# Adverse Selection, Moral Hazard and the Demand for Medigap Insurance

Michael Keane*
The University of Oxford

Olena Stavrunova†
University of Technology, Sydney

April 2014

**Abstract**

The size of adverse selection and moral hazard effects in health insurance markets has important policy implications. For example, if adverse selection effects are small while moral hazard effects are large, conventional remedies for inefficiencies created by adverse selection (e.g., mandatory insurance enrolment) may lead to substantial increases in health care spending. Unfortunately, there is no consensus on the magnitudes of adverse selection vs. moral hazard. This paper sheds new light on this important topic by studying the US Medigap (supplemental) health insurance market. While both adverse selection and moral hazard effects of Medigap have been studied separately, this is the first paper to estimate both in a unified econometric framework.

Our results suggest there is adverse selection into Medigap, but the effect is small. A one standard deviation increase in expenditure risk raises the probability of insurance purchase by 0.055. In contrast, our estimate of the moral hazard effect is much larger. On average, Medigap coverage increases health care expenditure by 24%.

Keywords: Health insurance, adverse selection, moral hazard, health care expenditure

JEL codes: I13, D82, C34, C35

*This research has been supported by Australian Research Council grant FF0561843 and by the ARC Centre of Excellence in Population Ageing Research (ARC grant CE110001029). But the views expressed are entirely our own.

†Corresponding author. Economics Discipline Group. University of Technology Sydney. PO Box 123, Broadway NSW 2007. Australia. Email: olena.stavrunova@uts.edu.au. T: +61295143597. F: +61295147711

# 1 Introduction

This paper studies adverse selection and moral hazard in the US Medigap insurance market. Medigap is a type of supplemental insurance sold by private insurers to cover gaps in Medicare, the primary social program providing health insurance coverage to senior citizens. While both the adverse selection and moral hazard effects of Medigap have been studied separately, this is the first paper to estimate both in a unified econometric framework.

Of course, private information is central to the analysis of insurance markets. For instance, "adverse selection" is the propensity of high-risk individuals to purchase more coverage. Rothschild and Stiglitz (1976) show that if people have private information about their risk type, the competitive equilibrium (if it exists) is not efficient: adverse selection drives up premiums, and low-risk individuals are underinsured. As a result, there may be scope for government intervention in insurance markets (e.g. mandatory insurance coverage).

But the functioning of insurance markets can also be distorted by "moral hazard," which is another type of informational asymmetry (Arrow (1963), Pauly (1968)). Moral hazard arises if ex-post risk of insured individuals is higher than the ex-ante risk. This occurs if insurance, by lowering the cost of health care, increases the rate of health care utilization (conditional on health outcomes) and/or decreases the incentive to avoid bad outcomes. In either case, insurance coverage tends to increase a person's health care utilization.[1]

Thus, both adverse selection and moral hazard manifest themselves in a positive relationship between ex-post realization of risk and insurance coverage (Chiappori and Salanie (2000)). This makes them challenging to disentangle empirically. But from a policy point of view the distinction between the two is very important. The same policies that can deal with adverse selection (e.g. mandatory enrolment) can lead to greatly increased aggregate health care costs if the moral hazard effect is strong.

[1] We focus on the first definition of moral hazard (i.e. insurance increasing health care utilization conditional on health outcomes), which is the most often used definition in the recent health economics literature.

Here, we study the Medigap insurance market, and develop a simultaneous equations model for the joint determination of (i) demand for health insurance, and (ii) health care expenditure. Joint modelling is important, because one can't estimate the moral hazard effect without quantifying the extent of selection (by risk type) into insurance coverage; nor can one measure the extent of adverse selection into insurance without quantifying how insurance coverage affects demand for services (conditional on health).

Our paper builds on the work of Fang, Keane and Silverman (2008), henceforth FKS, who studied selection into Medigap but did not estimate moral hazard. As FKS point out, a key advantage of the Medigap market for studying adverse selection is that it is relatively easy to measure private information about health expenditure risk. By law, insurers can only price Medigap policies based on age, gender, state of residence and smoking status. Thus, expenditure risk due to other factors, including health status, can be considered "private" information of individuals for the purposes of the analysis. The ability to observe private information enables us to estimate sources of selection into Medigap.[2] In addition, several of our private information variables generate plausibly exogenous variation in Medigap insurance coverage.[3] This allows us to identify the moral hazard effect.

Aside from estimating moral hazard jointly with selection, our paper contains at least five significant advances over FKS. First, we use a much more sophisticated model of health expenditure. To achieve a good fit to the health expenditure distribution we use the "smooth mixture of Tobits model," which generalizes the Smoothly Mixing Regressions framework of Geweke and Keane (2007). Second, like FKS, we merge data from two datasets in our analysis (the Medicare Current Beneficiary Survey (MCBS) and the Health and Retirement

[2]Aside from health measures, the private information variables we observe are cognitive ability, income, education, financial risk attitudes, financial planning horizon, longevity expectations, race and marital status.

[3]That is, while they enter the insurance equation, they can be excluded from the health expenditure equation, given our rich set of controls for health status. For example, we make the plausible assumption that, conditional on health status, financial risk attitudes do not enter the expenditure equation directly - they only affect health expenditure through their effect on insurance status.

study (HRS)). However, in contrast to the *ad hoc* imputation method used by FKS, we use a formal Bayesian approach. We construct a Markov Chain Monte Carlo (MCMC) algorithm, where variables missing from one dataset but present in the other are imputed within the algorithm as steps in the Markov chain. Third, we estimate not only an average moral hazard effect, but the entire distribution of effects across types of people. Fourth, in addition to the variables used in FKS, we also consider race and marital status as potential sources of adverse selection. These variables can affect both tastes for insurance and health care expenditure, but cannot be legally used to price Medigap policies. Finally, our model allows for correlation in the unobservable determinants of insurance choice and health care expenditures (which FKS assume to be uncorrelated conditional on observables).

Our main results are as follows: We find that, conditional on Medigap pricing variables only, there is *advantageous* selection into Medigap insurance. That is, contrary to classical theory, higher-risk individuals are *less* likely to buy insurance. But, conditional on a set of private information variables (including income, education, risk attitudes, cognitive ability, financial planning horizon, longevity expectations, race and marital status) there is adverse selection into Medigap insurance. This adverse selection effect is not very strong: a one standard deviation increase in expenditure risk in the Medicare only state (12.7 thousand dollars) increases the probability of buying insurance by only 5.5 percentage points (from the sample mean Medigap coverage rate of 50% up to 55.5%). We also find that, of the private information variables, cognitive ability and income are the most important factors explaining advantageous selection.[4] These findings regarding the magnitude and sources of selection are consistent with the main results of FKS. But we also find that race is an important source of *adverse* selection: blacks and Hispanics have both lower demand for Medigap insurance and lower health care expenditure.

[4]Both high income and high cognitive ability people tend to be (i) healthier and (ii) to demand more insurance conditional on health.

Our results imply that moral hazard effects of insurance are large. We find that, on average, a person with Medigap insurance spends about \$1,615 (24%) more on health care than his/her counterpart who does not have Medigap. As external validation of our model, it is notable that this is similar in magnitude to the moral hazard effect in the RAND Health Insurance Experiment. For example, Manning et al. (1987) find that decreasing the co-insurance rate from 25% to 0 increased total health care expenditure by 23%. The effect of adopting one of many typical Medigap insurance plans that cover co-pays is similar to this drop in the co-insurance rate,[5] and we predict it has a similar effect on expenditure.

Our model allows us to estimate the entire distribution of moral hazard effects. We find the moral hazard effect of Medigap varies in important ways with individual characteristics. In particular, the demand for health care is much more elastic for healthier people. As a result, given a universal extension of Medigap coverage, most of the increase in health care spending would be directed toward the healthiest seniors.

The paper is organized as follows: Section 2 contains the literature review; Section 3 describes the data; Section 4 presents our model of demand for Medigap insurance and health care expenditure; Section 5 presents the empirical results; Section 6 concludes.

# 2 Literature Review

## Part A: Literature on Health Insurance in General

Many studies examine either adverse selection or moral hazard in health insurance markets. Cutler and Zeckhauser (2000) review the literature that focuses on selection, and conclude that most studies find evidence for adverse selection. These studies often use data from em-

[5]For example, average out of pocket expenses of people with Medigap is about 1.8 thousand dollars (Kaiser Family Foundation 2005), which corresponds to about 23% of the average total health care expenditure. In contrast, basic Medicare alone leaves about 55% of costs uncovered. Thus, on average, adopting a Medigap policy decreases out-of-pocket costs by 32 percentage points (from 55% to 23%), or by 60%.

ployers who offer different insurance plans to their employees, and examine risks across plans with different generosity. There is also empirical evidence that points to the importance of moral hazard. For example, Manning et al. (1987) use data from the RAND Health Insurance Experiment and find that individuals who were randomly given more generous plans had higher health care expenditure. Chiappori et al. (1998) find that an exogenous increase in the generosity of health insurance coverage in France had a positive effect on some categories of health care expenditure. Several studies estimate substantial moral hazard effects of insurance by employing parametric multiple equation models with exclusion restrictions (e.g., Munkin and Trivedi (2008, 2010), Deb et al. (2006)).

Only a few papers have estimated selection and moral hazard effects within a single structural model of health insurance choice and demand for health care. Cardon and Hendel (2001) were the first to adopt this approach. Using data from National Medical Expenditure survey, they find evidence of little adverse selection but of substantial moral hazard. But to estimate their model they rely on the strong assumptions that the insurance choice set faced by an individual is exogenous, and that health shocks are lognormal.

In contrast, recent papers by Bajari et al. (2011a,b) develop a semiparametric method for inference in a structural model of health insurance and health expenditure choice. They find evidence of substantial moral hazard and adverse selection in the HRS and in the insurance claims data from a large self-insured employer. However, while Bajari et al. (2011a,b) are flexible with respect to the distribution of expenditure risk, their framework is restrictive in that it does not allow for heterogeneity in risk preferences, or correlation of risk preferences with expenditure risk. Such features have been found to be important for explaining data regularities in several insurance markets (e.g., Fang et al. (2008), Finkelstein and McGarry (2006)). An extensive review of empirical studies of selection and moral hazard in other insurance markets is given in Cohen and Siegelman (2010).

## Part B: Literature on Medigap in Particular

As we noted earlier, it is difficult to disentangle selection and moral hazard effects empirically. So it is not surprising that existing studies of the Medigap market disagree on their magnitudes. For example, Wolfe and Goddeeris (1991) find evidence of adverse selection and moral hazard in their 1977-1979 sample of Retirement History Survey respondents. In particular, they find that a one standard deviation health expenditure shock[6] increases the probability of supplemental insurance by roughly 12 percentage points over a two year period. They also find that the moral hazard effect of supplemental insurance is a substantial 37% increase in expenditure on hospital and physician services.

Ettner (1997) also finds both adverse selection and moral hazard using the 1991 MCBS. In particular, she finds that total Medicare reimbursements of seniors who purchased Medigap independently were about 500 dollars higher than for those who obtained Medigap coverage through an employer. Assuming the former group is less healthy, this implies adverse selection. She also reported moral hazard effects of 10% and 28% of average total Medicare reimbursements for plans with lower and higher generosity of coverage, respectively.

On the other hand, Hurd and McGarry (1997), who use the 1993-1994 Asset and Health Dynamics Survey, find that adverse selection is small, and that higher health care use by people with Medigap is mostly due to moral hazard. And, using HRS data on inpatient care, Dardanoni and Donni (2012) find that Medigap insurance increases the annual probability of hospital admissions by 4 percentage points (12% increase from the sample average of 0.33).

Recently, Fang, Keane and Silverman (2008) actually find evidence of *advantageous* selection into Medigap insurance. That is, seniors who purchase Medigap are (on average) in better health than those who have only Medicare. This finding contradicts much of the empirical work mentioned earlier, as well as the predictions of classic asymmetric informa-

[6] This is defined as the expenditure residual left after controlling for self-assessed health, disability, wealth and demographics

tion models of insurance markets (e.g., Rothschild and Stiglitz, 1976). These models predict that, given private information about risk, the riskier types should buy more insurance.

But more recent theoretical work (de Meza and Webb (2001)) shows that advantageous selection can arise if people are heterogeneous on multiple dimensions (not only risk type), and there exists private information that is positively correlated with both health and demand for insurance. In our context "private" information includes both true unobservables and information that cannot legally be used for pricing Medigap policies. FKS used the term "sources of advantageous selection" or SAS to refer to such quantities. The SAS variables proposed by FKS include "behavioral" variables (risk tolerance, cognitive ability, financial planning horizon), and demographics that are not priced on (income, education).

To test for selection, FKS first estimate an insurance demand equation that includes only pricing variables (e.g., age, gender, state) and a measure of expenditure risk based on an extensive set of health measures. This regression yields the puzzling negative coefficient on expenditure, implying advantageous selection. Of course, the demand for insurance may depend on other factors, like income or education, but, from the point of view of an insurance company the existence of adverse/advantageous selection depends only on the partial correlation between insurance purchase and expenditure risk conditional on pricing variables. By law, the price of insurance cannot be based on income or other private information.

Next, to test if their set of private or "SAS" variables can explain advantageous selection, FKS include them in the insurance equation and test if the expenditure coefficient turns from negative to positive. This does indeed occur. Thus, among individuals who are similar in terms of the SAS variables, it is indeed the less healthy who are more likely to buy Medigap, just as classical asymmetric information models predict. Cognitive ability and income are the most important of the SAS variables.[7] Interestingly, risk tolerance is not very important,

[7] FKS argue that people with higher cognitive ability may demand more insurance because (i) they better understand the rules of Medicare and the benefits of purchasing supplemental insurance and/or (ii) they are more aware of future health care expenditure risks.

as it affects demand for insurance but is not correlated with expenditure risk.

The main limitations of the FKS analysis are (i) they did not account for possibly non-random (conditional on observables) selection into insurance when estimating the prediction model for expenditure risk, and (ii) they did not attempt to estimate moral hazard. In the present paper we seek to address these issues.

# 3 Data: The HRS and MCBS Datasets

The conventional order of exposition in econometric analysis is model followed by data. But, in the present case, so many of our modelling choices depend on the features of the Medicare program and the nature of the available data that it is necessary to discuss the data first. We begin with a brief description of Medicare and Medigap, and then turn to data.

Medicare is the primary health insurance program for most seniors in the US, but on average it only covers about 45% of health care costs of beneficiaries. To cover the large gaps in Medicare, private companies offer Medigap insurance plans - private policies which cover some of the co-pays and deductibles associated with Medicare as well as various services not covered by Medicare. Importantly for our purposes, the Medigap market is heavily regulated. Only 10 standardized Medigap plans are offered (denoted A through J),[8] and insurers can only price policies based on age, gender, smoking status and state of residence. This means we can treat any other observed characteristics of individuals as "private" information.

