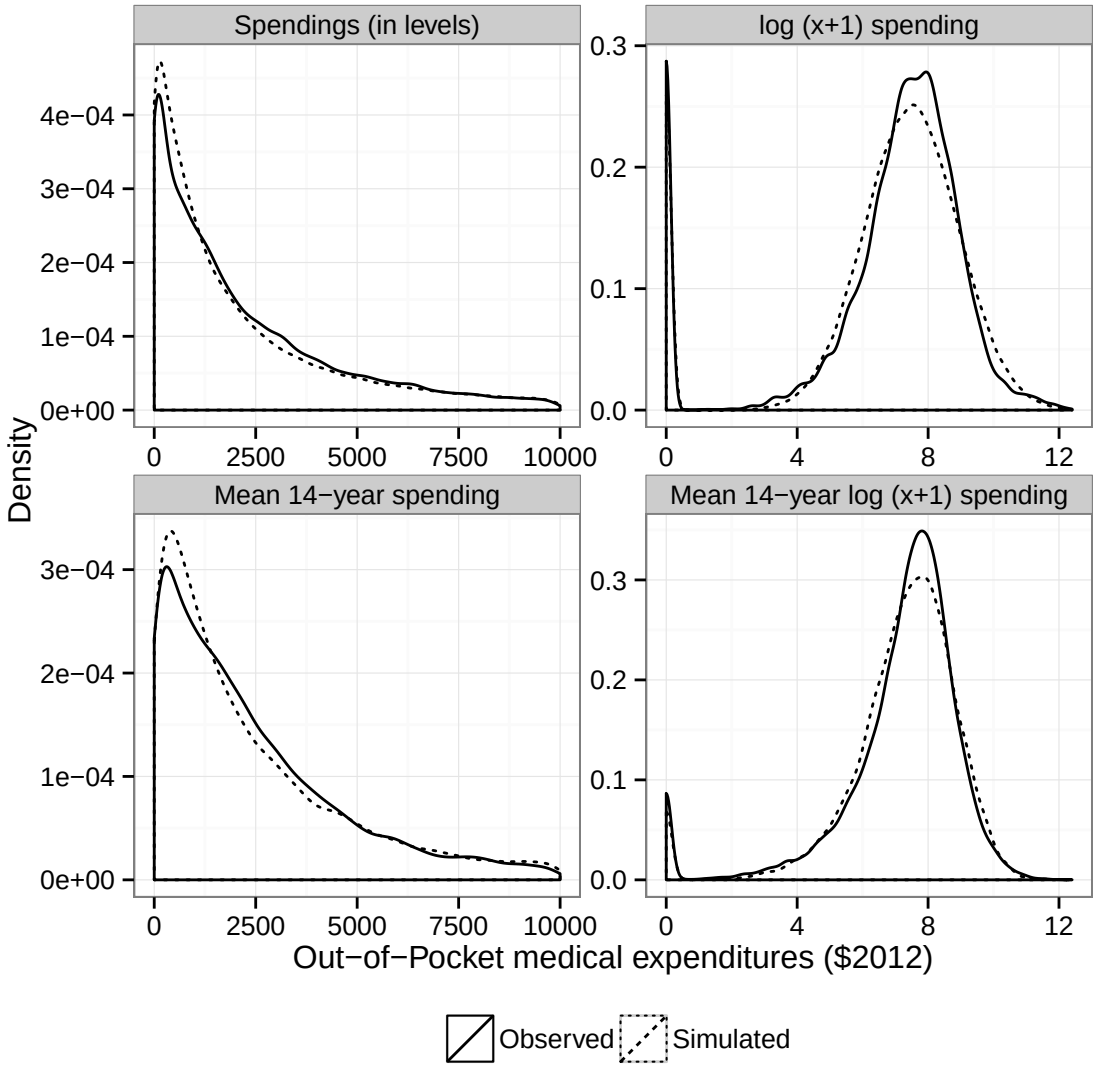

Figure 3: Out-of-Pocket Expenditure Density Estimates

Notes: Spending (in levels) is truncated at $10000 (approximately the 95th and 96th percentile of mean 2-year and 14-year spending respectively) for graphical purposes.

distribution of expenditures in wave t and survival until wave $t+4$.[18][19] The plot is similar to Figure 2 in that it presents the evolution of an individual's expenditures across waves. Each of the five panels in the plot shows the proportion of individuals in one of five spending

[18]For example, this compares expenditures in wave 5 to wave 9, wave 6 to wave 10, and wave 7 to wave 11.

[19]Results from other wave differences (i.e. comparing waves t and $t+1$) are similar.

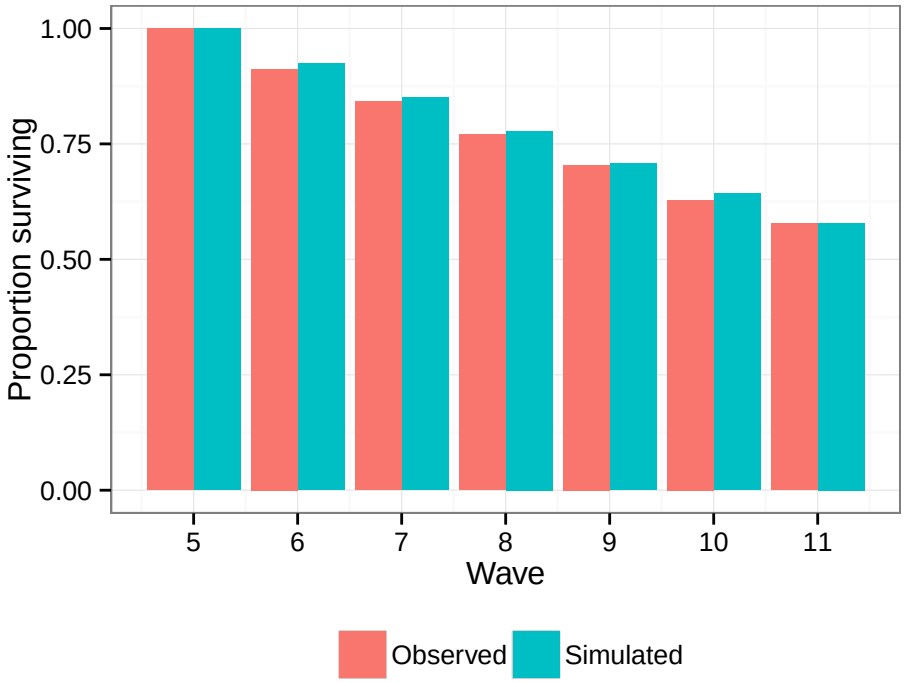

Figure 4: Survival Estimates

quintiles at wave $t+4$ given their spending quintile at wave t.

The results show that there is significant mean reversion in spending but that spending is still somewhat persistent over time. Individuals in the first expenditure quintile in wave t are much more likely to remain in the first quintile 4 waves later. Likewise, individuals in the highest quintile in wave t are more likely to remain in the higher quintiles in wave $t+4$. That said, the majority of individuals move to new spending quintiles.

The simulated data tends to replicate these patterns but there are a few discrepancies. For instance, compared to the observed data, individuals simulated to be in the fourth quintile in wave t are slightly less likely to remain in the fourth quintile in wave $t+4$. That being said, the model does a good job of simulating transitions from one quintile to another. For additional model checks see Appendix F.

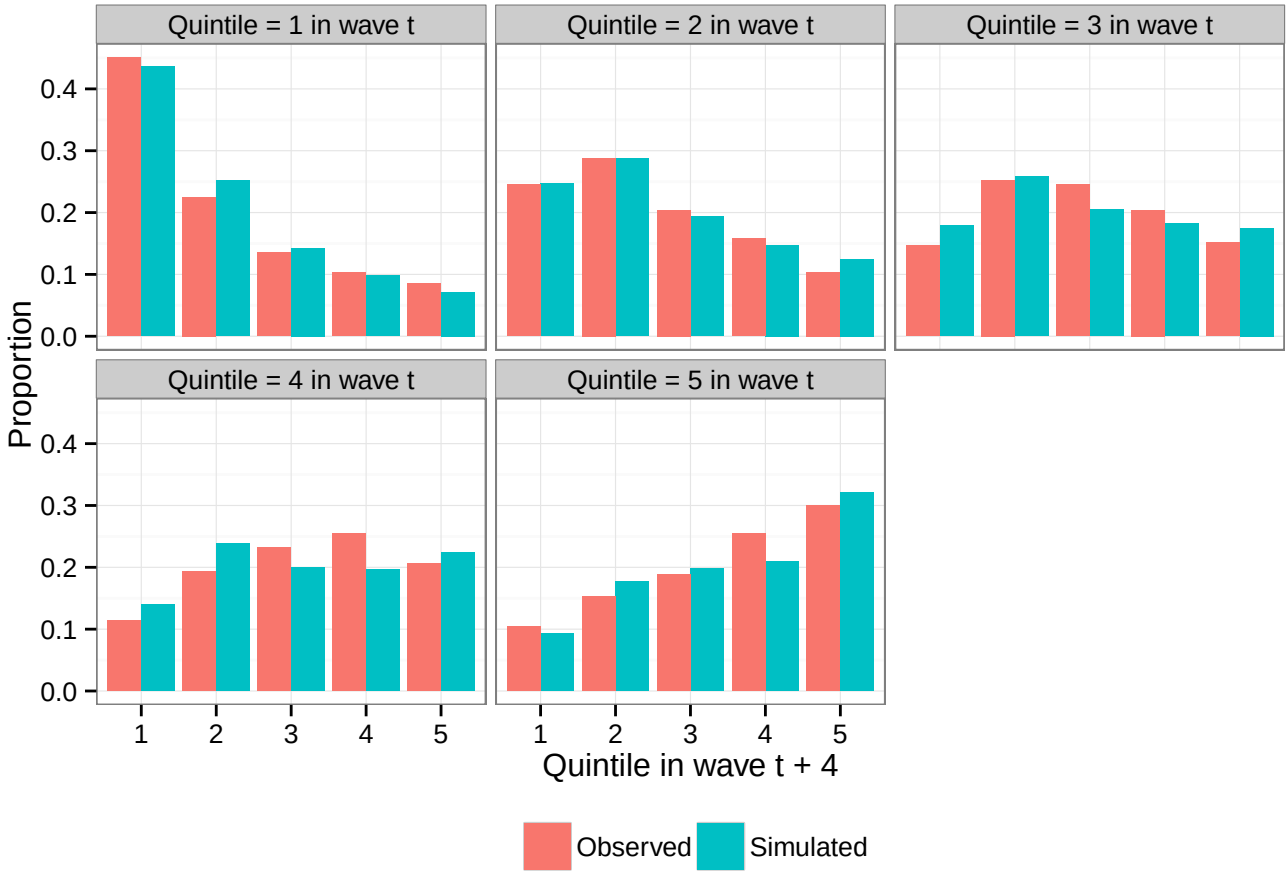

Figure 5: Distribution of Wave t + 4 Expenditures Conditional on Wave t Expenditures

Notes: The figure only contains expenditure data for those who survive to wave $t + 4$.