To conduct our analysis of demand for Medigap and determinants of medical expenditure, we would ideally need a dataset that contained all of the following variables: Medigap

[8]For example, Medigap plan A only covers Medicare co-insurance costs, 365 additional hospital days during a life-time, and blood products. In contrast, Medigap plan F, the most popular plan (with a 37% market share), also covers all Medicare deductibles, physician "balance billing," skilled nursing facility co-insurance and foreign travel emergency expenses. However, Plan F does not cover costs of preventative, home recovery or hospice care not covered by Medicare (Kaiser Family Foundation 2005). During the period of our study Medicare did not cover prescription drugs, and several Medigap plans offered drug coverage. After the Medicare drug plan was introduced the design of the Medigap options was revised, and the number of options increased.

insurance status, a comprehensive measure of all health care expenditures (not just covered expenses), a rich set of health measures (to predict expenditure), Medigap pricing variables and other demographics, and a set of "private" information or SAS variables. However, as FKS point out, such a dataset does not exist. Instead, we use the following two datasets:

The Medicare Current Beneficiary Survey (MCBS) contains comprehensive information about respondents' health care costs, as well as very detailed measures of health status. It also contains Medigap insurance status and demographics. But unfortunately it contains little information on potential SAS variables.

The Health and Retirement Study (HRS), contains detailed measures of health status similar to those found in the MCBS. It also has information on Medigap insurance status and demographics. Furthermore, the HRS contains an extensive set of potential SAS variables (including cognitive ability, education, measures of risk attitudes, etc.). But unfortunately the HRS has no information on health care expenditure.[9]

Given these data limitations, the empirical strategy of FKS was to first use the MCBS to estimate the relationship between expenditure and a rich set of health measures. Then, they used the estimated relationship to impute expected health care costs for respondents in the HRS. Finally, they estimated a demand equation for Medigap using this imputed measure of health expenditure risk. In contrast to this two-step procedure, we use formal Bayesian data imputation methods to merge the two datasets, as we describe in Section 4.

Our analysis uses data from the MCBS for 2000 and 2001, and from the HRS for 2002.

[9]Recently Medicare claims data were linked to the HRS. But expenditures not covered by Medicare cannot be linked. For example, Medicare covers skilled nursing care, but not custodial nursing care, so the latter is not in the HRS-Medicare linked data. Also, we were advised by HRS staff that prescription drug expenditures are not yet available, even for years after Medicare incorporated drug coverage. Prescription drugs constitute a large fraction of the total health care expenditures of Medicare beneficiaries (e.g., 29% on average in our MCBS data), and several Medigap plans sold before January 1, 2006 provided prescription drug coverage. More generally, FKS show that Medicare-reimbursed expenditures are very different from total medical expenditures in the MCBS (see their Table 1). And, of course, the purpose of Medigap is to cover at least some of the expenses not reimbursed by Medicare. Thus, the HRS-Medicare linked data may give a distorted view of the health expenditure risk relevant for Medigap insurance.

We include only people who rely on Basic Medicare as their primary source of coverage.[10] Descriptive statistics for selected variables are presented in Table 1. For the sake of comparability we use the same MCBS sample as FKS, and the same HRS sub-sample that FKS used to obtain column (3) of Table 6 in their paper.[11] This is the sub-sample in which all individuals have non-missing information about all potential SAS variables, including risk aversion, cognitive ability, financial planning horizon, and longevity expectations.

Risk attitude is a variable of particular interest because it has been suggested as a likely source of advantageous selection by de Meza and Webb (2001) among others. Our measure of risk attitude is the risk tolerance parameter estimated by Kimball et al. (2008) for all HRS respondents using their choices over several hypothetical income gambles. Thus, it measures financial risk aversion, not aversion to health related risks.

Cognitive ability, which was found to be an important SAS variable by FKS, is measured by several variables in the HRS: the Telephone Interview for Cognitive Status score, the word recall ability score, the numeracy score and the subtraction score. To conserve on parameters we extract a common factor from these variables and use it as a scalar measure of cognitive ability. The HRS also contains data on two other "behavioral" SAS variables used in our analysis: longevity expectations and financial planning horizon.[12] The "demographic" SAS variables (income, education, race, marital status) are contained in both datasets.

Both the MCBS and HRS contain a rich set of 76 health status measures which are

[10]That is, we drop people with other sources of coverage such as Veterans administration benefits or employer provided coverage.

[11]FKS limited their analysis to the 2000 and 2002 waves of the HRS because at the time of their study these were the latest years for which a full version of the HRS data for respondents aged 65 or more was available. FKS used three HRS samples in their analysis: (i) the full sample of 9973 observations, all of which have information on health, demographics and socioeconomic variables, but can have missing data on the SAS variables; (ii) a subsample of 3467 observations that have complete information on the risk-tolerance variable; (iii) a subsample of 1695 observations with information on all potential SAS variables. In our analysis we use the third HRS subsample. This subsample only contains observations from wave 2002 because several of the cognitive ability measures were not available in the earlier wave.

[12]From the HRS we estimate the distribution of these SAS variables conditional on exogenous characteristics common in the HRS and MCBS. We then impute the missing SAS variables in the MCBS sub-sample.

detailed in the Data Appendix of FKS. These include self-reported health, smoking status, long-term health conditions (diabetes, arthritis, heart disease, etc.) and difficulties and help received for Instrumental Activities of Daily Living (IADLs). We use factor analysis to reduce these 76 variables to ten factors that best explain expenditure.[13]

Table 1 shows descriptive statistics for our HRS and MCBS samples. Individuals in our HRS sample tend to be younger, healthier (i.e., they have a lower sample average of unhealthy factor 2 and a higher sample average of healthy factor 3) and higher income than those in the MCBS sample.[14] Table 1 also shows that individuals with Medigap are on average healthier than those without Medigap in both the HRS and the MCBS (compare again the means of factors 2 and 3). This implies advantageous selection into Medigap.

Table 2 presents results of regressions of expenditure on various sets of health status measures. Strikingly, demographics explain only 1.7% of the variance of expenditure. But inclusion of the 76 health measures increases this to 21%. Note also that inclusion of the health measures increases the Medigap coefficient from \$979 to \$1,951.[15] This implies that people with Medigap tend to be healthier, so failure to control for health understates the moral hazard effect.[16]

---

[13]We first factor-analyze these 76 variables to extract 38 factors (using data in both the HRS (full sample) and MCBS samples). We then regress health care expenditure in the MCBS on these 38 factors (along with demographics). We select 16 factors that are significant predictors of expenditure. Finally, we select the 10 factors from among these 16 that give the highest possible R-squared. The 10 factors that are selected are # 2, 3, 7, 8, 10, 11, 17, 20, 22 and 23 (not factors 1-10). Thus, the factors that explain the most covariance of the health indicators are not the same as the ones that explain most of the variance in expenditure.

[14]The fact that our HRS and MCBS samples have different characteristics does not create a problem for our analysis, provided the distribution of the SAS variables conditional on the exogenous characteristics used for imputation (including age and health) is the same in both.

[15]In Table 2 of FKS the Medigap coefficient changes from negative to positive as health controls are added to the expenditure equation. The reason for the discrepancy is that FKS use different subsamples for regressions with and without health controls. In particular, the regression without health controls uses 15,945 observations, while the regression with health controls uses 14,129 observations for which health status information is available. Table 2 in our paper uses the FKS sample of 14,129 observations to obtain the results both with and without health controls.

[16]Some caution is in order in interpreting this result however, as Medigap status may not be exogenous: It may be correlated with unobserved determinants of health expenditure that we do not control for. On the other hand, given our extensive controls for health status, this may not be a problem. Our structural model results will shed light on this issue.

Table 1: Descriptive Statistics

| | MCBS | | | HRS | | |
|---|---|---|---|---|---|---|
| Variable | All | Medigap | No Medigap | All | Medigap | No Medigap |
| Medigap | 0.50 | 1.00 | 0 | 0.43 | 1.00 | 0 |
| Female | 0.59 | 0.60 | 0.58 | 0.56 | 0.58 | 0.55 |
| Age | 76.57 | 77.02 | 76.11 | 68.70 | 68.67 | 68.72 |
| | (7.50) | (7.29) | (7.69) | (3.10) | (2.98) | (3.20) |
| Black | 0.10 | 0.04 | 0.17 | 0.14 | 0.06 | 0.20 |
| Hispanic | 0.08 | 0.03 | 0.12 | 0.07 | 0.02 | 0.11 |
| Married | 0.48 | 0.54 | 0.43 | 0.66 | 0.71 | 0.63 |
| Education: Less than high school | 0.36 | 0.27 | 0.45 | 0.28 | 0.22 | 0.33 |
| Education: High School | 0.27 | 0.31 | 0.24 | 0.38 | 0.41 | 0.35 |
| Education: Some college | 0.21 | 0.24 | 0.18 | 0.18 | 0.18 | 0.17 |
| Education: College | 0.08 | 0.10 | 0.06 | 0.08 | 0.08 | 0.08 |
| Household Income ($ in 1000s) | 25.7 | 31.2 | 20.0 | 43.8 | 54.4 | 35.7 |
| | (46.5) | (51.4) | (40.1) | (68.7) | (79.6) | (57.7) |
| Health factor 2 (Unhealthy) | 0.04 | -0.06 | 0.13 | -0.32 | -0.37 | -0.28 |
| | (1.01) | (0.89) | (1.10) | (0.51) | (0.43) | (0.56) |
| Health factor 3 (Healthy) | -0.12 | -0.09 | -0.15 | 0.17 | 0.23 | 0.13 |
| | (-0.93) | (0.97) | (0.86) | (0.72) | (0.70) | (0.74) |
| Cognition | | | | 0.46 | 0.54 | 0.40 |
| | | | | (0.31) | (0.25) | (0.33) |
| Risk tolerance | | | | 0.234 | 0.228 | 0.236 |
| (estimate from Kimball et al. (2008)) | | | | (0.142) | (0.138) | (0.146) |
| Financial planning horizon, years | | | | 4.46 | 4.83 | 4.18 |
| (finpln) | | | | (4.05) | (4.12) | (3.98) |
| Subjective probability to live to 75 or more | | | | 67.32 | 69.57 | 65.59 |
| (praliv75) | | | | (28.33) | (25.91) | (29.96) |
| Total medical expenditure ($) | 8,085 | 8,559 | 7,605 | | | |
| | (14,599) | (14,301) | (14,881) | | | |
| Number of observations | 14128 | 7113 | 7015 | 1671 | 726 | 945 |

Note: "Total medical expenditure" includes all expenditure, both covered and out-of-pocket. Standard deviations are in parentheses.

Table 2: OLS results of total medical expenditure on Medigap coverage, demographic and health status characteristics in the MCBS

| Variable | A. Without Health Controls | B. With Direct Health Controls | C. With Health Factor Controls | D. With Health Controls and Income |
|---|---|---|---|---|
| Medigap | 979.4*** | 1951.2*** | 1948.2*** | 1889.2*** |
| | (291.0) | (255.6) | (257.8) | (259.6)*** |
| Female | -933.6*** | -834.7*** | -734.3*** | -707.3** |
| | (304.9) | (290.7) | (282.3) | (282.7) |
| Age-65 | 501.5*** | 408.0*** | 437.3*** | 438.6*** |
| | (125.8) | (115.1) | (116.5) | (116.5) |
| $(\text{Age-65})^2$ | -23.3** | -28.8*** | -31.0*** | -31.0 |
| | (9.8) | (9.1) | (9.2) | ( 9.2) |
| $(\text{Age-65})^3$ | 0.43** | 0.50** | 0.51*** | 0.5*** |
| | (0.21) | (0.20) | (0.20) | (0.20) |
| Black | 1212.9* | 579.8 | 770.4 | 808.5 |
| | (639.3) | (550.3) | (596.2) | (596.4) |
| Hispanic | -576.7 | -843.8* | -622.2 | -568.3 |
| | (511.7) | (431.6) | (467.4) | (467.4) |
| Married | -779.9*** | -325.2 | -213.5 | -305.7 |
| | (299.0) | (268.7) | (275.3) | (276.3) |
| Health factor 2 | | | 4565.0*** | 4581.5*** |
| | | | (252.4) | (252.3) |
| Health factor 3 | | | -2544.6*** | -2568.4*** |
| | | | (226.4) | (226.7) |
| Health factor 7 | | | 2049.0*** | 2041.7*** |
| | | | (241.5) | (241.4) |
| Health factor 8 | | | 711.7*** | 718.6 |
| | | | (213.1) | (212.9) |
| Health factor 10 | | | -2047.0*** | -2044.7*** |
| | | | (535.5) | (535.8) |
| Health factor 11 | | | -961.6*** | -964.2 |
| | | | (207.8) | (207.8) |
| Health factor 17 | | | 1176.3 | 1180.8 |
| | | | (931.4) | (930.9) |
| Health factor 20 | | | -1339.2*** | -1350.4 |
| | | | (363.7) | (364.6) |
| Health factor 22 | | | 2144.6*** | 2136.6*** |
| | | | (382.4) | (382.8) |
| Health factor 23 | | | 1254.7*** | 1249.7*** |
| | | | (414.1) | (414.2) |
| Household Income | | | | 6.96*** |
| ($ in 1000s) | | | | (2.25) |
| Health status dummy | No | Yes | No | No |
| Region dummy | Yes | Yes | Yes | Yes |
| Year dummy | Yes | Yes | Yes | Yes |
| Observations | 14128 | 14128 | 14128 | 14128 |
| Adjusted $R^2$ | 0.017 | 0.21 | 0.18 | 0.19 |

Note: "Total medical expenditure" is measured in dollars and includes all expenditure, both covered and out-of-pocket. The regressions are weighted by cross-section sample weights. Robust standard errors clustered at the individual level are in parentheses. Statistical significance is indicated by * (10 percent), ** (5 percent) and *** (1 percent).

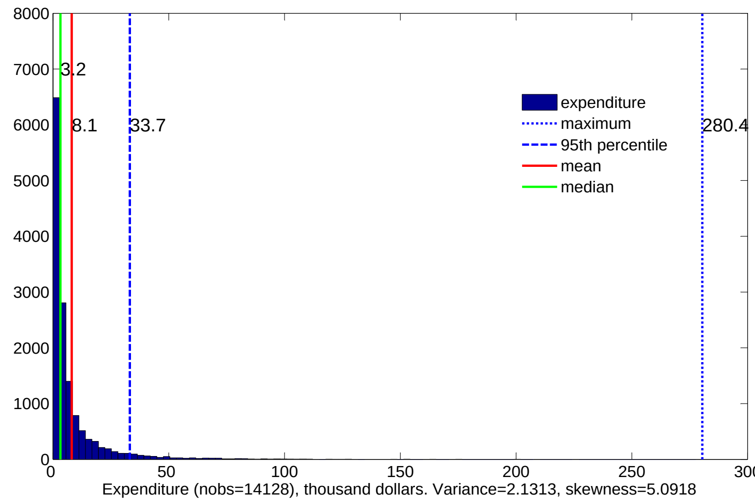
In column C we see that, when the 76 health status measures are replaced by our ten health factors, the adjusted R-squared drops from 0.21 to 0.18. This seems a reasonable price to reduce the number of covariates by 66. Health factors 2 and 3 are the most quantitatively important for predicting expenditure. Factor 2 loads heavily on deterioration in health as well as difficulties and help with IADLs, and so is an unhealthy factor. It increases expenditures by about $4,500 per standard deviation. Factor 3 loads positively on good and improving self-reported health and negatively on difficulties with IADLs and thus is a healthy factor. It decreases expenditure by $2,500 per standard deviation.

The last column of Table 2 adds household income to the previous specification. The inclusion of income has little effect on the health status and demographic variables, and it only slightly reduces the Medigap coefficient (by $59). The estimated effect of income on expenditure is quite small: a one standard deviation increase in income ($46,500) increases average expenditure by only about $324, which is 4% of the mean. Thus income appears to have little impact on health expenditure (conditional on health status). This is consistent with a view that health spending is largely driven by institutional factors, like clinical practice and Medicare/Medigap reimbursement rules, rather than by consumer choice.

# 4 A Model of Medigap Status and Health Expenditure

This section presents a model for the joint determination of insurance status and health care expenditure, in which we account for endogeneity of insurance choice by allowing the unobservable determinants of insurance status and expenditure to be correlated. But before developing the full model we first need to select a specification for the distribution of medical expenditure. It is well-known that econometric modelling of health care expenditures is challenging because of the properties of their empirical distribution. In particular, health care expenditures are non-negative, highly skewed to the right and have a point mass at

Figure 1: Histogram of total health care expenditure




zero. The histogram in Figure 1 shows that the empirical distribution of total health care expenditure of Medicare beneficiaries in our MCBS sample exhibits all these characteristics. The sample skewness is about 5.1 and the distribution has a long right tail. The proportion of observations with zero expenditure is about 0.025.

The literature on modelling health care expenditure has mainly focused on modelling it's conditional *expectation*, given the problems of extreme skewness and mass at zero (e.g., Manning (1998); Mullahy (1998); Blough et al. (1999); Manning and Mullahy (2001); Buntin and Zaslavsky (2004); Gilleskie and Mroz (2004); Manning et al. (2005)). The problem of modelling the entire *distribution* of expenditure has been less frequently addressed. The usual approach is to adopt a two-part model where positive outcomes are log-normal (e.g. Deb et al. (2006)). But we adopt a new approach that is much more flexible and that provides a much better fit to the expenditure distribution:

After trying several models of the distribution of expenditure, we decided to adopt a discrete mixture of Tobits where the probability of a mixture component depends on a person's observed characteristics. This is a generalization of the Smoothly Mixing Regressions (SMR) framework of Geweke and Keane (2007) to the case of a Tobit-type limited dependent variable, so we call it the Smooth Mixture of Tobits or "SMT." We find that the SMT can capture both the extreme skewness and non-negativity of the health care expenditure distribution. It provides a very good fit to many aspects of the MCBS expenditure data, including the conditional (on covariates) mean, variance, quantiles and probability of an extreme outcome. In section 5 we will discuss how the number of mixture components for SMT was selected, and examine the fit of the model to the distribution of expenditure.

Next, in section 4.1 we present our model of insurance status and expenditure, abstracting from the fact that not all variables of interest are available in both our datasets. Sub-section 4.1.A discusses our identification assumptions. Then, in section 4.2 we discuss our approach to dealing with missing data. Section 4.3 presents the posterior simulation algorithm.

## 4.1 Complete data

We assume there are $m$ types of individuals, indexed by $j$, $j = 1, \ldots, m$. A person's type is private information, i.e., people know their type, but from the point of view of the researcher types are latent: given a person's observed characteristics only her probability of belonging to type $j$ can be inferred. Our specification of heterogeneity is very general: types differ in the mapping of health factors and demographics to mean expenditure, the effect of insurance status on expenditure, and the variance of expenditure.

Let $I_i^*$ denote the utility that individual $i$ derives from Medigap insurance and let $E_i^*$ denote her total expected health care expenditure without Medigap. We assume that $E_i^*$ is the expenditure risk relevant when individual $i$ decides whether to purchase Medigap insurance, so henceforth we will refer to $E_i^*$ as "expenditure risk". While $E_i^*$ is the person's expected medical cost if uninsured, it also determines the expected cost of person $i$ to an insurance company.[17] Thus, for a given level of premiums, Medicare supplemental plans would like to attract the healthiest possible client base; i.e., "good risks" with low expected cost. Both $I_i^*$ and $E_i^*$ are known to the individual but are unobserved by the econometrician, so they enter the model as latent variables.

Let $I_i$ be an indicator equal to one if individual $i$ has Medigap insurance, and zero otherwise. Assume that $I_i = 0$ if $I_i^* < 0$ and $I_i = 1$ if $I_i^* >= 0$. Let $Y_i$ denote household income (in deviation form from the sample mean). Also, let $\widehat{E}_i$ denote *notional* health care expenditure of individual $i$ (as in "notional demand", which can be negative). We assume that healthcare expenditure $\widehat{E}_i$ is a function of latent health expenditure risk $E_i^*$, income $Y_i$, insurance status $I_i$ and a surprise health care cost shock $\eta_i$. Specifically, we have:

$$\widehat{E}_i|j = E_i^*|j + \gamma_{1j} Y_i + \gamma_{2j} I_i + \eta_i|j \tag{1}$$

where $j$ denotes type, $\gamma_{1j}$ denotes the type-specific effect of income on notional health care expenditure and $\gamma_{2j}$ denotes the type-specific effect of health insurance (i.e. moral hazard). We let $\boldsymbol{\gamma}_j$ denote the vector of parameters $[\gamma_{1j}, \gamma_{2j}]'$, and let $\mathbf{yi}_i$ denote the vector of covariates $[Y_i, I_i]'$. Finally, we assume that, given the individual's type $j$, the health care cost shock $\eta_i|j$ is normally distributed with zero mean and variance $\sigma_j^2$:

$$\eta_i|j \sim N(0, \sigma_j^2)$$

Note that $\eta_i|j$ can also be interpreted as the health expenditure forecast error of person $i$. The term $\sigma_j^2$ denotes the variance of notional expenditure around expected expenditure for a person of risk level $E_i^*$, conditional on their income and insurance status. Thus $\sigma_j^2$ can be

[17] Note that, due to deductibles and uncovered services, an insurer will not typically cover all costs. As we noted in footnote 5, the typical Medigap plan covers about 60% of costs not covered by basic Medicare.

interpreted as the variance of the health care expenditure forecast error.

Equation (1) captures the notion that total health care expenditure may not be driven entirely by health status $E_i^*$, but also by insurance and income. People with insurance and/or higher income may demand more health care, conditional on health status, for two reasons: First, those who are insured face a lower price of care (because they don't bear the full price of treatment). As a result they may utilize more care (”moral hazard”). Second, those with higher income may demand better quality care (e.g., private rooms) or elective treatment (e.g., cosmetic surgery) that may not be covered by insurance.

Thus, $E_i^*$ captures the intrinsic health status of a person, once we have purged their observed expenditure of any affects of insurance or income. From the point of view of an insurance company, $E_i^*$ captures the expected cost or ”riskiness” of a client.[18]

It is tempting to interpret equation (1) as a conventional demand function. Indeed, it includes the quantities we would expect in a demand function for health expenditure: (i) a measure of health status ($E_i^*$), (ii) a price shifter (i.e., insurance) and (iii) income. But (1) is not a standard demand function, because it does not depend only on consumers' preferences and budget constraints. It also depends on the incentives and constraints created by the U.S. health care delivery system. This includes physician treatment protocols, the Medicare and Medigap reimbursement rules, legal constraints on service provision, etc.

For instance, the moral hazard or (inverse) price effect of insurance on demand for services, $\gamma_{2j}$, is not necessarily positive. Indeed, a key rationale for the existence of private Medigap plans is the idea that private insurers have an incentive to manage care so as to keep costs down. For example, a Medigap plan may try to reduce costs by encouraging pre-

[18] In other words, we seek a risk measure $E_i^*$ that excludes costs of elective procedures that insurance will not cover, or any extra spending induced by insurance coverage itself. It would be a problem for our approach if income mattered for *covered* health care expenditure, so that $E_i^*$ depended on income. For example, say a Medigap plan had co-pays. Then, if higher income people are less price sensitive, they would demand more services, making $E_i^*$ an increasing function of income. However, co-pays are not a problem here, because the typical Medigap plan has a small deductible beyond which everything is reimbursed. For instance, as we noted earlier, even the most basic Medigap plan A covers basic Medicare co-insurance costs.

ventive care, thus reducing the incidence of costly but avoidable conditions and/or preventing unnecessary emergency room visits (O'Grady et al. (1985)). These arguments highlight the fact that even the sign of $\gamma_{2j}$ is of policy interest, not only its magnitude.

Similarly, even if health care is a normal good, higher income will not necessarily lead to higher health care spending, *ceteris paribus*. Treatment protocols and Medicare/Medigap reimbursement rules play a key role in determining the treatment of patients in the US health care system. As a result, it is plausible that treatment is fairly standardized (at least within geographic areas), conditional on health and insurance status.[19] Then, any impact of income on expenditure (conditional on health) may arise only because higher income individuals tend to purchase more comprehensive insurance. Indeed, this is consistent with the very small effect of income on expenditure that we found in Table 2.[20]

Returning to the exposition of the model, *realized* expenditure $E_i|j$ is given by:

$$E_i|j = max\{0, \widehat{E}_i|j\}. \tag{2}$$

where $\widehat{E}_i$ is *notional* expenditure from (1). Hence, conditional on type $j$, the model for realized expenditure $E_i$ is a Tobit. This ensures that predicted expenditure is always positive.[21]

[19] There is a large literature finding that treatment protocols vary substantially by region, in ways not explained by differences in health status, income or demographics; see Zuckerman et al. (2010), Fisher et al. (2003), Welch et al. (1993). This is one reason we include region dummies in the expenditure equation. National treatment protocols (i.e., best practice protocols) can be found at www.guideline.gov.

[20] It is worth noting how our equation (1) differs from FKS. They specify the health expenditure equation as $E_i = \mathbf{H}_i\boldsymbol{\beta} + \gamma I_i + \epsilon_{1i}$ and use this to predict expenditure risk for HRS respondents as $\widehat{E}_i = \mathbf{H}_i\widehat{\boldsymbol{\beta}}$, where $\mathbf{H}_i$ is a vector of health measures and demographic characteristics. Our $E^*$ is a counterpart of FKS's $\widehat{E}_i$. There are two key differences between (1) and the prediction equation of FKS. The first is that they use a simple linear model while we adopt the better fitting SMT specification. The second is that FKS did not include income. That is, FKS assume expenditure is uncorrelated with income, conditional on the health status H. We decided to include income in (1) as we were concerned that omitting it might lead us to exaggerate moral hazard. This would occur if income has a positive effect on both demand for health services and demand for insurance. But as we shall see, the effect of income in (1) is quite small, so this is not a major issue.

[21] In our data 2.5% of observations have zero expenditure, so notional expenditure $\widehat{E}$ is equal to the realized expenditure $E$ in most cases.

The model for the latent vector $[I_i^*, E_i^*]'$, conditional on type $j$, is specified as follows:

$$I_i^*|j = \alpha_0 E_i^*|j + \alpha_1 \sigma_j^2 + \alpha_2 \sigma_j^2 \cdot c_{1i} + \boldsymbol{\alpha}_3' \mathbf{xi}_i + \boldsymbol{\alpha}_4' \mathbf{c}_i + \varepsilon_{1i} \tag{3}$$

$$E_i^*|j = \boldsymbol{\beta}_j' \mathbf{xe}_i + \varepsilon_{2i}, \tag{4}$$

In the insurance demand equation (3) the vector $\mathbf{xi}_i$ includes insurance pricing variables (age, gender, location of residence), as well as SAS variables present in both datasets (income, education, ethnicity and marital status). The vector $\mathbf{c}_i$ includes SAS variables present in the HRS only (risk tolerance, financial planning horizon, cognition and longevity expectation). The first element of $\mathbf{c}_i$, denoted $c_{1i}$, is risk aversion. It enters both linearly and interacted with the variance of health expenditure shocks $\sigma_j^2$.

In the expenditure risk equation (4) the vector $\mathbf{xe}_i$ includes the ten health factors discussed in section 3, along with certain demographic characteristics that we assume may also affect health (i.e., age, gender, location of residence, marital status, race and ethnicity). The variables in $\mathbf{xe}_i$ are present in both datasets.

Thus, demand for insurance depends on: (i) expected costs in the uninsured state, $E_i^*$, (ii) the price of insurance, which is governed by the vector of legal pricing variables contained in $\mathbf{xi}_i$, (iii) income and other "demographic" SAS variables also included in $\mathbf{xi}_i$, (iv) the "behavioral" SAS variables in $\mathbf{c}_i$, such as risk aversion, cognitive ability, etc.,[22] and (v) the variance of the health expenditure forecast error $\sigma_j^2$ and its interaction with risk tolerance.

Note that the expenditure risk $E_i^*$ consists of a part that depends on observed health status and demographics ($\boldsymbol{\beta}_j' \mathbf{xe}_i$) and a part that depends on unobserved characteristics ($\varepsilon_{2i}$). The coefficients ($\boldsymbol{\beta}_j$) capture how each health status and demographic measure influences

[22] Recall that, in the terminology of FKS, SAS variables are variables that (i) affect demand for insurance but that can't be used by insurance companies to price policies, and (ii) that are correlated with health status (and hence with health expenditure risk). We use the terminology "potential" SAS variables to indicate variables that may plausibly be expected to satisfy these conditions. Of course, whether a variable is an actual SAS variable will not be known until we see the empirical results. But for simplicity we ignore this distinction in expositing the model.

expected costs. In our SMT framework there is heterogeneity in these effects, as the $\boldsymbol{\beta}_j$ differ across different types of individuals. Thus, the SMT model generates different marginal effects of covariates on expenditure for individuals of different types.

The degree of selection on observed health is captured by the sensitivity of insurance demand to expenditure risk $E_i^*$, conditional on other variables. That is, it is captured by $\alpha_0$ in the Medigap demand equation (3).[23] A negative $\alpha_0$ indicates advantageous selection (on observed health), while a positive value indicates adverse selection.