5 Long-term Expenditure Risk

5.1 Equity

How unequal is the distribution of out-of-pocket expenditures? Most studies analyzing this question have focused on the distribution of expenditures over short periods of time (typically a single year) (e.g. Berk and Monheit 2001; Monheit 2003; Stanton and Rutherford 2006). Here, I compare the distribution of out-of-pocket expenditures in a single (two-year) period

to the distribution of out-of-pocket expenditures over multiple periods.

One way to measure spending inequality is by using the approach developed in the income inequality literature. The standard graphical representation of inequality is a Lorenz curve which plots the cumulative share of spending on the y-axis and the cumulative portion of the population on the x-axis. The figure also typically contains a 45 degree line called "the line of perfect equality", which depicts the distribution of expenditures when everyone spends the same amount.

Lorenz curves for the raw HRS data are shown in Figure 6. A curve is plotted for both expenditures in a single wave and mean expenditures over waves 4 to 11 per wave alive. Estimates of single-wave spending are taken from each individual's midpoint period, or period $T_i/2$ rounded to the smallest integer not smaller $T_i/2$. The figure shows that the distribution becomes more equal when considering long-term spending patterns, but that out-of-pocket spending is still highly concentrated. For example, in a particular two-year wave the top 5% of spenders account for 50% of total spending; over 8 waves the top 5% of spenders still account for 38% of total spending.

One way to quantify inequality is with the Gini coefficient. The Gini coefficient is the ratio of the area between the line of perfect equality and the Lorenz curve to the total area under the line of perfect equality. Its value is 0 when everyone spends the same amount and 1 when only one person has positive expenditures. The Gini coefficient for 16-year spending (0.63) is substantially lower than for single-wave spending (0.74). Yet, both are fairly close to 1 which suggests that spending inequality persists over long periods of time.

The long-term out-of-pocket spending distribution might be more equal if one examines total usage over the 16-year period since high cost users are more likely to die and become zero-cost users. However, coding expenditures after death as \$0 has a negligible impact on the inequality estimates as the Gini coefficient only decreases from 0.63 to 0.60. Thus, long-term estimates of inequality are not very sensitive to whether spending is tracked until death or over a given period of time.

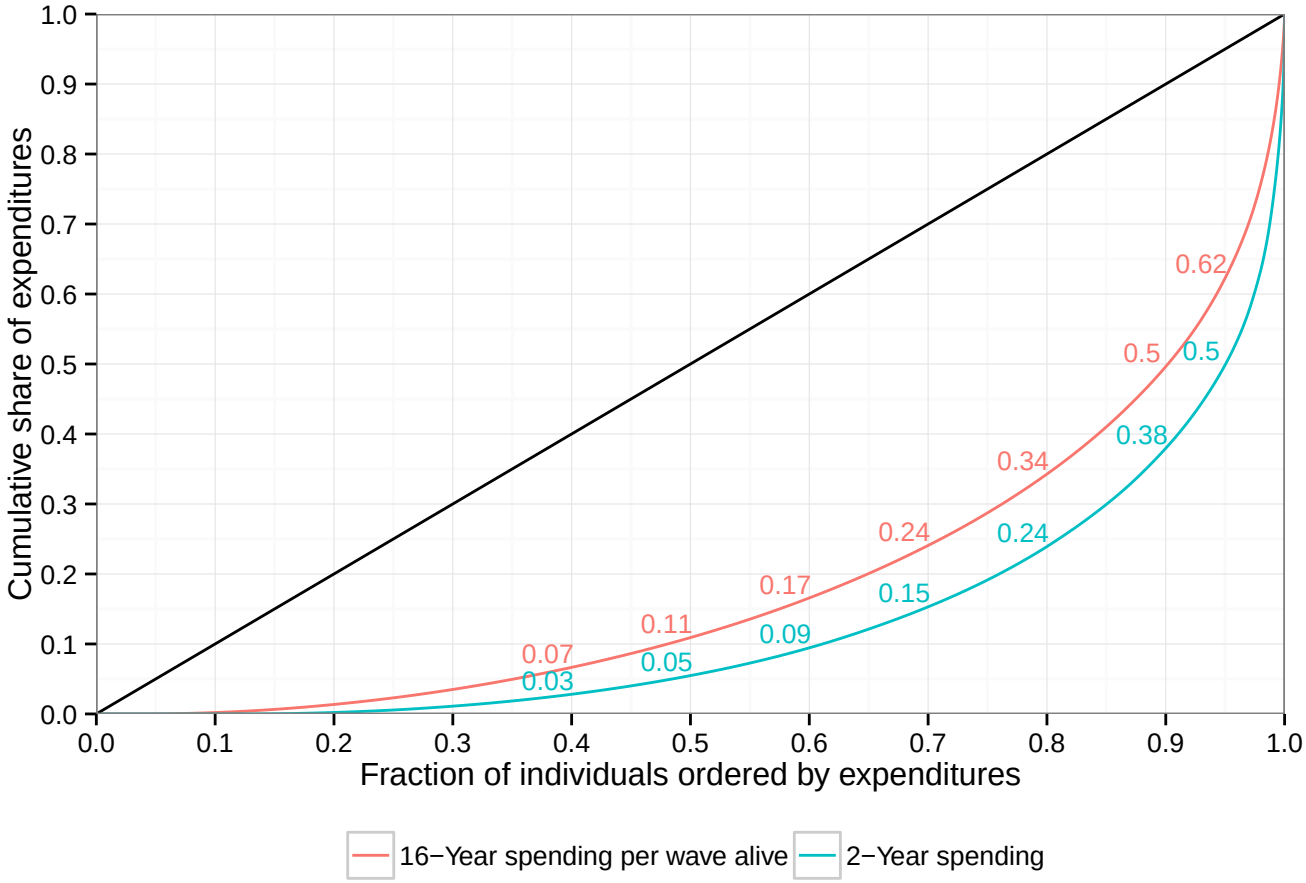

Figure 6: Empirical Lorenz Curves for Out-of-Pocket Expenditures per Wave Alive

Notes: The 45 degree line is the "line of perfect equality", which depicts the case where everyone spends the same amount out-of-pocket on healthcare.

A related question is whether the distribution of spending over a given number of periods would becomes more equal as the number of periods increased. To examine this question, I simulated expenditures (and mortality) over 13 periods (26 years) for individuals age 60 or below during wave 5. Lorenz curves for spending per period alive were then calculated in a rolling fashion over different period lengths.

Figure 7 plots some of these simulated Lorenz curves. The figure shows that increasing the number of periods reduces inequality in out-of-pocket expenditures but that inequality

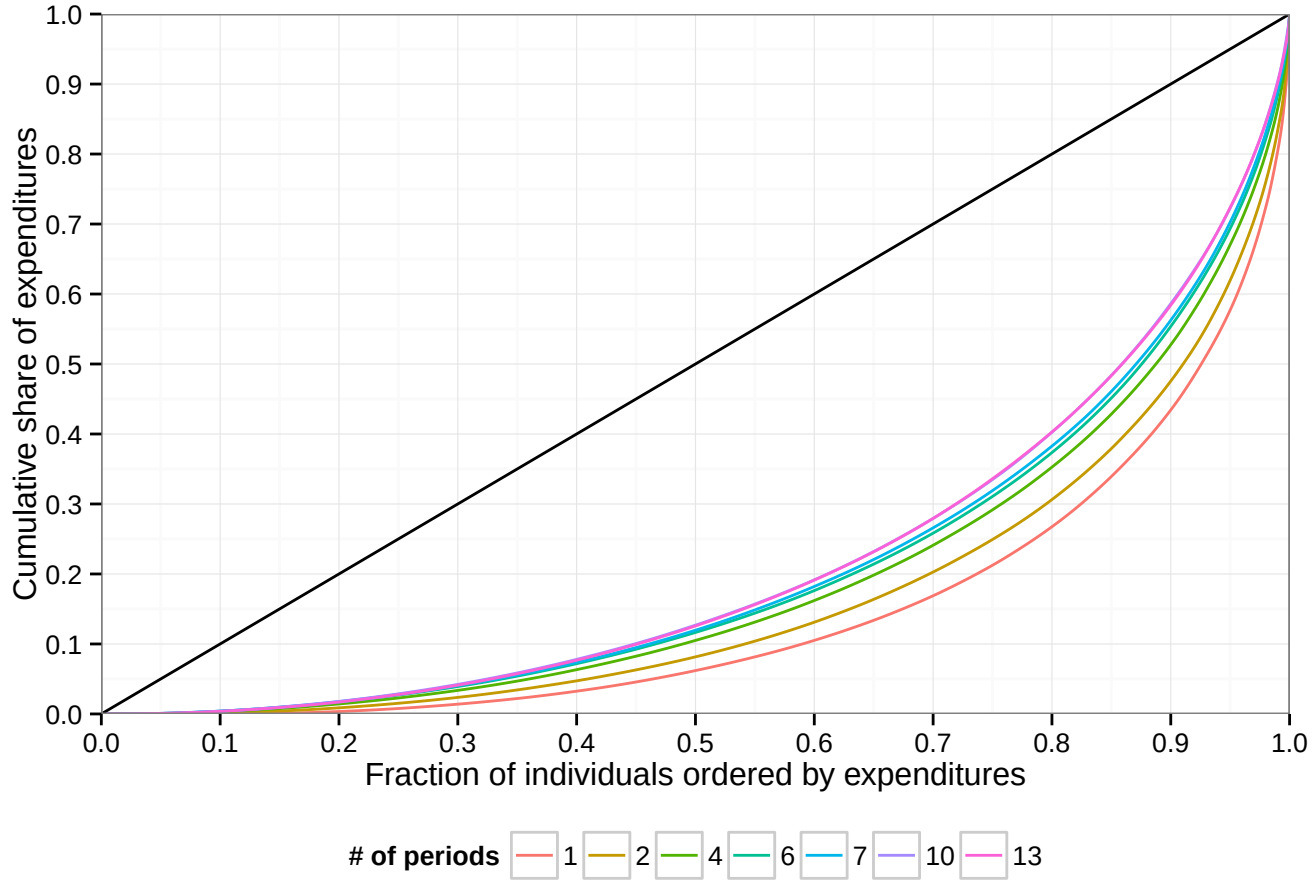

Figure 7: Simulated Lorenz Curves for Out-of-Pocket Expenditures per Period Alive, Age $\leq$ 60 in Period 1

Notes: Expenditures for individuals less than or equal to age 60 in wave 5 were simulated for various lengths of time using the simulation procedure described in the text. The Lorenz curves are calculated using spending averaged across the number of periods in which an in individual is simulated to be alive. The 45 degree line is the "line of perfect equality" in which every individuals spends the same amount out-of-pocket on healthcare.

decreases at diminishing rates as the number of periods increases. Once spending is measured over 6 periods additional increases in the number of periods have very little impact on the distribution of spending.

Differences in expenditures across individuals are driven by three primary factors: predictable variation, unobserved heterogeneity and random shocks. In order separate predictable variation from variation due to uncertainty, three cases were simulated. In the first

case, out-of-pocket spending variation is driven entirely by predicted variation from observable individual characteristics. Predictions are made using equation 10, without accounting for state dependence, initial conditions, or unobserved individual effects. In the second case, expenditures are still driven entirely by predictable variation, but spending is persistent because the model includes lagged dependent variables and initial conditions. The third case simply estimates out-of-pocket expenditure variation using the simulation procedure used throughout the paper.[20]

Figure 8 plots Lorenz curves from data generated from these three cases. Spending was simulated over 7 periods in each case and average spending per period alive was computed for each individual. The legend is ordered from left to right in decreasing order of the Gini coefficient. The spending distribution is the most equal when there is no uncertainty or persistence, and variation is driven entirely by observable characteristics. The distribution of spending becomes more unequal when state dependence is incorporated into the simulation, but the transitory shocks and unobserved individual effects contribute the most to spending differences. It is no surprise that the model estimated in the paper generates a distribution that is considerably less equal than the distributions generated from predicted means.

One potential consequence of highly concentrated out-of-pocket spending is that certain individuals might be more at risk for catastrophic healthcare expenditures. It is of course an open question as to what constitutes a catastrophic expenditure. To this end, researchers have come up with various ways to measure financial protection in health systems. One measure suggested by Waters, Anderson and Mays (2004) for analyses in the United States is whether out-of-pocket expenditures exceed 10% of family income. Using this (and two other) definitions of financial protection, Waters, Anderson and Mays (2004) argue that there is a lack of financial protection in the United States—particularly for poor families and those with chronic conditions. The results reported so far would seem to suggest that this risk remains significant over longer periods of times as well.

[20] A separate model was estimated for each case. The first two cases use two-part models without unobserved heterogeneity and the third case uses the model described in the text.

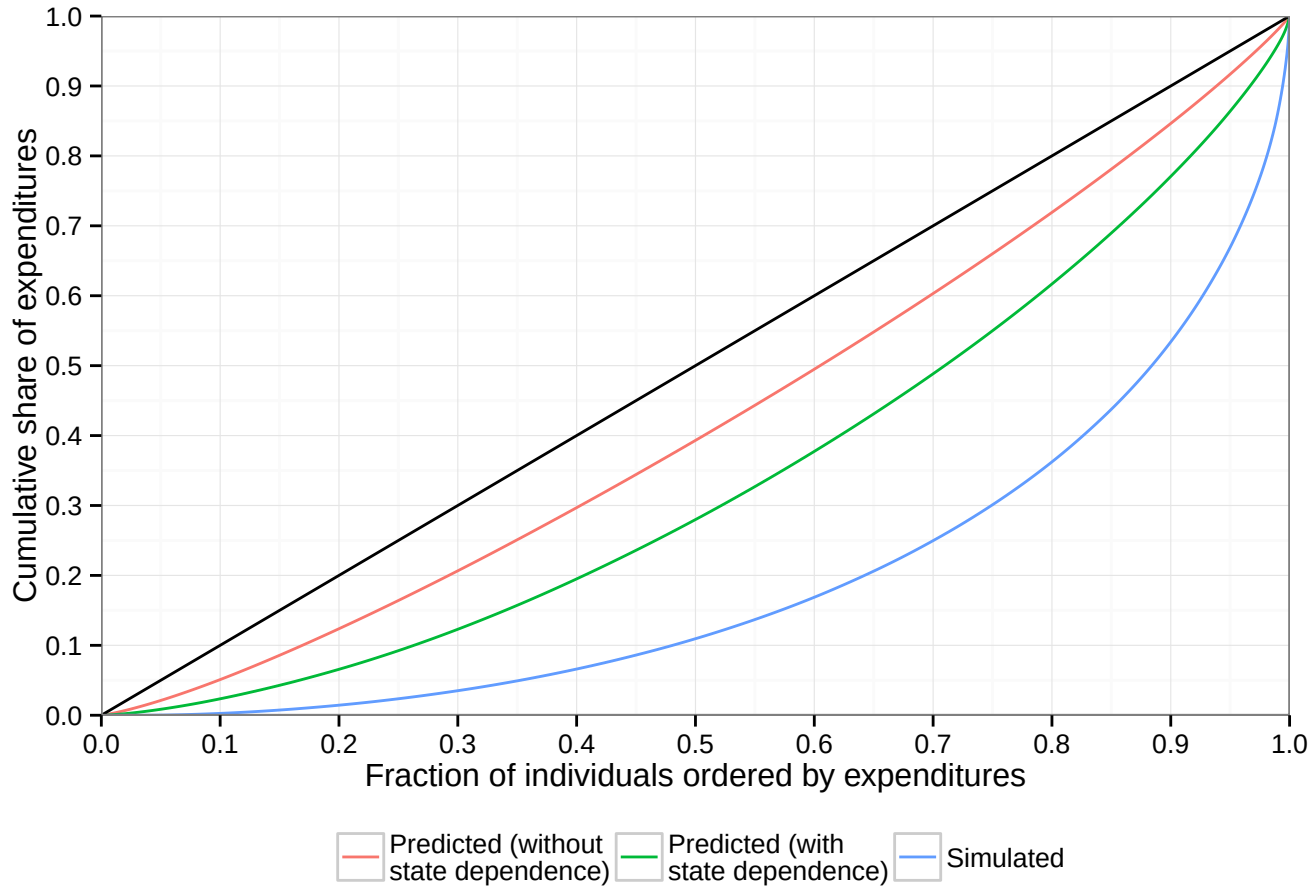

Figure 8: Simulated Lorenz Curves for 14-Year Out-of-Pocket Expenditures per Period Alive, by Type of Risk

Notes: The 45 degree line is the "line of perfect equality", which depicts the case where everyone spends the same amount out-of-pocket on healthcare.

Figure 9 looks at whether this is the case by plotting the empirical CDF of out-of-pocket expenditures as a percentage of family income.[21] The CDF's are plotted for both a single 2-year wave (the midpoint wave) and over 16 years. The CDF's are then further divided into four (annualized) income quartiles.

The distribution of spending as a percentage of income is, in general, not very sensitive to

[21]Some households reported earning \$0 in some wave, which is unrealistic because they should have earned income from either welfare or social security. Incomes are consequently truncated from below at the 1st percentile, although this has no impact on the results.

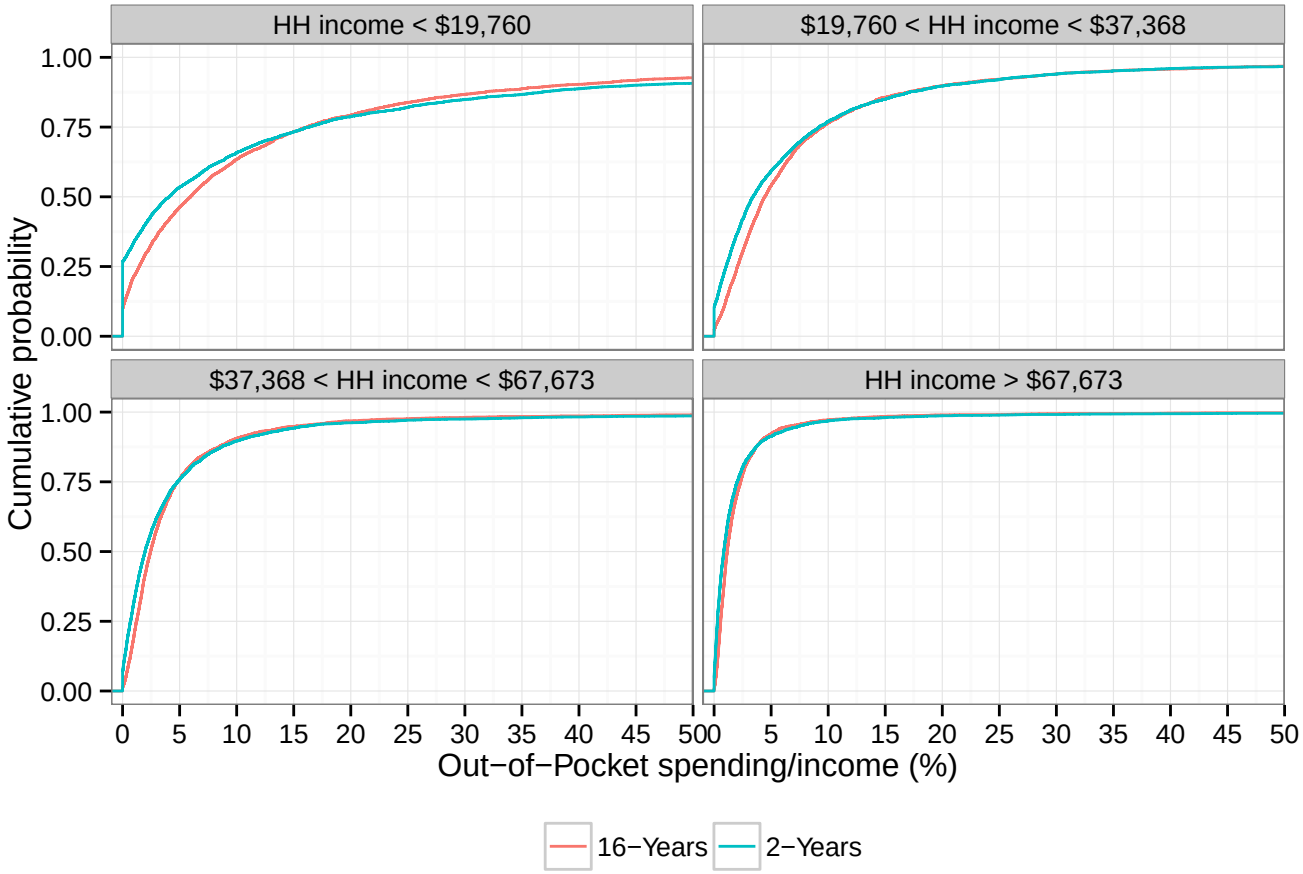

Figure 9: Empirical CDF of Out-of-Pocket Expenditure as a Percentage of Household Income by Income Quartile

Notes: Reported income quartiles are annualized.

whether the distribution is measured in a single wave or over 8 waves. The upper end of the distribution of out-of-pocket spending as a percentage of income is very similar regardless of income quartile or of the time period over which it is measured. For instance, 21% of the poor spend over 20% of income in a single wave; over 8 waves, 21% do as well. But at the far right tail, financial risk is somewhat less catastrophic over longer periods of time. For example, low-income individuals at the 95th percentile spend 96% of their income on out-of-pocket healthcare costs in a single wave; over 8 waves this number drops to 69%. Even so, healthcare costs are still a significant financial risk—especially for the poor— over long

periods of time.