The disturbances $\varepsilon_{1i}$ and $\varepsilon_{2i}$ capture unobserved heterogeneity in tastes for insurance and in health status, respectively, that are known to an individual, but not to the econometrician. We assume the vector of unobserved heterogeneity in tastes and health $\boldsymbol{\varepsilon}_{12i} = [\varepsilon_{1i}, \varepsilon_{2i}]'$ is independent of transitory health shocks $\eta_i$ and follows a bivariate normal distribution:

$$\boldsymbol{\varepsilon}_{12i}|j \sim BVN\left(\mathbf{0}, \begin{bmatrix} \sigma_{11} & \sigma_{12} \\ \sigma_{12} & \sigma_{22} \end{bmatrix}\right) \qquad \text{for all types } j = 1, \ldots, m.$$

Unlike FKS, we allow for $\varepsilon_{1i}$ and $\varepsilon_{2i}$ to be correlated with covariance given by $\sigma_{12}$. A negative $\sigma_{12}$ indicates advantageous selection on unobserved health, while a positive value indicates adverse selection.[24] If $\varepsilon_{1i}$ and $\varepsilon_{2i}$ are correlated and the model does not take that into account, then $\widehat{\gamma}_{2j}$ is a biased estimate of the moral hazard effect, and $\alpha_0$ is a biased estimate of the selection effect (on observables).[25]

The probability a person is latent type $j$ depends on his/her exogenous characteristics

[23]Equation (3) is our counterpart to the FKS's Medigap equation $I_i = \alpha_0 \widehat{E}_i + \mathbf{P}_i \boldsymbol{\alpha}_2 + \mathbf{SAS}_i \boldsymbol{\alpha}_3 + \epsilon_{2i}$, where $\mathbf{P}_i$ is a vector of variables that affect the price of Medigap insurance, and $\alpha_0$ measures selection. FKS show that $\alpha_0$ turns from negative (advantageous selection) to positive (adverse selection) as more SAS variables are added to the Medigap equation. In our equation (3) the pricing variables are included in $\mathbf{xi}_i$, while all other variables are part of the SAS vector.

[24]Note that in Medigap there may exist both selection on unobservables and selection on observables, because there are observables that insurance companies cannot legally price on (health status, race, etc.).

[25]FKS controlled for selection on observables but not on unobservables.

by way of a multinomial probit model, as in Geweke and Keane (2007):

$$
\begin{aligned}
\widetilde{W}_{ij} &= \boldsymbol{\delta}_j' \mathbf{xw}_i + \zeta_{ij} \qquad j = 1, \ldots, m-1 \\
\widetilde{W}_{im} &= \zeta_{im}.
\end{aligned} \tag{5}
$$

The $\widetilde{W}_{ij}$ are latent indices, and $\mathbf{xw}_i$ is a vector of individual characteristics including demographics and health status.[26] The $\zeta_{ij}$ are independent standard normal random variables. An individual $i$ is type $j$ iff $\widetilde{W}_{ij} \geq \widetilde{W}_{il}\ \forall\ l = 1, ..., m$. Type probabilities are given by:

$$
P(\text{type}_i = j | \mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m) = \int_{-\infty}^{\infty} \phi(y - \boldsymbol{\delta}_j' \mathbf{xw}_i) \prod_{l \neq j}^{m} \Phi(y - \boldsymbol{\delta}_l' \mathbf{xw}_i) dy, \tag{6}
$$

where $\Phi(.)$ and $\phi(.)$ denote the standard normal cdf and pdf, respectively. Finally, we adopt the location normalization $\boldsymbol{\delta}_m = \mathbf{0}$, fixing $m$ as the "base" alternative. This resolves the well-known identification issue in multinomial choice models that arises because only utility differences determine choices.[27]

### 4.1.A Identification

The model in equations (1)-(4) is a simultaneous equations model where the parameters of interest (i.e., the selection and moral hazard effects) are identified via cross-equation exclusion restrictions.[28] In order to identify the selection effect $\alpha_0$, we use the exclusion restriction

[26] In our empirical specification $\mathbf{xw}_i$ is almost identical to $\mathbf{xe}_i$, with the exception that the second and third powers of age and interactions of age with gender and of marital status with gender are included in $\mathbf{xe}_i$ but not in $\mathbf{xw}_i$, in order to reduce the number of parameters to be estimated. See Table A-2.

[27] Without restrictions on $\boldsymbol{\delta}_j$, the probability of being type $j$ would not change if all $\boldsymbol{\delta}_j$ were replaced by $\boldsymbol{\delta}_j + \Delta$, where $\Delta$ is a constant.

[28] Our approach to modelling health expenditure and Medigap insurance status is related to that of Munkin and Trivedi (2010), henceforth MT, who also estimate a simultaneous equations model. They study the market for supplemental drug insurance. The most obvious difference between our papers is that we study a different market (i.e., Medigap insurance). But from a modelling perspective the main difference is in how we model selection: MT only estimate selection on unobservables. But, in insurance markets, adverse/advantageous selection also involves selection on *observables* that cannot (legally) be used for pricing insurance. Thus, in contrast to MT, we estimate selection on both unobservables and observables. We find

that the individual health status variables affect demand for insurance only through their effect on the overall (scalar) expenditure risk $E_i^*$, not directly. If the health status variables were included *individually* in the insurance equation (3), we would not be able to isolate the effect of the expenditure risk $\alpha_0$ from the independent effects of the health status variables. This assumption appears plausible, as it is not clear why insurance demand would depend on health status measures *per se*, once one has conditioned on total expenditure risk.

The moral hazard effect is identified by the exclusion restriction that selected behavioral and demographic variables (risk aversion, cognitive ability, planning horizon, longevity expectations, education) enter the insurance demand equation, but do not affect expenditure risk directly (conditional on detailed health measures). Thus, these SAS variables induce exogenous variation in insurance choices conditional on expenditure risk $E_i^*$ and income. This identifies the moral hazard effect $\gamma_{2j}$ and the correlation between $\varepsilon_{1i}$ and $\varepsilon_{2i}$.

In our view it is plausible that these SAS variables can be excluded from equation (4), given the extensive set of health status controls we include in $\mathbf{xe}_i$. But this important assumption warrants further discussion. There is limited empirical evidence about the relationship between health care expenditure and the behavioral SAS variables, but what evidence there is does seem consistent with our assumptions: For instance, FKS found no significant relationship between *financial* risk aversion and health expenditure (conditional on health status). As we noted earlier, this is unsurprising given that financial risk aversion is a fundamentally different concept from health-related risk aversion. We would make a similar argument to exclude the *financial* planning horizon from (4). Our argument for ex-

that selection on "observable private information" is much more important. Our paper also differs from MT in other important ways: (i) we use a richer set of instruments for insurance status (not just price shifters but also the SAS variables); (ii) we use a much richer set of controls for health status in the expenditure equation (this is made feasible by our factor analysis procedure), (iii) MT only use the MCBS, while we merge the MCBS with the HRS in order to study effects of SAS variables, thus extending the application of MCMC methods to a rather novel selection/data fusion exercise; and (iv) as MT note (see their conclusion), the expenditure distribution that they assume could be improved upon, and we do this by using the SMT specification, which provides a very substantial improvement in fit (see Keane and Stavrunova (2011)).

cluding longevity expectations from (4) is that differences in life expectancy, holding actual health fixed, will shift a person's planning horizon, and hence their demand for insurance, but not affect the person's current health care spending needs directly.[29]

As for cognitive ability, a recent paper by Fang et al. (2010) shows that the cross-sectional correlation between cognitive ability and medical expenditure largely vanishes when one controls for health status in the HRS.[30] Finally, turning to education, our own analysis of the MCBS subsample suggests that education has no significant explanatory power for expenditure, conditional on other demographic and health controls.[31]

A general point is that we control for both objective and *subjective* health measures in (4). In fact, subjective health is a very good predictor of expenditure. Because we control for subjective health, it seems much less likely that our SAS variables should enter (4) just because they are correlated with private information about health.

In summary we believe our set of exclusion (or exogeneity) restrictions is plausible. Of course, as Koopmans et al (1950) note, "...the distinction between exogenous and endogenous variables is a theoretical, *a prior* distinction..." Thus, we cannot prove our exogeneity

[29]We argue that differences in life expectancy, conditional on actual health, measure differences in subjective beliefs or attitudes. This is why we call life expectancy a "behavioral" variable. Clearly, life expectancy may affect demand for insurance: The longer one expects to live, the more valuable is a Medigap plan to insure against possible future costs of nursing home care (and other late-in-life expenses) that Medicare generally fails to cover. But we argue that subjective life expectancy should not affect current health care costs, once we condition on actual health - justifying its exclusion from (4). The effect of life expectancy is similar to that of financial planning horizon: a person with a longer financial horizon will also be more likely to insure against future (late-in-life) financial risks. But the two concepts are different, e.g., one could expect to live a long life yet also be the type of person who does not like to plan ahead.

[30]It is tempting to argue people with higher cognitive ability, who know more about medical conditions, should be more likely to seek treatment. But this is not at all clear. For example, if one understands that most viruses are not treatable, but that they are self-regulating, one ought to be *less* likely to make an unnecessary doctor visit for virus-like symptoms.

[31]Also, as our model is cross-sectional, we implicitly assume the health status measures ($\mathbf{H}$) are not affected by insurance status *over time*. That is, we assume away any "dynamic moral hazard" effect, whereby insurance coverage leads to more risky behavior, which, in turn, causes health status to deteriorate over time. If such dynamics do exist, we will underestimate the moral hazard effect (at least in the long run). However, Khwaja (2001) shows that in a dynamic model health insurance has two opposite effects. There is the moral hazard effect, but there is also the "Mickey Mantle" effect: because insurance increases life expectancy, an individual has a greater incentive to invest in health. Khwaja finds that the two effects roughly cancel, so insurance has little effect on how health status evolves over time.

assumptions are correct: the extent to which they are credible is up to individual readers to decide.

## 4.2 Combining data from the MCBS and the HRS

To estimate the model in section 4.1, we would ideally like a dataset with information on $I_i, E_i$ and $\mathbf{c}_i$, along with health status, demographics and income. We denote the latter variables by $\mathbf{xz}_i : \mathbf{xz}_i \supseteq \{\mathbf{xi}_i, \mathbf{xe}_i, \mathbf{xw}_i, Y_i\}$. Unfortunately, such a dataset is not available. Instead we have the MCBS, which has information on $I_i, E_i$ and $\mathbf{xz}_i$ but not on $\mathbf{c}_i$, and the HRS, which has information on $I_i, \mathbf{c}_i$ and $\mathbf{xz}_i$ but not on $E_i$.

Our estimation strategy is to combine information from the MCBS and HRS by way of multiple data imputation. To this end, we specify an auxiliary prediction model for the SAS variables ($\mathbf{c}_i$) that are missing from the MCBS, conditional on exogenous variables common in the two datasets:[32]

$$c_{ki}|j = \mathbf{xc}_i'\boldsymbol{\lambda}_k + \varepsilon_{3ki}, \tag{7}$$

where $k = 1, \ldots, 4$. Here $\mathbf{xc}_i$ denotes the vector of exogenous variables common in the two datasets, such as demographics, income, health status and education.[33] The disturbances $[\varepsilon_{31i}, \ldots, \varepsilon_{34i}]' \equiv \boldsymbol{\varepsilon}_{3i}$ follow a multivariate normal distribution for all types $j = 1, \ldots, m$:

$$\boldsymbol{\varepsilon}_{3i}|j \sim N(0, V_c).$$

The disturbances $\boldsymbol{\varepsilon}_{3i}$ are independent of $\boldsymbol{\varepsilon}_{12i}$ and $\eta_i|j$. Hence,

$$\mathbf{c}_i|j = XC_i\Lambda + \boldsymbol{\varepsilon}_{3i}, \tag{8}$$

[32]We treat SAS variables as exogenous, so the model for insurance demand and expenditure is conditional on these variables. The auxiliary model for SAS variables is needed only for imputation of missing data.

[33]The vector $\mathbf{xc}_i$ includes most of the variables in $\mathbf{xi}_i$ and $\mathbf{xe}_i$. The exception is that the second and third powers of age and interactions of age with gender and of marital status with gender as well as time trend are included in $\mathbf{xe}_i$ but not in $\mathbf{xc}_i$ to reduce the number of parameters to be estimated. See Table A-2.

where

$$XC_i = \begin{pmatrix} \mathbf{xc}_i' & \mathbf{0} & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \mathbf{xc}_i' & \mathbf{0} & \mathbf{0} \\ \mathbf{0}' & \mathbf{0}' & \mathbf{xc}_i' & \mathbf{0} \\ \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & \mathbf{xc}_i' \end{pmatrix},$$

and $\Lambda = [\boldsymbol{\lambda}_1', ..., \boldsymbol{\lambda}_4']'$. Thus, the disturbances of the structural system of equations (1)-(8), conditional on type $j$, follow a multivariate normal distribution with zero mean and variance-covariance matrix given by:

$$\begin{pmatrix} \sigma_{11} & \sigma_{12} & 0 & \mathbf{0} \\ \sigma_{12} & \sigma_{22} & 0 & \mathbf{0} \\ 0 & 0 & \sigma_j^2 & \mathbf{0} \\ \mathbf{0}' & \mathbf{0}' & \mathbf{0}' & V_c \end{pmatrix}.$$

We have now defined all the model parameters. We denote the parameter vector by:

$$\boldsymbol{\theta} = [\alpha_0, \alpha_1, \alpha_2, \boldsymbol{\alpha}_3', \boldsymbol{\alpha}_4', \boldsymbol{\beta}_1', ..., \boldsymbol{\beta}_m', \sigma_1^2, ...\sigma_m^2, \boldsymbol{\gamma}_m', \boldsymbol{\delta}_1', ..., \boldsymbol{\delta}_m', \sigma_{12}, \sigma_{22}, V_c, \Lambda'],$$

Next, to deal with health expenditure data missing from the HRS, we use the expenditure distribution implicit in the joint model for insurance and expenditure.[34] Let $\mathbf{x}_i$ denote the vector of unique variables contained in $\mathbf{xz}_i$ and $\mathbf{xc}_i$. We assume: (i) that the joint distribution of $I_i^*, E_i^*, \widehat{E}_i, E_i, I_i, \mathbf{c}_i$ conditional on $\mathbf{x}_i$ and $\boldsymbol{\theta}$, is the same in both datasets, and is as specified in section 4.1 and equation (8), and (ii) that $\mathbf{c}_i$ and $E_i$ are missing from the MCBS and the HRS, respectively, completely at random (using the definition of Gelman et al. (1995)).