At the lower end of the distribution, healthcare spending as a percentage of income can actually be larger over long periods of time—again, especially for the poor. This occurs because a smaller portion of individuals have very small healthcare costs over extended periods. For individuals in the lowest income quartile a larger fraction of survey respondents spend approximately 5% of their income on out-of-pocket healthcare when measured over 8 waves than in a single wave. Likewise, the median low-income individual spends 4% of their household income on on out-of-pocket health costs in a single wave and 6% over 8 waves.

5.2 Uncertainty and Welfare

So far this paper has shown that there is significant variation in the long-term distribution out-of-pocket expenditures and that out-of-pocket spending in a given two-year period is largely unpredictable. In this section, I provide evidence on the degree of uncertainty in long-term spending. I then examine the extent to which future spending uncertainty makes it difficult to target high spenders. Lastly, I use a stylized utility framework to provide a rough estimate of potential welfare gains from eliminating exposure to risk.

To investigate the predictability of long term spending I estimated expected out-of-pocket expenditures over a 14-year period for each individual conditional on period 0 information. This was done by calculating the mean of the $1,000$ simulations described in Section 4. I then repeated the simulations assuming that each individual knew the value of his or her b_i. More precisely, in the first simulation the b_i's were drawn from their $N_2(0, \Sigma_b)$ distribution and in the second simulation the actual posterior draws of b_i's were used.

Table 5 examines the amount of variation in actual 14-year spending that each simulation is able to predict. The figure reports the R^2 value described in Section 4 for total 14-year out-of-pocket expenditures and out-of-pocket expenditures per period alive. Estimates are then further sub divided by whether the simulations estimated mortality or the simulations used the observed life expectancy for each individual.

Table 5: R-squared for 14-Year Out-of-Pocket Spending, by Simulation Type

	Total out-of-pocket expenditures		Out-of-pocket expenditures per period alive	
Simulation type	Simulated mortality	Observed mortality	Simulated mortality	Observed mortality
Unknown b_i	0.05	0.07	0.10	0.10
Known b_i	0.35	0.41	0.37	0.36

The main takeaway from the table is that only a small fraction of the variation in spending can be predicted using period 0 information. The R^2 values are generally quite low when there is unobserved heterogeneity (row 1 in the table). The simulation can explain around 10% of the variation in spending per period alive but less than 7% in total expenditures. On the other hand, when observed values of b_i are used in the simulations, the percent of explained variation increases noticeably. The model can explain over 35% of the variation in 14-year spending per wave alive and in total expenditures. But even so, a large fraction of the variation in long-term spending remains unaccounted for.

The low R^2 values suggest that it may be difficult for disease management programs to target high spenders. Figure 10 investigates this by comparing observed spending by predicted high spenders to observed spending by actual high spenders. As in Table 5, predictions are made using expected 14-year expenditures from the simulation. A high spender is defined as someone with spending above a minimum spending threshold. Predicted expenditures are larger than the threshold for predicted high spenders while observed expenditures are larger than the threshold for actual high spenders.

The figure considers a number of thresholds ranging from spending at the 70th to 99th percentile. At each minimum spending percentile, the figure divides observed spending by predicted high spenders by observed spending by actual high spenders. This ratio is the fraction of spending by high spenders that the model can account for and is an upper bound on the fraction of spending that disease management programs could eliminate.

As expected, the model can account for a much larger fraction of out-of-pocket expenditures when heterogeneity is known. The spending ratio decreases at an increasing rate as

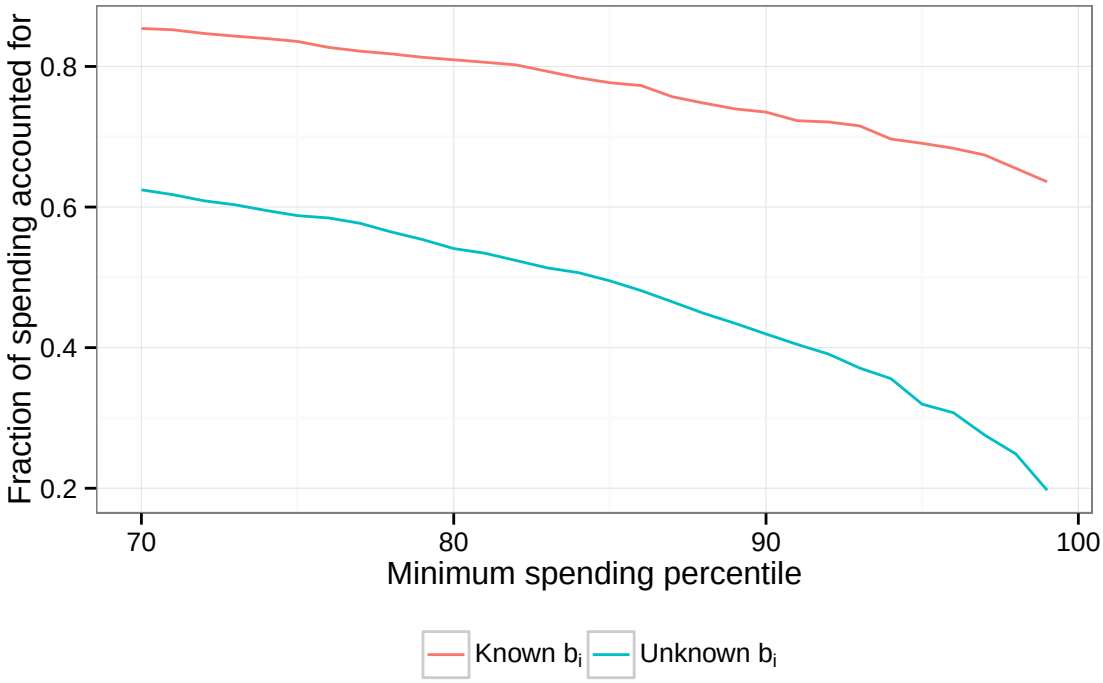

Figure 10: Targeting High Spenders

Notes: Predicted (actual) high spenders are individuals with predicted (actual) expenditures above a particular "minimum spending percentile." The "fraction of spending accounted for" is the sum of observed spending by predicted high spenders divided by the sum of observed spending by actual high spenders.

the minimum spending percentile increases, which suggests that there is a tradeoff between using more resources per individual and correctly identifying high spenders. For example, when b_i is unknown, the spending ratio decreases from 0.62 when the minimum threshold is at the 70th percentile to 0.20 when it is at the 99th percentile.

Although uncertainty makes targeting high spenders difficult, it also makes reducing risk more valuable. To quantify these benefits, I estimate the potential welfare gains from completely eliminating uncertainty. My approach is similar to a number of other studies that have estimated the utility gains associated with specific insurance plans (McClellan and Skinner 2006; Brown and Finkelstein 2008; Finkelstein and McKnight 2008; Engelhardt and Gruber 2011). In particular, I estimate welfare by calculating each individual's risk premium, or the maximum amount that they would be willing to pay at the beginning of

period 1 to completely insure against future risk.

Following Finkelstein and McKnight (2008) and Engelhardt and Gruber (2011), individuals are assumed to receive utility from consumption, c_t, which is subject to a per period budget constraint, $c_t = I_t - y_t$, where I_t is income and y_t is out-of-pocket spending. Over T periods, individuals are assumed to receive utility from the present value of consumption, $\sum_{t=1}^{T} \delta^t c_t$, where δ_t is the discount rate. The risk premium, π, is the amount of money an individual would need to receive to become indifferent between a world in which out-of-pocket costs are uncertain each period and a world in which he or she would pay the expected present value of future spending over T periods.[22] Given a utility function, $U(\cdot)$, and probability density function, $f(\cdot)$, π is the solution to the implicit equation,

$$U\left(\sum_{t=1}^{T} \delta_t[I_t - E(y_t)] - \pi\right) = \int U\left(\sum_{t=1}^{T} \delta_t[I_t - y_t]\right) f(y_t)dy_t. \tag{11}$$

This setup is obviously a simplification of reality since it does not allow for borrowing, saving, or non-income wealth. However, there are a number of reasons to believe that it might not be too unrealistic. First, a number of studies have shown that the relationship between consumption and age is humped shaped—consumption increases during young adulthood before peaking around age 45 and then declines until death—and tracks income closely (Gourinchas and Parker 2002; Fernández-Villaverde and Krueger 2007).[23] Second, it has been estimated that nearly four-fifths of net wealth is bequeathed (Kopczuk and Lupton 2007), so it is unreasonable to assume that individuals consume all of their wealth over their lifetime.

To solve the model, individuals are assumed to have a constant relative risk aversion

[22] Paying a fixed sum equal to the expected present value of future spending is equivalent to paying an actuarially fair insurance premium each period (a premium is actuarially fair if the expected present value of benefits is equal to the expected present value of premiums).

[23] Note that these empirical findings are inconsistent with the permanent income hypothesis. Common explanations within an expected utility maximization framework are that individuals have high discount rates, only save as a precaution against sharp falls in income, or that they are liquidity constrained. Shefrin and Thaler (1988) provide an alternative explanation in which individuals do not base their consumption choices on maximizing their lifetime utility. Instead, individuals follow a rule of thumb in which they they spend current income and only dip into their savings if absolutely necessary.

(CRRA) utility function,

$$U(C) = \begin{cases} \frac{1}{1-\gamma} C^{1-\gamma} & \text{if } \gamma > 0, \gamma \neq 1 \\ \ln C & \text{if } \gamma = 1, \end{cases} \tag{12}$$

where γ is a risk aversion parameter. Future consumption is discounted at 3% per year; that is, $\delta_t = 1/(1+0.03)^{2t-2}$. The integral in the second term of equations 11 is evaluated by using the probability distribution of out-of-pocket expenditures estimated from the Bayesian model. For instance, using the 1000 simulated values, y^{sim}, expected utility for the ith individual in period t is just $(1/1000)\sum_{s=1}^{1000} U(I_{it} - y_{it}^{sim,s})$. To remain consistent with previous studies (e.g. Finkelstein and McKnight 2008; Engelhardt and Gruber 2011), out-of-pocket spending is truncated from below at 0 and above at 80% of income. In addition, incomes are treated as fixed and truncated from below at the 1st percentile.[24] Household incomes are converted to individual incomes using the OECD equivalence scales.[25]

Figure 11 reports 14-year risk premiums for the entire sample of 15,273 survey respondents analyzed in this paper. The figure plots empirical CDF's of the risk premiums for $\gamma = 2$ and $\gamma = 3$, which are common estimates in the literature.

Variability across individuals stems from three primary factors. First, welfare gains are larger for poorer individuals because the CRRA utility function implies decreasing absolute risk aversion (DARA); that is, risk premiums are decreasing in income. Second, risk premiums are larger when individuals live longer lives. Third, risk premiums increase when the variance of out-of-pocket expenditures increases. This in turn implies that risk premiums are increasing in predicted (mean) out-of-pocket expenditures because the distribution of nonzero expenditures is assumed to have a lognormal distribution.

[24] Since incomes are set equal to observed values it is necessary to set mortality to observed values as well; otherwise income would need to be forecasted. See footnote 21 regarding truncating income at the 1st percentile.

[25] Equivalence scales measure the change in consumption needed to keep the welfare of a family constant when its size changes. OECD equivalence scales place a weight of 1 on the household head, 0.7 on subsequent adults over age 14, and 0.5 on each child under 14. Child ages are unavailable in the RAND HRS data so each household member other than the household head is assigned a weight of 0.7.

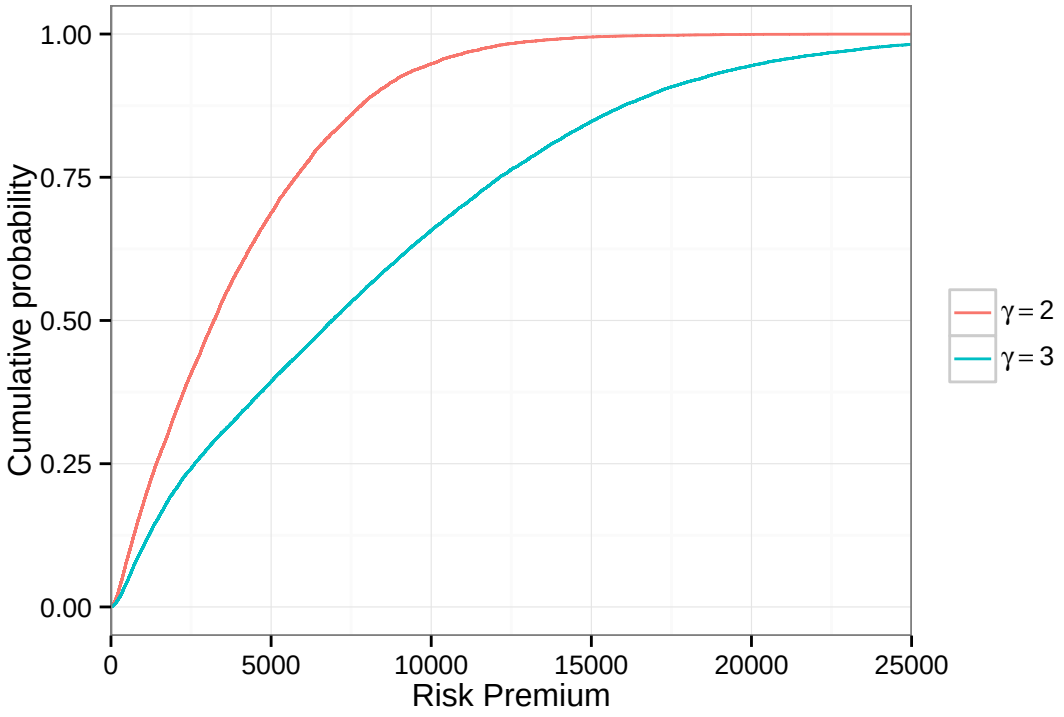

Figure 11: Empirical CDF of Estimated Risk Premiums

Notes: Risk premiums are calculated assuming a CRRA utility function. Individuals are assumed to be more risk averse when the risk aversion parameter, γ, is larger.

For a relative risk coefficient of 2, the long-term per period risk premium for the median individual is $\$3,215$ while the risk premium for an individual at the 90th percentile is $\$8,349$; for a risk premium of 3, these quantities increase to $\$6,898$ and $\$17,103$ respectively. These estimates are fairly large, which implies that there are potential welfare gains from additional insurance, especially for those at the upper end of the distribution. That said, any welfare gains from additional public insurance must be weighed against inefficient increases in spending due to moral hazard.

6 Conclusion

This paper analyzes out-of-pocket expenditures over a much longer time period than in previous studies. It develops a dynamic Bayesian models of out-of-pocket expenditures and

mortality and applies it to the HRS. It finds that the data generating process for expenditures is well represented by a dynamic two-part longitudinal model with correlated unobserved individual effects, and that mortality can be modeled using a probit model that is conditional on previous expenditures. The model accounts for state dependence and three potential sources of error: parameter uncertainty, unobserved heterogeneity and a transitory shocks.

The analysis has some important implications for cost containment policies. Expenditures are difficult to predict using observed data alone: both unobserved heterogeneity and transitory shocks explain a large proportion of the variance in spending. This suggests that it may be difficult for clinically oriented methods such as disease management programs to identify high spenders. At the same time, the uncertain nature of future costs makes health insurance valuable and suggests that individuals using financial instruments like health savings accounts may still be subject to significant financial risk. The distribution of spending also remains highly unequal—even over long periods of time—which means that increasing cost sharing would likely create policy pressure to redistribute healthcare costs in other ways. For consumer driven healthcare to succeed, it must successfully balance the benefits of reduced moral hazard against the costs of increased inequality in out-of-pocket spending and reduced welfare from risk reduction.[26]

There are many avenues for future research. First, the model could be extended in a number of ways. For example, analyses focusing on financial planning should model wealth alongside mortality and out-of-pocket expenditures. In addition, one could allow survival to depend on the unobserved individual effects in the expenditure model (e.g. Wulfsohn and Tsiatis 1997; Ibrahim, Chu and Chen 2010). Finally, the model should be applied to new data sources. There is, in general, a need for high quality longitudinal datasets of healthcare spending. Different mortality models might be appropriate if the data were less aggregated; other data may be less normal and require a relaxation of the lognormality assumption.

Second, it would be useful to analyze individuals with specific chronic conditions using

[26] Another potential cost is that increased cost sharing could cause individuals to forgo necessary medical care.

longitudinal data. The analyses in this paper do not use detailed clinical data and certain chronic conditions might be more suitable for disease management programs than suggested in this paper. Indeed, Monheit (2003) has shown that spending is more persistent when individuals have certain medical conditions like diabetes or mental disorders. Allowing persistence to depend on health status would aid efforts aimed at identifying individuals in need of medical and financial management.

Third, researchers should let long-term out-of-pocket spending depend on cost-sharing mechanisms like copayments, coinsurance and deductibles since there is convincing evidence that spending is sensitive to cost sharing (Manning et al. 1987; Aron-Dine, Einav and Finkelstein 2013). These behavioral responses could be incorporated into the simulation procedure used here and built into dynamic utility-based models (e.g. Brown and Finkelstein 2008) to analyze the long-term implications of specific health insurance reforms. This would, in turn, allow researchers to quantify the competing welfare effects of risk reduction and moral hazard. If possible, these simulations should allow the price elasticity of health expenditures to vary across individuals (e.g. Kowalski 2015) in order to properly analyze the distributional implications of different insurance policies.

Appendices

A Out-of-Pocket Spending Data

The HRS collects data on out-of-pocket spending in the following manner. First, each respondent is asked to provide a continuous estimate of spending in a number of medical categories. If the respondent is unsure they are then asked a series of unfolding bracket questions where they are asked whether spending falls within a number of ranges. According to careful analyses of the data by (Goldman, Zissimopoulos and Lu 2011), non-response on the continuous question is typically around 20%, although it can reach as high as 30% in

some categories such as hospital and nursing home care. The unfolding bracket question reduces the number of non-responses to below 5% for most categories.

The RAND HRS data files impute out-of-pocket expenditure data for all individuals in the sample. RAND uses a matching imputation method. They first use a linear regression to model the inverse hyperbolic sine transformation of out-of-pocket expenditures and then use the regression to predict out-of-pocket expenditures for every individual in the survey. These predicted values are then used to impute continuous expenditure values for respondents who answered the unfolding bracket questions but did not report a continuous expenditure value. More specifically, each individual whose expenditures were reported to fall within a given bracket were assigned the expenditure value of the respondent with the closest predicted expenditure value whose actual (continuous) expenditures fell with that bracket. A slightly different technique was used to impute data for open ended brackets. For further details on the imputations see [see St Clair et al. (2014, section 7.5, pp. 18–21)].

B Sample Selection

Here, I discuss two potential problems with dropping observations from the sample. The first problem is that the sample may become less representative of the population, which limits external validity. Table B.1 examines this by comparing three groups of survey respondents. Respondents in the first group have complete data until death or through wave 11. Members of the second group have missing data in some years (due to non-response), but still participated in the survey until either death or wave 11. Finally, survival times for individuals in the third group are right censored because they dropped out of the HRS completely.

Means of out-of-pocket expenditures, survival times, and time constant observed characteristics are reported for each group. The out-of-pocket expenditure statistics reported in the table are means of individual per wave averages from waves 4 to 11. Mean out-of-pocket expenditures for non-respondents and respondents with complete data are similar, although

those with complete data spend a little bit more. Non-respondents were three years younger during the initial wave, so they predictably survived 1 wave longer. Observed characteristics are also similar between the two groups.

Table B.1: Sample selection in the HRS

	Complete data	Non-respondents	Dropouts
Out-of-pocket expenditures	4889.79	4393.81	2919.36
Survival time (number of waves)	6.00	7.08	4.66
Initial age	67.41	64.32	63.47
Black	0.14	0.18	0.09
Female	0.57	0.58	0.55
Hispanic	0.07	0.09	0.08
Years of education	11.84	11.87	12.42

Notes: Means of observed characteristics are averages across individuals. Survival time refers to the number of waves in which a survey respondent was tracked by the HRS before either death or wave 11. Out-of-pocket expenditures are means of individual per wave averages.