Our approach to merging the two datasets can be described as "data fusion" - the com-

[34] We will show below that the missing expenditure data can be integrated out analytically without complicating our MCMC algorithm for simulation from the posterior of the model parameters. Therefore, we only have to perform multiple imputations of the SAS variables missing from the MCBS.

bination of data from distinct datasets, which can have some variables in common and some variables present in only one dataset. Rubin (1986) emphasized that the problem of data fusion can be cast as the problem of missing data, which, in turn, can be dealt with using Bayesian methods for multiple imputations from the posterior distribution of missing variables, conditional on common variables, as discussed in Gelman et al. (1995). This is the approach we adopt. Data fusion methods are often used in marketing to combine data from different surveys, such as product purchase and media exposure (e.g. Gilula et al. (2006)). Currently, there are few if any examples of data fusion in applied work in economics.

To proceed, let $\mathbf{C}^o$ denote the collection of $\mathbf{c}_i$'s that are observed, and $\mathbf{C}^m$ denote the collection of $\mathbf{c}_i$'s that are missing. Similarly, let $\mathbf{E}^o$ denote the collection of $E_i$'s that are observed, and $\mathbf{E}^m$ denote the collection of $E_i$'s that are missing. Thus, $\mathbf{c}_i \in \mathbf{C}^m$ iff $i \in MCBS$, and $\mathbf{c}_i \in \mathbf{C}^o$ iff $i \in$ HRS. Similarly, $E_i \in \mathbf{E}^m$ iff $i \in HRS$, and $E_i \in \mathbf{E}^o$ iff $i \in$ MCBS. The assumption that data are missing completely at random implies the missing data mechanism is independent of $I_i, E_i, \mathbf{c}_i, \mathbf{x}_i$. Hence, there is no need to specify an auxiliary missing data process that is separate from the structural model in (1)-(6). Assuming the HRS and MCBS are non-overlapping random samples from the same population, inference can be conducted out by stacking observations from the two datasets and imputing missing values using the data generating process in (1)-(6).

Let $S_i$ denote a survey indicator so that $S_i = 1$ if $i \in$ MCBS and $S_i = 0$ if $i \in$ HRS, and let $N^M$ and $N^H$ denote number of observations in the MCBS and HRS respectively. Let $N = N^M + N^H$ denote the number of observations in the combined dataset. Let $\mathbf{I} = [I_1, ..., I_N]'$ be a vector of Medigap indicators in the combined dataset. The probability density function of the observables $\mathbf{I}$, $\mathbf{E}^o$ and $\mathbf{C}^o$ conditional on exogenous variables $\mathbf{X} \equiv [\mathbf{x}_1, ..., \mathbf{x}_N]'$, survey indicators $\mathbf{S} \equiv [S_1, ..., S_N]$ and parameters $\boldsymbol{\theta}$ consists of two parts, corresponding to the MCBS and HRS subsets. To obtain the expression for probability density we: (i) substitute equations (4) and (8) into equation (3); and (ii) substitute (4) into (1). This gives us, conditional on type $j$, a system of equations for $I_i^*$, $E_i^*$, $\widehat{E}_i$, $\mathbf{c}_i$, in which the vector of disturbances has a multivariate normal distribution. At this point we can integrate out the latent variable $E_i^*$, which leaves us with the multivariate normal distribution of $I_i^*$, $\widehat{E}_i$ and $\mathbf{c}_i$. We also have to integrate out $\mathbf{c}_i$ from the MCBS subsample as these SAS variables are missing from the MCBS. So, in the MCBS subsample we are left with the following reduced-form model, conditional on type $j$:

$$\begin{aligned} I_i^*|j &= \alpha_0 \boldsymbol{\beta}_j' \mathbf{xe}_i + \alpha_1 \sigma_j^2 + \alpha_2 \sigma_j^2 \mathbf{xc}_i' \boldsymbol{\lambda}_1 + \boldsymbol{\alpha}_3' \mathbf{xi}_i + \boldsymbol{\alpha}_4' X C_i \boldsymbol{\Lambda} + \xi_{1i} && (9) \\ \widehat{E}_i|j &= \boldsymbol{\beta}_j' \mathbf{xe}_i + \boldsymbol{\gamma}_j' \mathbf{yi}_i + \xi_{2i} && (10) \\ I_i|j &= \iota(I_i^* > 0|j) && (11) \\ E_i|j &= max\{0, \widehat{E}_i|j\}, && (12) \end{aligned}$$

where $\iota(.)$ is an indicator function, and the errors $\xi_{1i}$ and $\xi_{2i}$ are given by:

$$\begin{aligned} \xi_{1i} &= \varepsilon_{1i} + \alpha_0 \varepsilon_{2i} + \alpha_2 \sigma_j^2 \varepsilon_{31i} + \boldsymbol{\alpha}_4' \boldsymbol{\varepsilon}_{3i} \\ \xi_{2i} &= \varepsilon_{2i} + \eta_i. \end{aligned}$$

The reduced-form errors $\xi_{1i}$ and $\xi_{2i}$ have a bivariate normal distribution:

$$\begin{matrix} \xi_{1i} \\ \xi_{2i} \end{matrix} |j \sim N \left( \begin{matrix} 0 \\ 0 \end{matrix}, \begin{bmatrix} \sigma_{11} + 2\alpha_0\sigma_{12} + \alpha_0^2\sigma_{22} + \boldsymbol{\alpha}_4' V_c \boldsymbol{\alpha}_4 + \alpha_2^2 \sigma_j^4 \cdot \mathrm{v}_c^{11} + 2\alpha_2\sigma_j^2 \sum_{l=1}^4 \cdot \alpha_{4l} \cdot \mathrm{v}_c^{1l} & \sigma_{12} + \alpha_0\sigma_{22} \\ \sigma_{12} + \alpha_0\sigma_{22} & \sigma_{22} + \sigma_j^2 \end{bmatrix} \right),$$

where $\mathrm{v}_c^{lk}$ denotes the $lk^{th}$ element of $V_c$.

Let $\mu_{1i} \equiv \alpha_0 \boldsymbol{\beta}_j' \mathbf{xe}_i + \alpha_1 \sigma_j^2 + \alpha_2 \sigma_j^2 \mathbf{xc}_i' \boldsymbol{\lambda}_1 + \boldsymbol{\alpha}_3' \mathbf{xi}_i + \boldsymbol{\alpha}_4' X C_i \boldsymbol{\Lambda}$ and let $s_\xi$ denote the standard deviation of $\xi_{1i}$. The joint probability density of $E_i$ and $I_i$, conditional on type $j$, in the MCBS subsample is that of a Tobit model (for $E_i$) with an endogenous binary explanatory variable ($I_i$). Its derivation is given in Wooldridge (ex.16.6):

$$
\begin{aligned}
& p(E_i, I_i|\mathbf{x}_i, j, \boldsymbol{\theta}, S_i = 1) = \\
& I_i \cdot \int_{-\mu_{1i}}^{\infty} g(E_i|\boldsymbol{\beta}_j'\mathbf{xe}_i + \gamma_{1j}Y_i + \gamma_{2j} + \frac{\sigma_{12}+\alpha_0\sigma_{22}}{s_\xi^2}\xi_{1i}, \sigma_{22}+\sigma_j^2 - \frac{(\sigma_{12}+\alpha_0\sigma_{22})^2}{s_\xi^2}) \cdot \frac{1}{s_\xi}\phi(\frac{\xi_{1i}}{s_\xi})d\xi_{1i} \\
+ \quad & (1-I_i) \cdot \int_{-\infty}^{-\mu_{1i}} g(E_i|\boldsymbol{\beta}_j'\mathbf{xe}_i + \gamma_{1j}Y_i + \frac{\sigma_{12}+\alpha_0\sigma_{22}}{s_\xi^2}\xi_{1i}, \sigma_{22}+\sigma_j^2 - \frac{(\sigma_{12}+\alpha_0\sigma_{22})^2}{s_\xi^2}) \cdot \frac{1}{s_\xi}\phi(\frac{\xi_{1i}}{s_\xi})d\xi_{1i},
\end{aligned}
$$

where

$$g(E|\mu, \sigma^2) = \left(\frac{1}{\sigma}\phi\left(\frac{(E-\mu)}{\sigma}\right)\right)^{\iota(E>0)} \left(1-\Phi\left(\frac{\mu}{\sigma}\right)\right)^{\iota(E=0)}. \tag{13}$$

In the HRS the SAS variables $\mathbf{c}_i$ are available, but $E_i$ is not. Hence, we have to integrate out $\widehat{E}_i$ and $E_i$. After the integration we are left with the following reduced-form model for the HRS subsample, conditional on type $j$:

$$
\begin{aligned}
I_i^*|j &= \alpha_0\boldsymbol{\beta}_j'\mathbf{xe}_i + \alpha_1\sigma_j^2 + \alpha_2\sigma_j^2 \cdot c_{1i} + \boldsymbol{\alpha}_3'\mathbf{xi}_i + \boldsymbol{\alpha}_4'\mathbf{c}_i + \nu_i & (14)\\
\mathbf{c}_i|j &= XC_i\Lambda + \boldsymbol{\varepsilon}_{3i} & (15)\\
I_i|j &= \iota(I_i^* > 0|j) & (16)
\end{aligned}
$$

where $\nu_i = \alpha_0\varepsilon_{2i} + \varepsilon_{1i}$ is normal with mean 0 and variance $s_\nu \equiv \sigma_{11} + 2\alpha_0\sigma_{12} + \alpha_0^2\sigma_{22}$. It is independent of $\boldsymbol{\varepsilon}_{3i}$.

The joint probability density of $I_i$ and $\mathbf{c}_i$ in the HRS subsample, conditional on type $j$, is given by the product of the likelihood of a probit model for $I_i$ and a multivariate normal probability density function for $\mathbf{c}_i$:

$$
\begin{aligned}
p(I_i, \mathbf{c}_i|\mathbf{x}_i, j, \boldsymbol{\theta}, S_i = 0) &= \Phi\left(\frac{\alpha_0\boldsymbol{\beta}_j'\mathbf{xe}_i + \alpha_1\sigma_j^2 + \alpha_2\sigma_j^2 \cdot c_{1i} + \boldsymbol{\alpha}_3'\mathbf{xi}_i + \boldsymbol{\alpha}_4'\mathbf{c}_i}{\sqrt{\sigma_{11} + 2\alpha_0\sigma_{12} + \alpha_0^2\sigma_{22}}}\right)^{I_i} \\
& \left(1 - \Phi\left(\frac{\alpha_0\boldsymbol{\beta}_j'\mathbf{xe}_i + \alpha_1\sigma_j^2 + \alpha_2\sigma_j^2 \cdot c_{1i} + \boldsymbol{\alpha}_3'\mathbf{xi}_i + \boldsymbol{\alpha}_4'\mathbf{c}_i}{\sqrt{\sigma_{11} + 2\alpha_0\sigma_{12} + \alpha_0^2\sigma_{22}}}\right)\right)^{1-I_i}
\end{aligned}
$$

$$\cdot \quad (2\pi)^{-\frac{K_c}{2}} |V_c|^{-\frac{1}{2}} \exp(-(\mathbf{c}_i - XC_i\mathbf{\Lambda})' V_c^{-1} (\mathbf{c}_i - XC_i\mathbf{\Lambda})/2).$$

To obtain the probability density of the observables unconditional on type $j$ we have to marginalize over the types by multiplying type-specific densities of observables by the type probabilities in (6) and summing the resulting products over the types. The probability density function of observables $\mathbf{E}^o$,$\mathbf{I}$, $\mathbf{C}^o$ conditional on exogenous variables $\mathbf{X}$, survey indicators $\mathbf{S} \equiv [S_1, ..., S_N]$ and parameters $\boldsymbol{\theta}$ is given by:

$$p(\mathbf{E}^o, \mathbf{I}, \mathbf{C}^o | \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}) = \prod_{i=1}^{N} (\sum_{j=1}^{m} (\int_{-\infty}^{\infty} \phi(y - \boldsymbol{\delta}_j' \mathbf{x} \mathbf{w}_i) \prod_{l \neq j}^{m} \Phi(y - \boldsymbol{\delta}_l' \mathbf{x} \mathbf{w}_i) dy$$
$$\cdot \quad p(E_i, I_i | \mathbf{x}_i, j, \boldsymbol{\theta}, S_i = 1)^{S_i=1} \cdot p(I_i, \mathbf{c}_i | \mathbf{x}_i, j, \boldsymbol{\theta}, S_i = 0)^{S_i=0})), \tag{17}$$

where $\boldsymbol{\delta}_m = \mathbf{0}$. It is easy to see that $\sigma_{11}$ is not identified separately from $\alpha_0$, $\alpha_1$, $\alpha_2$, $\boldsymbol{\alpha}_3$, $\boldsymbol{\alpha}_4$ and $\sigma_{12}$ in the sense that if we multiply $\sigma_{11}^{1/2}$ and all these parameters by a constant, the joint density will not change. Identification in such cases is usually achieved by the normalization $\sigma_{11} = 1$. But for the purposes of posterior simulation it is more convenient to normalize the variance of $\varepsilon_{1i} | \varepsilon_{2i}$, i.e. to set $\sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} = 1$, which implies the restriction $\sigma_{11} = 1 + \frac{\sigma_{12}^2}{\sigma_{22}}$.

## 4.3 Posterior Simulation Algorithm

Bayesian inference in this model can be simplified by data augmentation. In particular, both the MCBS and HRS subsamples are augmented by the latent vectors $\mathbf{I}^* = [I_1^*, ..., I_N^*]'$; the MCBS data are also augmented by the missing values $\mathbf{c}_i^m$, $i = 1, ..., N_M$ and by notional expenditure $\widehat{\mathbf{E}} = [\widehat{E}_1, ..., \widehat{E}_{N^M}]'$. The notional expenditure $\widehat{E}_i$ differs from actual expenditure $E_i$ only for observations with $E_i = 0$. Data augmentation which introduces artificial values of the dependent variable for observations with truncated outcomes is a standard approach to Bayesian inference in the Tobit model by way of the Gibbs sampler (due to Chib (1992)).

The fact that the $\mathbf{c}_i$ is missing from the MCBS subsample complicates simulation from the posterior distribution of parameters. In particular, if $\mathbf{c}_i^m$ is integrated out of the MCBS subsample, the usual normal and Wishart prior distributions for $\boldsymbol{\alpha}_4$ and $V_c$, respectively, are no longer conjugate to the probability density function of observables in the MCBS (as is clear from the expression for the variance of the reduced-form error $\xi_{1i}$ in equation (9)). There are no other known distributions which would serve as conjugate priors for $\boldsymbol{\alpha}_4$ and $V_c$. Hence, the conditional (on other parameters) posterior distributions of the Gibbs sampler blocks involving $\boldsymbol{\alpha}_4$ and $V_c$ would be of unknown form and would need to be sampled using a Metropolis-Hastings step. This involves a challenging task of choosing the proposal distribution for multidimensional vectors of parameters. For this reason in our algorithm we perform multiple imputations of $\mathbf{c}_i^m$ rather than integrating it our analytically.

Similarly, integrating out $E_i^*$, rather than augmenting the data with the latent $E_i^*$, would destroy conjugacy of the data density to the normal and gamma priors for $\sigma_{12}$, $\sigma^2$ and $\sigma_j^2$. However, these parameters are scalars, so proposal densities for Metropolis-Hastings steps are reasonably easy to choose. Hence, we analytically integrate out $E_i^m$ and $\widehat{E}_i$ in the HRS subsample, as well as $E_i^*$ in both the HRS and MCBS samples.

Both the HRS and MCBS are also augmented by latent type indicators $\mathbf{s} = [s_1, ..., s_N]'$, where $s_i = j$ if $i$'s type is $j$, and by latent type propensities $\mathbf{W} = [\widetilde{\mathbf{W}}_1', ..., \widetilde{\mathbf{W}}_N']$, where $\widetilde{\mathbf{W}}_i = [\widetilde{W}_{i1}, ..., \widetilde{W}_{im}]'$. Then the augmented data density conditional on $\mathbf{X}, \mathbf{S}$ and $\boldsymbol{\theta}$ can be written as follows:

$$\begin{aligned}
& p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W} | \mathbf{S}, \mathbf{X}, \boldsymbol{\theta}) \\
= \; & \prod_i^N \Big[ p(I_i^* | \mathbf{c}_i^m, \mathbf{x}_i, s_i = j, \boldsymbol{\theta}) \cdot p(I_i | I_i^*, \mathbf{c}_i^m, \mathbf{x}_i, s_i = j, \boldsymbol{\theta}) \cdot p(\widehat{E}_i | I_i^*, I_i, \mathbf{c}_i^m, \mathbf{x}_i, s_i = j, \boldsymbol{\theta}) \\
\cdot \; & p(E_i^o | \widehat{E}_i, I_i^*, I_i, \mathbf{c}_i^m, \mathbf{x}_i, s_i = j, \boldsymbol{\theta}) \cdot p(\mathbf{c}_i^m | \mathbf{x}_i, s_i = j, \boldsymbol{\theta}) \Big]^{S_i} \\
\cdot \; & [p(I_i^* | \mathbf{c}_i^o, \mathbf{x}_i, s_i = j, \boldsymbol{\theta}) \cdot p(I_i | I_i^*, \mathbf{c}_i^o, \mathbf{x}_i, s_i = j, \boldsymbol{\theta}) \cdot p(\mathbf{c}_i^o | \mathbf{x}_i, s_i = j, \boldsymbol{\theta})]^{1-S_i} \\
\cdot \; & p(s_i = j | \widetilde{\mathbf{W}}_i, \boldsymbol{\theta}) \cdot p(\widetilde{\mathbf{W}}_i | \mathbf{x}_i, \boldsymbol{\theta}), \qquad (18)
\end{aligned}$$

where superscript $o$ indicates values of $E_i$ and $\mathbf{c}_i$ that are observed and hence do not have to be imputed. The first two lines of (18) are for the MCBS observations, the third line is for the HRS observations, and the last line which involves type probabilities, is relevant for all observations.

After substituting in the expressions for the probability densities implied by the model in (1)-(6) the expression in (18) becomes:

$$\begin{aligned}
p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \boldsymbol{\theta}) = \prod_i [&\frac{1}{\sqrt{2\pi(1+\frac{\sigma_{12}^2}{\sigma_{22}}+2\alpha_0\sigma_{12}+\alpha_0^2\sigma_{22}-\frac{(\sigma_{12}+\alpha_0\sigma_{22})^2}{\sigma_{22}+\sigma_{s_i}^2})}} \\
\cdot \ \exp(-&\frac{(I_i^* - \alpha_0\boldsymbol{\beta}'_{s_i}\mathbf{xe}_i - \alpha_1\sigma_{s_i}^2 - \alpha_2\sigma_{s_i}^2 c_{1i}^m - \boldsymbol{\alpha}'_3\mathbf{xi}_i - \boldsymbol{\alpha}'_4\mathbf{c}_i^m - \frac{\sigma_{12}+\alpha_0\sigma_{22}}{\sigma_{22}+\sigma_{s_i}^2}(\widehat{E}_i - \boldsymbol{\beta}'_{s_i}\mathbf{xe}_i - \boldsymbol{\gamma}'_{s_i}\mathbf{yi}_i))^2}{2(1+\frac{\sigma_{12}^2}{\sigma_{22}}+2\alpha_0\sigma_{12}+\alpha_0^2\sigma_{22}-\frac{(\sigma_{12}+\alpha_0\sigma_{22})^2}{\sigma_{22}+\sigma_{s_i}^2})}) \\
\cdot \ (2\pi)^{-4/2}|V_c|^{-1/2}&\exp(-(\mathbf{c}_i^m - XC_i\Lambda)'V_c^{-1}(\mathbf{c}_i^m - XC_i\Lambda)/2) \\
\cdot \ \frac{1}{\sqrt{2\pi(\sigma_{22}+\sigma_{s_i}^2)}}&\exp(-\frac{(\widehat{E}_i - \boldsymbol{\beta}'_{s_i}\mathbf{xe}_i - \boldsymbol{\gamma}'_{s_i}\mathbf{yi}_i)^2}{2(\sigma_{22}+\sigma_{s_i}^2)}) \cdot (\iota(E_i^o = \widehat{E}_i)\cdot\iota(\widehat{E}_i \geq 0) + \iota(E_i^o = 0)\cdot\iota(\widehat{E}_i < 0))]^{S_i} \\
\cdot \ [\frac{1}{\sqrt{2\pi(1+\frac{\sigma_{12}^2}{\sigma_{22}}+2\alpha_0\sigma_{12}+\alpha_0^2\sigma_{22})}}&\exp(-\frac{(I_i^* - \alpha_0\boldsymbol{\beta}'_{s_i}\mathbf{xe}_i - \alpha_1\sigma_{s_i}^2 - \alpha_2\sigma_{s_i}^2 c_{1i}^o - \boldsymbol{\alpha}'_3\mathbf{xi}_i - \boldsymbol{\alpha}'_4\mathbf{c}_i^o)^2}{2(1+\frac{\sigma_{12}^2}{\sigma_{22}}+2\alpha_0\sigma_{12}+\alpha_0^2\sigma_{22})}) \\
\cdot \ (2\pi)^{-4/2}|V_c|^{-1/2}&\exp(-(\mathbf{c}_i^o - XC_i\Lambda)'V_c^{-1}(\mathbf{c}_i^o - XC_i\Lambda)/2)]^{1-S_i} \\
\cdot \ (\iota(I_i^* \geq 0)\cdot\iota(I_i = 1) + \iota(I_i^* < 0)&\cdot\iota(I_i = 0)) \cdot \left(\sum_{j=1}^m \prod_{k=1}^m \iota(\widetilde{W}_{ik} \in (-\infty, \widetilde{W}_{ij}])\right) \\
\cdot \ (\frac{1}{\sqrt{2\pi}})^m \cdot \exp(-&\sum_{i=1}^N(\widetilde{W}_{im}^2/2)) \cdot \exp(-\sum_{j=1}^{m-1}(\tilde{\mathbf{w}}_j - \mathbf{XW}\boldsymbol{\delta}_j)'(\tilde{\mathbf{w}}_j - \mathbf{XW}\boldsymbol{\delta}_j)/2),
\end{aligned} \tag{19}$$

where $\mathbf{XW} = [\mathbf{xw}_1, ..., \mathbf{xw}_N]'$ and $\tilde{\mathbf{w}}_j = [\widetilde{W}_{1j}, ..., \widetilde{W}_{Nj}]'$ for $j = 1, ..., m$.

For the purposes of Bayesian inference via MCMC it is convenient to split the parameter vector $\boldsymbol{\theta}$ into the following blocks:

1. $\alpha_0$
2. $\boldsymbol{\alpha} \equiv [\alpha_1, \alpha_2, \boldsymbol{\alpha}'_3, \boldsymbol{\alpha}'_4]'$
3. $\boldsymbol{\beta}_j$ for $j = 1, ..., m$

4. $\boldsymbol{\gamma}_j$ for $j = 1, ..., m$

5. $h_j \equiv \frac{1}{\sigma_j^2}$, $j = 1, ..., m$.

6 $\boldsymbol{\Lambda} \equiv [\boldsymbol{\lambda}_1^{'}, ..., \boldsymbol{\lambda}_4^{'}]'$,

7. $H_c \equiv V_c^{-1}$;

8. $h_{22} \equiv \sigma_{22}^{-1}$;

9. $\sigma_{12}$

10. $\boldsymbol{\delta}_j$, $j = 1, ..., m-1$;

Where possible, we specify natural conjugate prior distributions for these parameters blocks, and specify that in the prior these blocks are independent, i.e

$$p(\boldsymbol{\theta}) = p(\alpha_0)p(\boldsymbol{\alpha}) \prod_{j=1}^{m} p(\boldsymbol{\beta}_j) \prod_{j=1}^{m} p(\boldsymbol{\gamma}_j) \prod_{j=1}^{m} p(h_j) \prod_{k=1}^{4} p(\boldsymbol{\lambda}_k) p(\sigma_{12}) p(h_{22}) p(H_c) \prod_{j=1}^{m-1} p(\boldsymbol{\delta}_j). \quad (20)$$

We specify the hyperparameters of these prior distributions so as to allow substantial prior uncertainty about the parameter values. The priors are discussed in detail in Appendix A-1.

Let **data** denote the collection $\langle \mathbf{I}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{X}, \mathbf{S} \rangle$. Then the joint posterior distribution of the parameters and the latent and missing data $p(\boldsymbol{\theta}, \mathbf{I}^*, \widehat{\mathbf{E}}, \mathbf{C}^m, \mathbf{W}, \mathbf{s} | \mathbf{data})$ is proportional to the product of (19) and (20). To simulate from this posterior distribution we construct a Gibbs sampling algorithm with Metropolis within Gibbs steps which cycles between the conditional posterior distributions of blocks of parameters and vectors of latent and missing variables $\mathbf{I}^*, \widehat{\mathbf{E}}, \mathbf{C}^m, \mathbf{W}, \mathbf{s}$. The details of the algorithm are given in Appendix A-2.

# 5 Results

The exact specification of the equations of the model in terms of the demographic and health status characteristics included in each equation is given in Appendix A-3. In particular, in Table A-2 we show our exclusion restrictions in tabular form. As for the expenditure distribution, we have specified $m = 5$. In a companion paper (Keane and Stavrunova (2011)) we discuss the SMTobit model in detail and show that a 6 component mixture provides the best fit to the expenditure distribution in the MCBS sample used in this paper, while a model with 5 components fares only slightly worse.[35] The number of components in this study is a compromise between model fit and the mixing properties of the posterior simulator. We use m=5 because the posterior simulator exhibited slow convergence with $m = 6$.

In order to investigate sources of advantageous selection we estimate a sequence of five models that progressively add more potential SAS variables to the insurance equation. This procedure is similar to FKS. In the first model the insurance equation contains only expenditure risk and insurance pricing variables. The second model adds income and education. The third adds cognitive ability, financial planning horizon and longevity expectations. The fourth adds risk tolerance, the variance of the expenditure forecast error and an interaction between risk tolerance and variance. Finally, the fifth model adds ethnicity, marital status and an interaction of gender with age. To assess the bias in the selection and moral hazard effects that arises from failure to account for correlation between unobserved determinants of Medigap status and expenditure, we also estimate the fifth model with $\sigma_{12}$ set to zero.

Due to the presence of latent variables and mixture components, the output of the posterior simulator exhibits a high degree of autocorrelation. To allow the simulator to explore the parameter space adequately, the algorithm was allowed to run for an extended period

[35] To compare models with different numbers of components we use the modified cross-validated log-scoring rule developed in Geweke and Keane (2007), which is less computationally demanding than the comparison based on marginal likelihoods, which is a standard approach to model selection in Bayesian statistics.

of time. We obtained 1,200,000 draws from the posterior distribution, discarded the first 200,000 draws as a burn-in and using every 1000th of the remaining 1,000,000 draws for analysis. In the fifth model the autocorrelation in these 1,000 draws ranges from 0 (for parameters of the $\mathbf{c}^m$ distribution and the coefficients of the exogenous covariates in the insurance equation) to 0.12, 0.29 and 0.68 (for $\alpha_0$, $\sigma_2$, $\sigma_{12}$) and to 0.70 (for parameters of the expenditure distribution for the type with the lowest probability). The serial correlation coefficients for the parameters $\sigma_j^2$ are between 0.33 and 0.02, while those for $\gamma_{2j}$ are between 0.28 and 0.05. Thus, the serial correlation is low for the parameters that are most important for our analysis (i.e. the insurance equation parameters and $\gamma_{2j}$). The relative numerical efficiency ranges from 0.08 (for the parameters with the highest autocorrelation) to 1.6. All parameters pass the formal test of convergence suggested in Geweke (1992).

## 5.1 Model Fit

In order to examine the fit of the model, we simulate artificial samples of $E_i$ and $I_i$, conditional on $\mathbf{x}_i$, for 1000 draws from the posterior distribution of parameters. In particular, for each $\boldsymbol{\theta}^k$ drawn from the posterior distribution we simulate artificial data for each $i = 1, ..., N$ as follows:

1. Latent types $s_i^k \sim p(s_i|\mathbf{xw}_i, \boldsymbol{\delta}_1^k, ..., \boldsymbol{\delta}_m^k)$ using (5);

2. Missing SAS variables in the MCBS subsample $\mathbf{c}_i^{mk} \sim p(\mathbf{c}_i^m|XC_i, \Lambda^k, V_c^k)$ using (8);

3. Latent data $[I_i^{*k}, E_i^{*k}]' \sim p(I_i^*, E_i^*|\mathbf{xi}_i, S_i \cdot \mathbf{c}_i^{mk} + (1 - S_i) \cdot \mathbf{c}_i^o, \mathbf{xe}_i, s_i^k, \boldsymbol{\beta}_{s_i^k}^k, \sigma_{12}^k, \sigma_{22}^k, \sigma_{s_i^k}^{2k})$ using (3) and (4). To obtain insurance status set $I_i^k = \iota(I_i^{*k} > 0)$.

4. Notional expenditure $\widehat{E}_i^k \sim p(\widehat{E}_i^k|Y_i, E_i^{*k}, I_i^k, s_i^k, \boldsymbol{\gamma}_{s_i^k}^k, \sigma_{s_i^k}^{2k})$ using (1). To obtain expenditure set $E_i^k = \widehat{E}_i^k \cdot \iota(\widehat{E}_i^k > 0)$.