Expenditures are lower for dropouts, but this is likely because they were not observed during the end of the 16-year survey period. This biases the out-of-pocket expenditure estimate downward because spending increases with age and prior to death. In fact, there are a few reasons to believe that excluding dropouts will only have a minor impact on the sample. First, differences between dropouts and individuals in the other two groups decrease significantly if one restricts the sample to those who did not die during the survey: survivors with complete data spent $\$3,434$ while non-respondent survivors spent $\$3,826$. Second, observed characteristics are generally similar across all three groups, although dropouts tend to be slightly better educated, a little more likely to be male, and less likely to be black. Finally, excluding dropouts only eliminates 820 individuals. It is important to note however that dropping right censored observations will bias survival times downward, although this bias should be small given that only a small fraction of the sample is right censored.

The second problem with excluding observations is that it might bias the model parameters. However, the differences in expenditures in Table B.1 are relatively small and are likely due to differences in observable characteristics between individuals. It seems reasonable to assume that the probability of non-response or dropout is unrelated to current spending after

controlling for demographics, lagged and initial spending, and unobserved heterogeneity. In other words, the missing data can plausibly be thought of as missing at random (MAR).

C MCMC Algorithm

The prior distributions for the model are,

$$\begin{aligned}
\alpha &\sim N(\alpha_0, V_\alpha), \\
\beta &\sim N(\beta_0, V_\beta), \\
\sigma_\epsilon^{-2} &\sim \text{Ga}(\text{a}_0, \text{b}_0), \\
\Sigma_b &\sim IW(S_0, v_0).
\end{aligned}$$

To facilitate posterior computations, it is convenient to augment the joint posterior distribution in equation 9 with the latent variable d_{it}^*. Denote n as the number of individuals in the data, $T_i - 1$ as the number of periods for which expenditure data is available for individual i, and N as the total number of observations. Furthermore, let y, d, and d^* be the $N \times 1$ stacked vectors of y_{it}, d_{it} and d_{it}^* respectively. The joint posterior distribution is then,

$$\begin{aligned}
p(d^*, \alpha, \beta, b_i, \sigma_\epsilon^{-2}, \Sigma_b | y, d) &\propto p(\alpha)p(\beta)p(\sigma_\epsilon^{-2})p(\Sigma_b) \\
&\quad \prod_{i=1}^{n}\prod_{t=1}^{T_i-1} p(d_{it}|d_{it}^*)p(d_{it}^*|\alpha)p(y_{it}|d_{it}^*, \alpha, \beta, \sigma_\epsilon^{-2}, b_i)p(b_i|\Sigma_b) \\
&= p(\alpha)p(\beta)p(\sigma_\epsilon^{-2})p(\Sigma_b) \\
&\quad \times \prod_{i=1}^{n}\prod_{t=1}^{T_i-1} [I(d_{it}=1)I(d_{it}^*>0) + I(d_{it}=0)I(d_{it}^*\leq 0] \\
&\quad \times N(d_{it}^*; x_{1it}^T\alpha + b_{1i}, 1) \\
&\quad \times \left[LN(y_{it}; x_{2it}^T\beta + b_{2i}, \sigma_\epsilon^2)I(d_{it}^*>0) + I(y_i=0)I(d_{it}^*\leq 0)\right] \\
&\quad \times N\left(\begin{bmatrix} b_{1i} \\ b_{2i} \end{bmatrix}; \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \Sigma_b\right),
\end{aligned}$$

where $I(\cdot)$ is an indicator function equal to 1 if the statement in parentheses is true and 0 if it is false. Starting with initial values, the MCMC algorithm iterates through the following steps by drawing from the following conditional distributions:

1. **Draw α:**. To draw α, make use of the augmented data and first draw d_{it}^* from its full conditional,

$$d_{it}^*|\cdot \sim \begin{cases} TN_{(-\infty,0)}(x_{1it}^T\alpha + b_{1i}, 1) & \text{if } d_{it} = 0 \\ TN_{(0,\infty)}(x_{1it}^T\alpha + b_{1i}, 1) & \text{if } d_{it} = 1, \end{cases}$$

where "$\cdot$" is a simplifying notation used to denote the remaining parameters in the model. α can then be drawn from its full conditional,

$$\alpha|\cdot \sim N(\overline{\alpha}, \overline{V}_\alpha),$$

where

$$\begin{aligned} \overline{V}_\alpha &= \left(V_\alpha^{-1} + X_1^T X_1\right)^{-1} \\ \overline{\alpha} &= \overline{V}_\alpha \left[V_\alpha^{-1}\alpha_0 + X_1^T(D^* - b_1)\right], \end{aligned}$$

and X_1 is an $N \times k_1$ design matrix and b_1 is an $N \times 1$ column vector of unobserved individual effects for the binary component (i.e. b_{1i} is repeated $T_i - 1$ times for individual i).

2. **Draw β:** Let $\ln y^+$ denote the $N^+ \times 1$ stacked vector of y_{it} for $y_{it} > 0$ where N^+ is the total number of positive observations. Similarly, let X_2^+ be the corresponding $N^+ \times k_2$ design matrix and b_2^+ be the $N^+ \times 1$ concatenated column vector of unobserved individual effects for the continuous component. Then β can be drawn from its conditional

distribution using results for a standard normal linear regression with a normal prior,

$$\beta|\cdot \sim N(\overline{\beta}, \overline{V}_\beta),$$

where,

$$\overline{V}_\beta = \left(V_\beta^{-1} + \sigma_\epsilon^{-2} X_2^{+^T} X_2^+\right)^{-1}$$
$$\overline{\beta} = \overline{V}_\beta \left[V_\beta^{-1}\beta_0 + \sigma_\epsilon^{-2} X_2^{+^T}\left(\ln y^+ - b_2^+\right)\right]$$

3. **Draw σ_ϵ^{-2}:** With the gamma prior, the conditional distribution is also a gamma distribution,

$$\sigma_\epsilon^{-2}|\cdot \sim \mathrm{Ga}\left(\mathrm{a}_0 + \mathrm{N}^+/2, \mathrm{b}_0 + \frac{1}{2}\left[\ln \mathrm{y}^+ - \mathrm{X}_2^+ - \mathrm{b}_2^+\right]^{\mathrm{T}}\left[\ln \mathrm{y}^+ - \mathrm{X}_2^+ - \mathrm{b}_2^+\right]\right)$$

4. **Draw Σ_b:** With the inverse Wishart prior, the conditional posterior is also inverse Wishart,

$$\Sigma_b \sim IW\left(S_0^{-1} + b^T b, n + v_0\right),$$

where b is an $n \times 2$ matrix with the first and second columns contain unobserved individual effects from the binary and continuous components respectively.

5. **Draw b_i:** The full conditional (marginalized over d_{it}^*) is,

$$p(b_i) \propto N_2\left(b_i; 0, \Sigma_b\right) \prod_{t=1}^{T_i - 1} f(y_{it}|y_{it-1}, \alpha, \beta, \sigma_\epsilon^2, b_i).$$

It is not possible to sample directly from this distribution so a random walk Metropolis algorithm is used to update b_i using a multivariate normal proposal density centered at the value from the previous iteration, b_i^{old}. The variance of the proposal density is

tuned to achieve desired acceptance rates.

D Simulation Procedure

Each simulation is conditional on data in period 0 (wave 4). For a given individual i and a set of parameter values from the posterior distribution, a single simulation from period 1 until period k proceeds as follows:

1. At $t = 1$:
   (a) Draw unobserved effects: $b_i \sim N_2(0, \Sigma_b)$.
   (b) Draw expenditures:
      i. $d^*_{i1} \sim N(x^T_{1i1}\alpha + b_{1i}, 1)$. If $d^*_{i1} \geq 0$, then $d_{i1} = 1$. Else $d_{i1} = 0$
      ii. If $d_{i1} = 0$, then $y_{i1} = 0$. Else $y_{i1} \sim \text{LN}(x^T_{2i1}\beta + b_{2i}, \sigma^2_\epsilon)$
2. For $2 < t < k$:
   (a) For $j = 1, 2, M$, update x_{jit} to reflect y_{it-1}. Increment *age* by 2 and update x_{jit} accordingly.
   (b) Draw death indicators:
      i. $m^*_{it} \sim N(x^T_{Mit}\kappa, 1)$. If $m^*_{it} \geq 0$, then $m_{it} = 1$. Else $m_{it} = 0$
   (c) If $m_{it} = 0$, draw expenditures:
      i. $d^*_{it} \sim N(x^T_{1it}\alpha + b_{1i}, 1)$. If $d^*_{it} \geq 0$, then $d_{it} = 1$. Else $d_{it} = 0$.
      ii. If $d_{it} = 0$, then $y_{it} = 0$. Else $y_{it} \sim \text{LN}(x^T_{2it}\beta + b_{2i}, \sigma^2_\epsilon)$.
   (d) If $t < k$, increment t by 1. Else, stop simulation.