We compare these simulated data to the actual data, focussing on the fit of the full model, which includes all potential SAS variables.

In Figure 2 panel (a) we compare model predictions vs. actual values of health expenditure. To construct panel (a), we partition the actual data by deciles of predicted expenditure, and calculate means of actual and predicted expenditure within each decile. Figure 2a plots the predicted vs actual means (along with the 5th and 95th percentiles of the posterior of predicted expenditure).[36] Figure 2 panel (b) shows the fit of the model to the probability of Medigap coverage, using an analogous procedure.

As we see in Figures 2a and 2b, the model provides a good fit to both health expenditure and insurance coverage. When we plot predicted vs. actual values for mean expenditure and probability of insurance coverage, the points (for each decile) are very close to the 45 degree line. And the actual values are almost always contained within the 5th and 95th percentiles of the predictions.

Next, in Figure 2c, we examine how well the model fits the relationship between health expenditure and insurance coverage. We split the MCBS data into 10 expenditure groups (as in panel (a)). For each group, we plot the probability of Medigap coverage against average expenditure (solid red line). It is interesting that, in the MCBS data, the probability of Medigap coverage is rising with expenditure at low levels of expenditure (adverse selection), but falling with expenditure at higher levels of expenditure (advantageous selection).

Figure 2c also shows model predictions for average expenditure $AE_g^k$ and insurance coverage $AI_g^k$ within each expenditure decile. We plot predictions based on 20 random draws $\boldsymbol{\theta}^k$ from the posterior. The predictions are the blue and black dots in Figure 2c. As we see,

[36] More precisely, we first find the mean of predicted expenditure over draws $k$ for each person $i$, that is $\overline{E}_i \equiv 10^{-3} \sum_{k=1}^{10^3} (E_i^k)$. We then we split the MCBS sample into ten deciles based on the sample distribution of $\overline{E}_i$. Finally, within each decile $g$, $g = 1, ..., 10$, and for each draw $k$, we compute mean expenditure $AE_g^k \equiv N_g^{-1} \sum_{i \in g} (E_i^k)$, where $N_g$ is the number of observations in subsample $g$. In panel (a) we plot the average of $AE_g^k$ over $k$ against the mean of *actual* expenditure for individuals falling into decile $g$ (red dots). For each $g$, we also plot the 5th and 95th percentiles of the series $AE_g^k$ to show the uncertainty about the predictions due to the posterior distribution of parameters (blue dots).

the predicted relationship between insurance and expenditure is very similar to the actual relationship. The model captures the inverted-U shape we see in the MCBS data quite well.

Finally, in panel (d) we show the fit of the model to the variance of expenditure. To construct panel (d) we use the same expenditure deciles as for panel (a), and plot the variance of the actual vs. predicted expenditure within each decile. Panel (d) shows that the fit of the model to the variance of expenditure is also quite good. Note that the variance increases as we move to higher expenditure deciles.

Figure 3 presents kernel density estimates for actual expenditure in the MCBS (red line), as well as the density of predicted expenditure ($E_i^k$) generated by our model (black line). Panel (a) shows the density plot for the entire support of the expenditure distribution. Notice that the distribution has an extremely long right tail, and the smooth mixture of Tobits does a good job in capturing this complex shape - i.e. in panel (a) the predicted and actual data densities are almost indistinguishable. Panels (b) and (c) present density plots for the [$0, $20,000] and [$20,000, $100,000] intervals of the support of the expenditure distribution (these intervals together contain more than 99% of the sample distribution of expenditure). The fit is very good even for these more narrowly defined intervals.

In order to examine the differences between latent types, we order the mixture components by the level of the health expenditure forecast error variance. Type 1 has the lowest variance, while type 5 has the highest.[37] We order by variance because it differs substantially

[37]The parameters of mixture components are not identified with respect to permutations of component labels. For example, the value of the likelihood function of a mixture of two types will not change if type 1 is relabelled as type 2, and vice versa. As a result, the likelihood function and the posterior distribution of parameters is multimodal with $m!$ modes corresponding to $m!$ permutations of the $m$ component labels. This can create problems for posterior simulation via the Gibbs sampler, because the simulator can get stuck in one of the posterior modes and not fully explore the entire posterior distribution (Celeux et al. (2000)). One solution to this problem, proposed by Fruhwirth-Schnatter (2001), is random permutation of component labels after each iteration of the Gibbs sampler. Another solution, proposed by Geweke (2006), is to use a permutation-augmented simulator. For permutation-sensitive functions of interest this amounts to reordering of the output from the usual Gibbs sampler (i.e., without the random permutation step) according to inequality constraints which identify the component labels, and using the reordered output for inference. In this paper we use the approach of Geweke (2006). We run the Gibbs sampling algorithm with the Metropolis-Hastings steps as described in Appendix A-2. We then use the resulting output directly for

Figure 2: Model fit: Conditional Moments of Medigap Probability and Expenditure

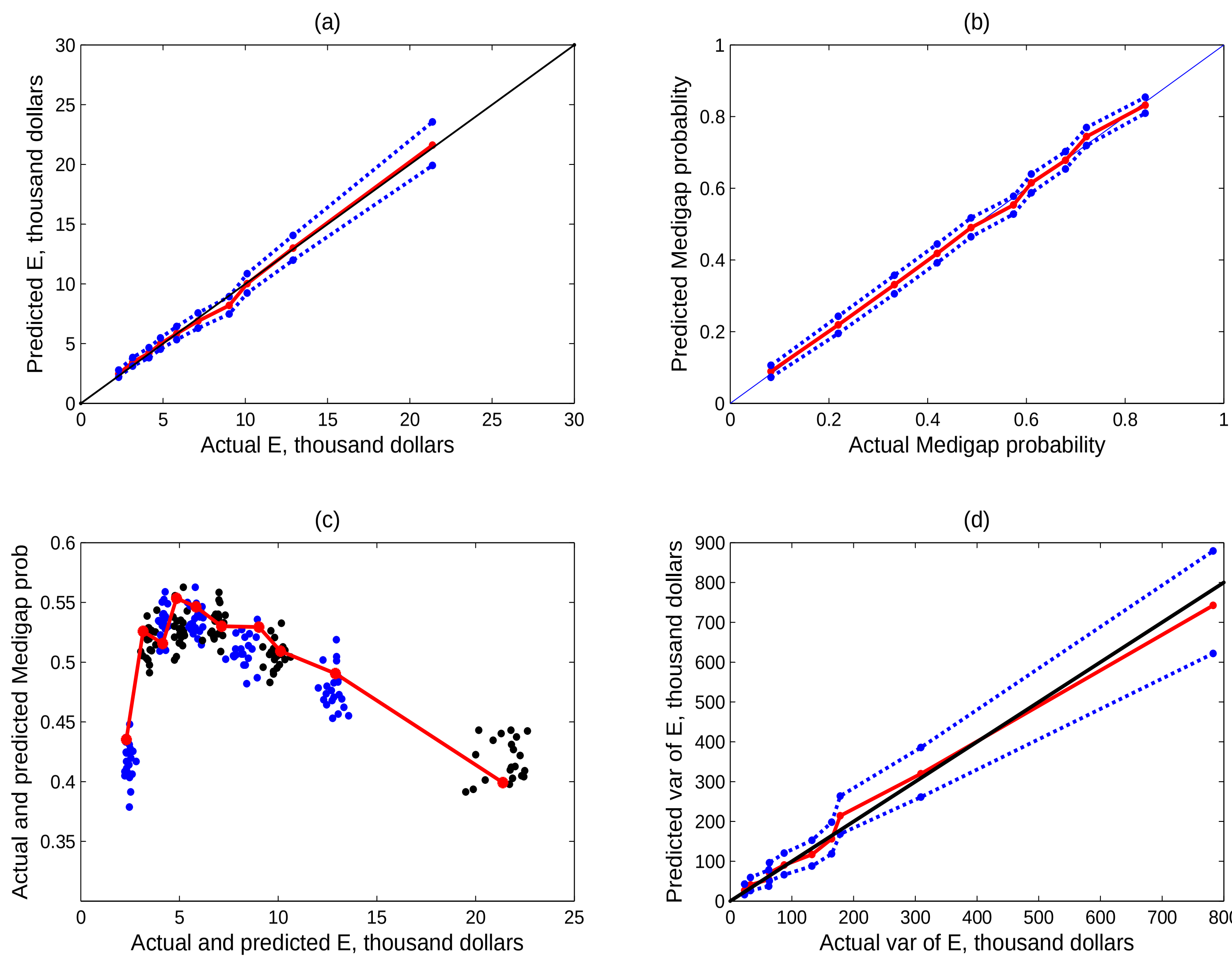
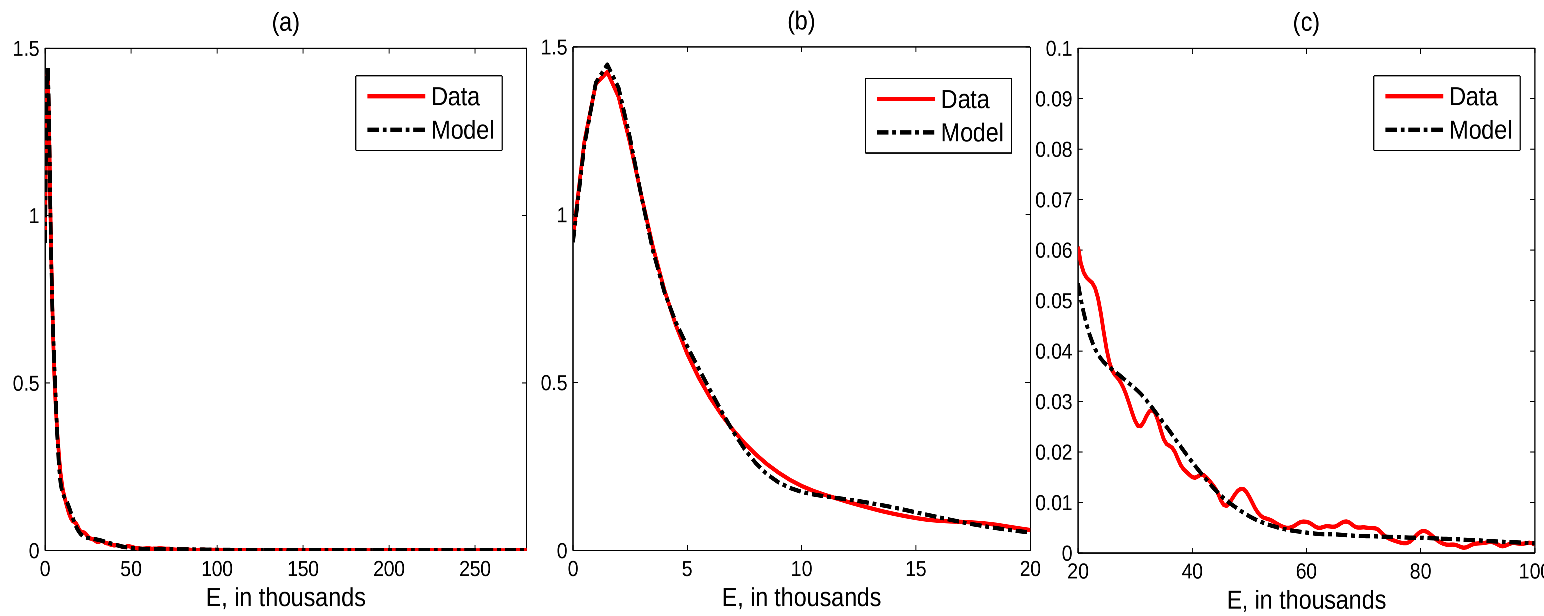


Figure 3: Model fit: Expenditure Distribution



across types and separates the components of the mixture quite well. Munkin and Trivedi (2010) also found that mixture components were most easily separated by variance in their application of a discrete mixture model to drug expenditure of Medicare beneficiaries. Ordering by variance also has the advantage that it renders our types easily interpretable, because the conditional (on covariates) means and variances of health expenditures tend to be strongly positively related (Deb et al., 2010).

Table 3 reports some key type-specific parameters and functions of interest.[38] As expected, the ranking of types by expenditure risk corresponds closely to that by variance. In fact, there is a perfect rank correlation (see the $1^{st}$ and $2^{nd}$ rows of Table 3). Types 1 and 2 have the lowest expenditure risks $E(E_i^*|\text{type}_i = j, \mathbf{data})$, so they are the healthiest types. Together they make up about 71% of the sample. Type 5, which makes up only 3% of the sample, is the unhealthiest type. Their expected expenditure is \$63,000 per year.

The posterior distributions of the type-specific income effects $\gamma_{1j}$ are presented in row 3 of Table 3. The income effect is generally small, and is increasing as health status deteriorates, with the exception of the unhealthiest type. For example, for the healthiest type a one standard deviation increase in income (i.e. roughly $\$50,000$ in the combined HRS/MCBS sample) would lead to only a $0.007 \cdot 5 = 0.035$ thousand dollars (or \$35) increase in health care expenditure. For the type with the largest income effect (type 4) a one standard deviation increase in income raises expenditure by \$1,165. Averaging across all types, the mean effect of a \$50,000 increase in income on expenditure is only \$327. This is very close to the OLS

inference about permutation-invariant functions of interest, such as the population mean of the moral hazard effect. But for inference about permutation-sensitive functions of interest, such as moral hazard effects for different types $j$, we use the output reordered according to the inequality restrictions on $\sigma_j^2$.

[38] The posterior mean of the type probabilities, $p(\text{type}_i = j|\mathbf{data})$, was computed as the average of $\iota(s_i^k = j)$ over $i$ and $k$, while the 5th and 95th percentiles are those of the series $\frac{1}{N}\sum_{i=1}^{N} \iota(s_i^k = j)$ for $k = 1, ...10^3$. These computations approximate the posterior mean and percentiles (over the posterior of parameters) of $\frac{1}{N}\sum_{i=1}^{N} P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)$. Similarly, the posterior mean of the expenditure risk $E(E_i^*|\text{type}_i = j, \mathbf{data})$ was computed as the average of $E_i^{*k}$ over $i$ and $k$ such that $\iota(s_i^k = j)$. The 5th and 95th percentiles are those of the series of $E_i^{*k}$ averaged over $i$ such that $\iota(s_i^k = j)$ for $k = 1, ...10^3$. These computations approximate the posterior mean and percentiles of $\frac{1}{N}\sum_{i=1}^{N} \mathbf{xe}_i \boldsymbol{\beta}_j \cdot P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)$.