E Additional Tables and Figures

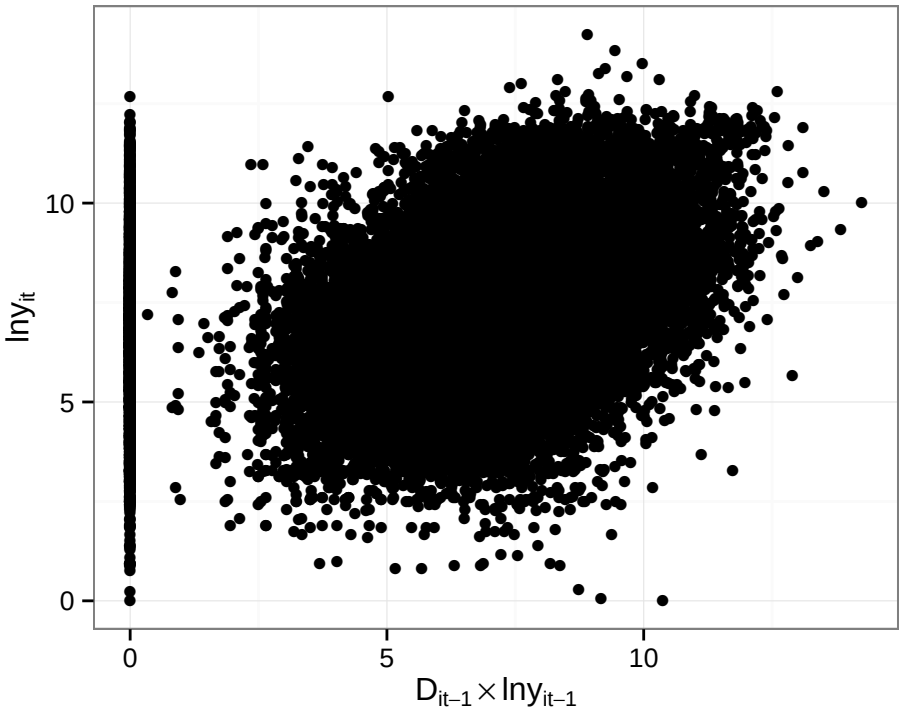

Figure E.1: Plot of $\ln y_{it}$ Against $d_{it-1} \times \ln y_{it-1}$

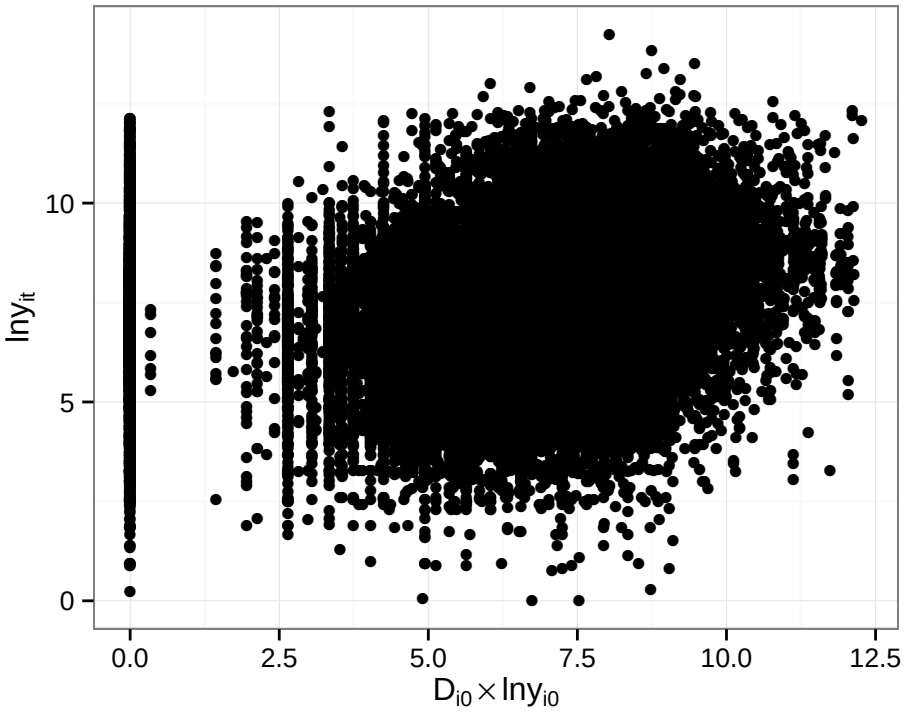

Figure E.2: Plot of $\ln y_{it}$ Against $d_{i0} \times \ln y_{i0}$

F Additional Model Checking

A few diagnostics for longitudinal two-part models have been proposed in the literature. I use some of these to further check the fit of the model. First, as suggested by Olsen and Schafer (2001), I compare the observed and predicted logarithm of non-zero expenditures for each individual averaged across waves and the actual and predicted number of waves with positive expenditures. In both cases predictions are made using the posterior means of the relevant posterior predictive distributions. The predicted versus actual plots are shown in Figure F.1. The plots don't show any major flaws with the model although the predictions tend to underestimate (overestimate) values at the upper (lower) end of the distribution. This is not unexpected since the hierarchical model partially pools the random effects.

I conducted two additional checks to ensure that the model was able to predict sample means accurately. The first test quantity is the observed proportion of nonzero observations. The second test quantity is the mean of nonzero expenditures, which is not always easy to predict because the log of expenditure needs to be transformed. This transformation can be sensitive to the lognormality assumption and the homoscedasticity assumption. Model fit is measured with a Bayesian p-value, which is just the proportion of simulations for which the test quantity is greater than or equal to observed value. P-values close to 0 or 1 indicate that the model does not capture a particular aspect of the data.

Figure F.2 compares the observed test quantities, $T(y, \theta)$, to the simulated test quantities, $T(y^{sim}, \theta)$. Figure (a) graphs the simulated distribution of the percentage of observations with nonzero spending. The Bayesian p-value is equal to the shaded error below the density curve. The p-value is relatively close to 0.5 which indicates that the model predicts the number of nonzero observations adequately. (Note that this p-value is consistent with the 95% credible interval reported for the proportion of positive values of y^{sim} reported in Section 4.4).

Figure (b) repeats figure (a) for mean expenditures. The p-value is again far from 0 or 1

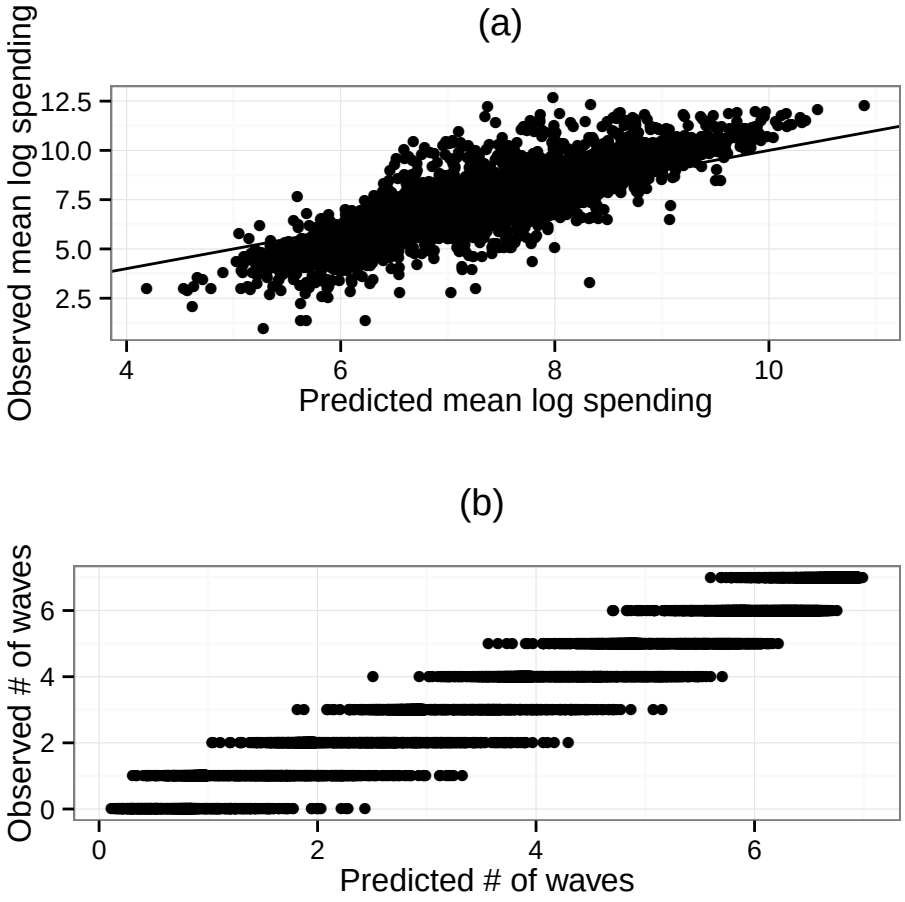

Figure F.1: Actual versus Predicted Plots

Notes: Figure (a) compares observed versus expected average log-spending for each individual across waves 5 to 11. Figure (b) plots the observed number of waves with positive expenditures against expected number of waves with positive expenditures.

which suggests that the continuous component of the expenditure model is able to simulate the sample mean fairly well.

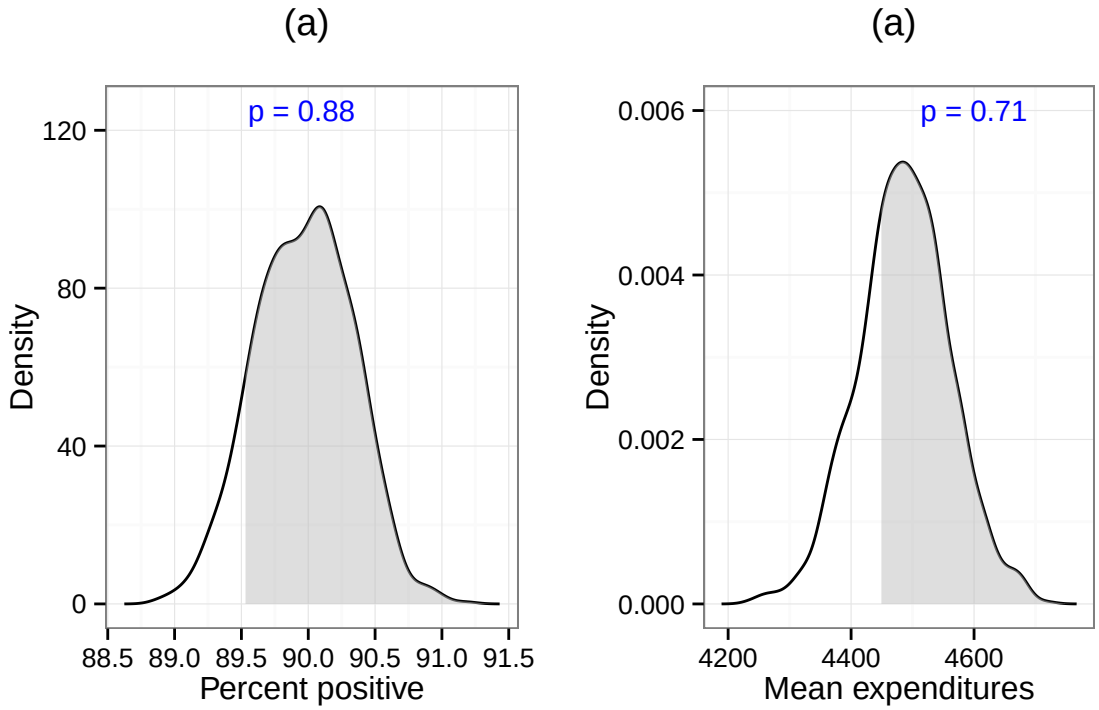

Figure F.2: Posterior Predictive Checks for Summary Measures of Fit

Notes: Figure (a) plots the simulated distribution of nonzero observations and figure (b) plots simulated mean expenditures. The p-value in each figure is the proportion of simulations in which the simulated data is greater than the observed value.

References

Albert, James H and Siddhartha Chib. 1993. "Bayesian analysis of binary and polychotomous response data." *Journal of the American statistical Association* 88(422):669–679.

Aron-Dine, Aviva, Liran Einav and Amy Finkelstein. 2013. "The RAND Health Insurance Experiment, Three Decades Later." *The Journal of Economic Perspectives* pp. 197–222.

Berk, M.L. and A.C. Monheit. 2001. "The concentration of health care expenditures, revisited." *Health Affairs* 20(2):9–18.

Bernard, Didem, Cathy Cowan, Thomas Selden, Liming Cai, Aaron Catlin and Stephen

Heffler. 2012. "Reconciling medical expenditure estimates from the MEPS and NHEA, 2007." *Medicare & medicaid research review* 2(4).

Blanco, Carlos, Sapana R Patel, Linxu Liu, Huiping Jiang, Roberto Lewis-Fernández, Andrew B Schmidt, Michael R Liebowitz and Mark Olfson. 2007. "National trends in ethnic disparities in mental health care." *Medical Care* 45(11):1012–1019.

Breyer, F., MK Bundorf and MV Pauly. 2012. "Health Care Spending Risk, Health Insurance, and Payment to Health Plans." *MV Pauly, TG McGuire, & PP Barros, Handbook of health economics* 2:691–762.

Brown, Jeffrey R and Amy Finkelstein. 2008. "The Interaction of Public and Private Insurance: Medicaid and the Long-Term Care Insurance Market." *The American Economic Review* 98(3):1083.

Cameron, A Colin and Pravin K Trivedi. 2005. *Microeconometrics: methods and applications*. Cambridge university press.

Centers for Medicare & Medicaid Services. N.d. "National Health Care Expenditure Data." https://www.cms.gov/Research-Statistics-Data-and-Systems/Statistics-Trends-and-Reports/NationalHealthExpendData/index.html. Accessed August 13, 2015.

Cook, Benjamin Lê and Willard G Manning. 2009. "Measuring racial/ethnic disparities across the distribution of health care expenditures." *Health services research* 44(5p1):1603–1621.

Cook, Benjamin, Thomas McGuire and Jeanne Miranda. 2007. "Measuring trends in mental health care disparities, 2000–2004." *Psychiatric Services* 58(12):1533–1540.

Cox, David R. 1972. "Regression models and life-tables." *Journal of the Royal Statistical Society. Series B (Methodological)* pp. 187–220.

Eichner, M.J., M.B. McClellan and D.A. Wise. 1997. Health expenditure persistence and the feasibility of medical savings accounts. In *Tax Policy and the Economy, Volume 11*. MIT Press pp. 91–128.

Engelhardt, Gary V and Jonathan Gruber. 2011. "Medicare Part D and the financial protection of the elderly." *American Economic Journal: Economic Policy* 3(4):77–102.

Feenberg, Daniel, Jonathan Skinner et al. 1994. "The Risk and Duration of Catastrophic Health Care Expenditures." *The Review of Economics and Statistics* 76(4):633–47.

Fernández-Villaverde, Jesús and Dirk Krueger. 2007. "Consumption over the life cycle: Facts from consumer expenditure survey data." *The Review of Economics and Statistics* 89(3):552–565.

Finkelstein, Amy and Robin McKnight. 2008. "What did Medicare do? The initial impact of Medicare on mortality and out of pocket medical spending." *Journal of public economics* 92(7):1644–1668.

French, Eric and John Bailey Jones. 2004. "On the distribution and dynamics of health care costs." *Journal of Applied Econometrics* 19(6):705–721.

Gelman, Andrew, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari and Donald B Rubin. 2013. *Bayesian data analysis*. CRC press.

Goldman, Dana P, Julie Zissimopoulos and Yang Lu. 2011. Medical Expenditure Measures in the Health and Retirement Study. In *Forum for health economics & policy*. Vol. 14.

Gourinchas, Pierre-Olivier and Jonathan A Parker. 2002. "Consumption over the life cycle." *Econometrica* 70(1):47–89.

Gross, Cary P, Benjamin D Smith, Elizabeth Wolf and Martin Andersen. 2008. "Racial disparities in cancer therapy." *Cancer* 112(4):900–908.

Hartman, Micah, Aaron Catlin, David Lassman, Jonathan Cylus and Stephen Heffler. 2008. "US health spending by age, selected years through 2004." *Health Affairs* 27(1):w1–w12.

Hsiao, Cheng. 2014. *Analysis of panel data*. Vol. 54 Cambridge university press.

Hurd, Michael D and Susann Rohwedder. 2009. "The level and risk of out-of-pocket health care spending." *Michigan Retirement Research Center Working Paper* 218.

Ibrahim, Joseph G, Haitao Chu and Liddy M Chen. 2010. "Basic concepts and methods for joint models of longitudinal and survival data." *Journal of Clinical Oncology* 28(16):2796–2801.

Keehan, Sean P. 2006. "Reconciling medical expenditure estimates from the MEPS and NHEA, 2002." *Health Care Financing Review* 28(1):25.

Koop, Gary, Dale J Poirier and Justin L Tobias. 2007. *Bayesian econometric methods*. Cambridge University Press.

Kopczuk, Wojciech and Joseph P Lupton. 2007. "To leave or not to leave: The distribution of bequest motives." *The Review of Economic Studies* 74(1):207–235.

Kowalski, Amanda. 2015. "Censored quantile instrumental variable estimates of the price elasticity of expenditure on medical care." *Journal of Business & Economic Statistics* (just-accepted):00–00.

Li, Tong and Xiaoyong Zheng. 2008. "Semiparametric Bayesian inference for dynamic Tobit panel data models with unobserved heterogeneity." *Journal of Applied Econometrics* 23(6):699–728.

Lynch, Scott M and J Scott Brown. 2005. "A new approach to estimating life tables with covariates and constructing interval estimates of life table quantities." *Sociological Methodology* 35(1):177–225.

Manning, W.G., J.P. Newhouse, N. Duan, E.B. Keeler and A. Leibowitz. 1987. "Health insurance and the demand for medical care: evidence from a randomized experiment." *The American Economic Review* pp. 251–277.

Marshall, Samuel, Kathleen McGarry and Jonathan S Skinner. 2011. The risk of out-of-pocket health care expenditure at the end of life. In *Explorations in the Economics of Aging*. University of Chicago Press pp. 101–128.

Martinez, Steve R, Anthony S Robbins, Frederick J Meyers, Richard J Bold, Vijay P Khatri and James E Goodnight. 2008. "Racial and ethnic differences in treatment and survival among adults with primary extremity soft-tissue sarcoma." *Cancer* 112(5):1162–1168.

McClellan, Mark and Jonathan Skinner. 2006. "The incidence of Medicare." *Journal of Public Economics* 90(1):257–276.

Monheit, A.C. 2003. "Persistence in health expenditures in the short run: prevalence and consequences." *Medical Care* 41(7):III.

Neelon, Brian, A James O'Malley and Sharon-Lise T Normand. 2011. "A Bayesian Two-Part Latent Class Model for Longitudinal Medical Expenditure Data: Assessing the Impact of Mental Health and Substance Abuse Parity." *Biometrics* 67(1):280–289.

Neelon, Brian H, A James O'Malley and Sharon-Lise T Normand. 2010. "A Bayesian model for repeated measures zero-inflated count data with application to outpatient psychiatric service use." *Statistical Modelling* 10(4):421–439.

Newhouse, J.P. 2004. *Pricing the priceless: a health care conundrum*. Vol. 2 MIT Press.

Olsen, Maren K and Joseph L Schafer. 2001. "A two-part random-effects model for semicontinuous longitudinal data." *Journal of the American Statistical Association* 96(454):730–745.

Pope, Gregory C, John Kautter, Randall P Ellis, Arlene S Ash, John Z Ayanian, Melvin J Ingber, Jesse M Levy, John Robst et al. 2004. "Risk adjustment of Medicare capitation payments using the CMS-HCC model.".

Shefrin, Hersh M and Richard H Thaler. 1988. "The behavioral life-cycle hypothesis." *Quasi Rational Economics* pp. 91–126.

Siebert, Uwe, Oguzhan Alagoz, Ahmed M Bayoumi, Beate Jahn, Douglas K Owens, David J Cohen, Karen M Kuntz, ISPOR-SMDM Modeling Good Research Practices Task Force et al. 2012. "State-transition modeling: a report of the ISPOR-SMDM modeling good research practices task force-3." *Value in Health* 15(6):812–820.

St Clair, Patricia, Delia Bugliari, Philip Pantoja, Serhii Ilchuk, Gabriela Lopez, Steven Haider, Michael Hurd, David Loughran, Constantijn Panis, Monika Reti et al. 2014. "RAND HRS Data Documentation Version N." *RAND Corporation, Santa Monica* .

Stanton, Mark W and MK Rutherford. 2006. *The high concentration of US health care expenditures*. Agency for Healthcare Research and Quality Washington, DC.

Su, Li, Brian DM Tom and Vernon T Farewell. 2009. "Bias in 2-part mixed models for longitudinal semicontinuous data." *Biostatistics* 10(2):374–389.

Tooze, Janet A, Gary K Grunwald and Richard H Jones. 2002. "Analysis of repeated measures data with clumping at zero." *Statistical methods in medical research* 11(4):341–355.

Van Vliet, René CJA. 1992. "Predictability of individual health care expenditures." *Journal of Risk and Insurance* pp. 443–461.

Waters, H.R., G.F. Anderson and J. Mays. 2004. "Measuring financial protection in health in the United States." *Health Policy* 69(3):339–349.

Webb, Anthony and Natalia Zhivan. 2010. How Much Is Enough? The Distribution of Lifetime Health Care Costs. Technical report Center for Retirement Research.

Wooldridge, Jeffrey M. 2005. "Simple solutions to the initial conditions problem in dynamic, nonlinear panel data models with unobserved heterogeneity." *Journal of applied econometrics* 20(1):39–54.

Wulfsohn, Michael S and Anastasios A Tsiatis. 1997. "A joint model for survival and longitudinal data measured with error." *Biometrics* pp. 330–339.

Zhang, Min, Robert L Strawderman, Mark E Cowen and Martin T Wells. 2006. "Bayesian inference for a two-part hierarchical model: An application to profiling providers in managed health care." *Journal of the American Statistical Association* 101(475):934–945.

Zuvekas, Samuel H and Gary L Olin. 2009. "Validating household reports of health care use in the medical expenditure panel survey." *Health Services Research* 44(5p1):1679–1700.

Zuvekas, Samuel H and Joel W Cohen. 2007. "Prescription drugs and the changing concentration of health care expenditures." *Health Affairs* 26(1):249–257.