Table 3: Type-specific characteristics: posterior means and 5th and 95th percentiles

| Variable | Type 1 | Type 2 | Type 3 | Type 4 | Type 5 |
|---|---|---|---|---|---|
| Std. deviation of $\eta_i$, $\sqrt{\sigma_j^2}$, thousand dollars | 0.72<br>(0.68, 0.75) | 1.9<br>(1.8, 2.0) | 4.5<br>(4.2, 4.9) | 11.5<br>(10.1, 12.5) | 33.1<br>(31.0, 35.5) |
| $E(E^*\|\text{type}_i = j, \mathbf{data})$, thousand dollars | 0.60<br>(0.54, 0.66) | 2.92<br>(2.72, 3.13) | 9.46<br>(8.7, 10.3) | 24.5<br>(22.3, 29.9) | 63.3<br>(56.7, 70.5) |
| Income effect $\gamma_{1j}$, thousand dollars | 0.007<br>(0.003,0.013) | 0.055<br>(0.042, 0.068) | 0.126<br>(0.029, 0.225) | 0.233<br>(-0.017, 0.728) | 0.074<br>(-1.72, 2.05) |
| Moral hazard effect $\gamma_{2j}$, thousand dollars | 1.21<br>(1.1, 1.3) | 1.88<br>(1.7, 2.1) | 1.74<br>(1.1, 2.5) | 2.145<br>(-0.2, 4.3) | 1.64<br>(-7.8, 10.8) |
| $P(\text{type}_i = j\|\mathbf{data})$ | 0.394<br>(0.38, 0.41) | 0.315<br>(0.30, 0.33) | 0.166<br>(0.15, 0.18) | 0.094<br>(0.08, 0.10) | .031<br>(0.027, 0.036) |

* Note: We measure expenditure in thousands of dollars. In row 3 the income effect on expenditure $\gamma_{1j}$ is for income measured in tens of thousands of dollars. So, e.g., the type-specific effect of a one standard deviation increase in income ($50,000) would be computed as $\gamma_{1j} \cdot 5$ thousand dollars (i.e., $35 for Type 1).

estimate reported in Table 2.

In the next two sub-sections we discuss, in turn, the selection and moral hazard effects implied by the model.

## 5.2 The Adverse (or Advantageous) Selection Effect

One key focus of this paper is the selection effect; i.e., the relation between expenditure risk and Medigap insurance status, conditional on pricing variables and potential SAS variables. Figure 4 shows how this relationship changes as we progressively add potential SAS variables to the insurance equation. This figure plots the distribution of the marginal effects of a one standard deviation increase in expected expenditure $E_i^*$ (12.7 thousand dollars)[39] on the probability of having Medigap insurance.[40]

[39] This is the standard deviation of $N \cdot 10^3$ simulated values $E_i^{*k}$.

[40] We evaluate the marginal effects for all individuals $i = 1, ..., N$ and for 1000 draws from the posterior distribution of parameters, replacing $E_i^*$ and the unobserved components of $\mathbf{SAS}_i$ ($\sigma_{s_i}^2$ and $\mathbf{c}_i^m$) with $E_i^{*k}$, $\sigma_{s_i^k}^{2k}$ and $\mathbf{c}_i^{mk}$ simulated as discussed in the previous section. Figure 4 plots the histograms of the resulting $N \cdot 10^3$ marginal effects, and shows sample averages and standard deviations of these effects.

The results in Figure 4 Panel (a) correspond to the benchmark model, where the insurance equation (3) contains only expenditure risk and pricing variables (no SAS variables). Here the relationship between expenditure risk and Medigap is negative; an increase in expenditure risk by 12.7 thousand dollars *decreases* the probability of Medigap coverage on average by 0.027. This implies advantageous selection, which is consistent with the findings of FKS.

Next we follow FKS and include potential SAS variables that insurance companies cannot legally use in pricing. Adding income and education (panel (b)) weakens the relationship between risk and insurance to almost zero. Adding cognitive ability, financial planning horizon and longevity expectations (panel (c)) changes the sign of the relationship - it becomes positive, but the effect is small: a one standard deviation increase in $E_i^*$ increases the probability of Medigap coverage on average by 0.026. In panel (d) we add risk tolerance and the variance of the forecast error, as well as their interaction. This further increases the marginal effect to 0.066. Thus, our SAS variables explain advantageous selection in the sense that, once we condition on these additional dimensions of private information, we find adverse selection on health as predicted by theory.

Next, in addition to the variables used in FKS, we also consider race and marital status as potential sources of adverse/advantageous selection. These variables can affect both tastes for insurance and health care expenditure, but cannot be legally used to price Medigap policies. Thus, race and marital status are also potential SAS variables. In Figure 4 panel (e) we add these variables to the insurance equation. Doing so reduces the average marginal effect of $E_i^*$ from 0.066 to 0.055. Thus, these variables are a source of adverse selection. In particular, blacks and Hispanics have a relatively low probability of purchasing Medigap, and they have relatively low expected expenditure (see Table 5).

Overall, these results support the results of FKS: conditional on SAS variables we also find adverse selection into Medigap. But we find that additional demographics like race are an important source of selection, and our estimate of the adverse selection effect is 1/3 smaller than theirs (5.5% vs. 9%).

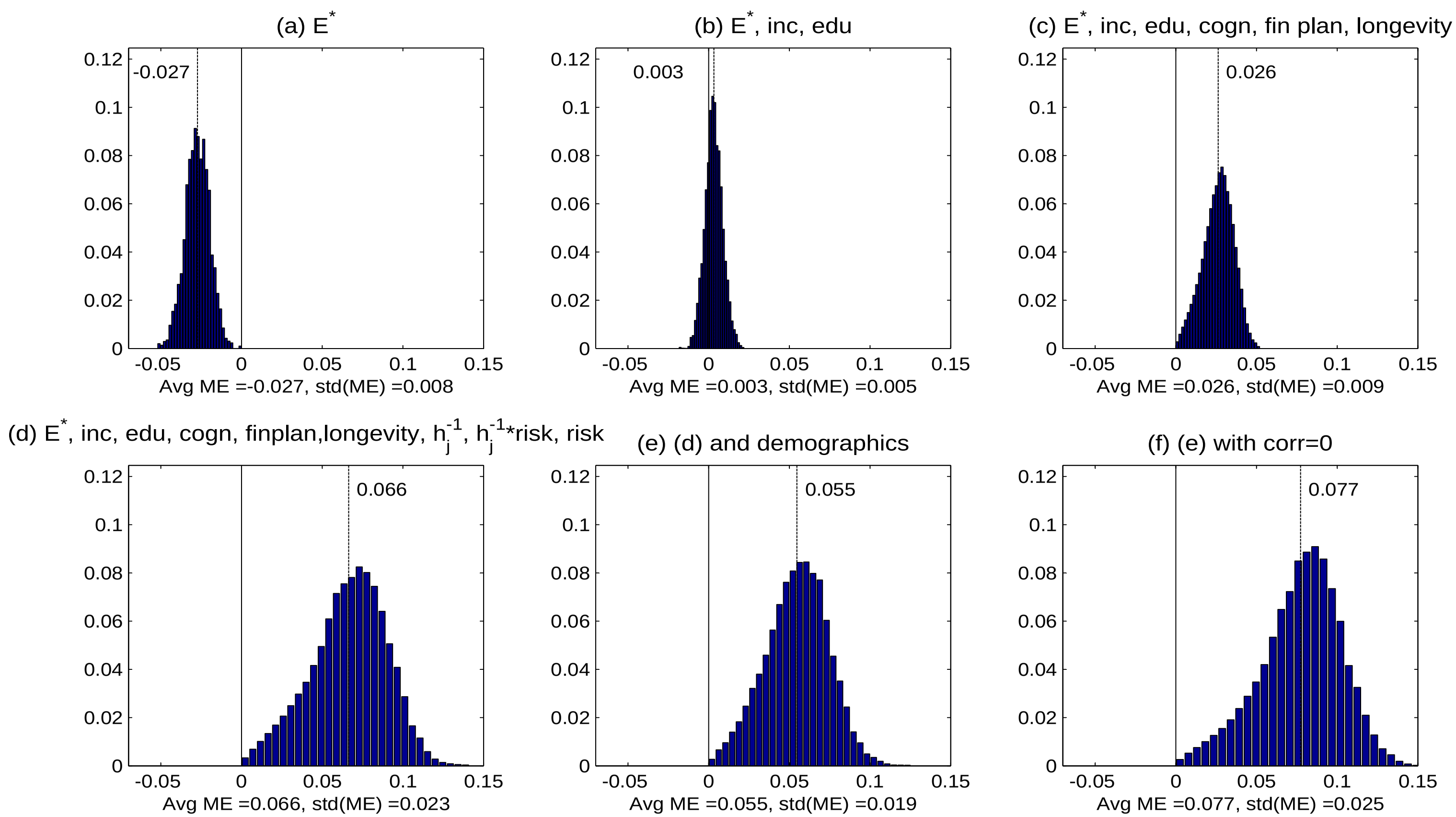
We next examine in more detail how particular variables affect the demand for Medigap insurance. Table 4 presents marginal effects of covariates on the probability of Medigap coverage. The effects are evaluated for a "median" individual for whom: (i) exogenous characteristics are set to their sample medians, (ii) the $E_i^*$ and $\sigma^2_{s_i}$ are set to their medians over $E_i^{*k}$ and $\sigma^{2k}_{s_i^k}$, and (iii) $\mathbf{c}_i$ is set to its median in the HRS subsample. We present the mean and the $5^{th}$ and $95^{th}$ percentile of the series of effects evaluated at 1000 draws from the posterior distribution. For continuous variables we report the change in Medigap probability brought about by a one standard deviation increase in the variable of interest from these median levels. For $E_i^*$ and $\sigma^2_{s_i}$ the marginal effects correspond to a one standard deviation increase in $E_i^{*k}$ and $\sigma^2_{s_i^k}$, while for $\mathbf{c}_i$ they correspond to a one standard deviation increase in the HRS subsample.[41] Table 4 also summarizes the posterior distributions of the correlation coefficient between $\varepsilon_1$ and $\varepsilon_2$ ($\rho$) and the variance of $\varepsilon_2$, ($\sigma_{22}$).

The first column of Table 4 is for the basic model containing only expected expenditure ($E_i^*$) and pricing variables (gender, age, region) in the insurance equation. The subsequent columns progressively add demographic and behavioral variables that are potential sources of selection (SAS variables). Consistent with Figure 4, the effect of $E_i^*$ goes from -0.03 in the basic model (advantageous selection) to +0.06 in the full model (adverse selection).

The results in Table 4 column (2) suggest that, conditional on expenditure risk, the probability of Medigap coverage is higher for females, increases with age, education and income, and varies substantially by region. But in column (3) the inclusion of the behavioral SAS variables eliminates the effect of education and greatly reduces the positive effect of income. Among the behavioral variables, we see that cognitive ability has by far the largest effect on the probability of Medigap coverage.

[41] Note that the marginal effects of $E_i^*$ in Table 4 do not correspond to those in Figure 4, as the former are for a median individual, while the latter correspond to the whole sample distribution.

Figure 4: Posterior distribution of the marginal effect of $E_i^*$ on the probability of Medigap coverage



In the 4th column we add the risk tolerance and variance measures. This has little effect on the impacts of other variables, but raises the effect of a one standard deviation increase in $E_i^*$ to 0.08, which clearly implies adverse selection. To assess the magnitude of this effect, note that the probability of Medigap coverage in the combined HRS/MCBS sample is 50%. Thus, 0.08 corresponds to roughly a 16% increase in Medigap coverage when expenditure risk increases by one standard deviation, *ceteris paribus.*

Finally, the inclusion of race and marital status variables in column (5) causes the effects of cognition to drop from 0.17 to 0.08, and the effect of $E_i^*$ to fall from 0.08 to 0.06. The indicators for black and Hispanic are among the most important determinants of Medigap status - they both decrease the probability of Medigap by 0.24, which is roughly a 50% drop.

As we noted, cognition has a much bigger effect on Medigap coverage than other behavioral variables (e.g. risk tolerance, longevity expectation, etc.). In the most general model (i.e., column (5)), an increase in the cognitive ability factor to one standard deviation above the median increases probability of Medigap coverage by 0.08 on average. This effect is estimated rather precisely - 90% of the mass of the posterior is between 0.06 and 0.10.

The variance of the expenditure forecast error, $\sigma_j^2$, also has a large effect on demand for insurance. As we see in rows 2 and 3 of Table 4, a one standard deviation increase in $\sigma_j^2$ (holding $E_i^*$ constant) decreases the probability of Medigap by 0.05 for people at the median of the risk tolerance distribution, and by 0.06 for people at the 90th percentile of risk tolerance distribution. Thus, variance is an important source of advantageous selection. FKS give several potential explanations for the surprising negative effect of variance, including crowding out of Medigap by Medicaid in the case of catastrophic health care expenses and the underweighting of small probabilities of a large loss.

The probability of Medigap coverage also differs substantially by region. According to the full model in column (5), residents of New England, the West South Central and Mountain census divisions are 7 to 21 percentage points less likely to have Medigap than people in the reference category (i.e., non-response). People in the East North Central, South Atlantic and Pacific regions are 15 to 16 percentage points more likely to have Medigap. For most regions the relative probability of Medigap is not greatly affected by controls for the SAS variables. But the results for the Pacific region are greatly affected. In the baseline model of column (1), the Pacific region is 24 percentage points below the reference group, while in the full model it is 16 points higher. This implies that residents of the Pacific region have SAS characteristics that make them unlikely to buy Medigap.

Finally, we examine the issue of selection on *unobservable* determinants of insurance coverage and expenditure risk, an issue that was not considered by FKS. Interestingly, in all models the correlation between the unobservables $\varepsilon_1$ and $\varepsilon_2$ is strongly *negative*. This means that selection with respect to *unobserved* expenditure risk is advantageous, even when all the SAS variables are included.[42] *Prima facie*, this result appears to contradict the finding of FKS that *observed* SAS variables can explain advantageous selection. This, in turn, raises a puzzle of why we obtained similar results to FKS in Figure 4.

The most plausible explanation for the similarity between our results and those of FKS is that the health status variables included in the prediction model of FKS capture most of the information relevant when individuals form an expectation about future health care costs and make a decision about Medigap insurance status. Indeed, our results indicate that the standard deviation of the unobservable component of expenditure risk, $\varepsilon_{2i}$, is very small compared to the standard deviation of expenditure risk $E_i^*$ itself (i.e., 0.55[43] vs. 12.7 thousand dollars). This suggests that any systematic difference in expenditure risk between individuals with and without Medigap that is left unexplained by the observable health status characteristics is also small. Hence, results about the extent of adverse selection obtained from a model that does not account for the correlation between $\varepsilon_{1i}$ and $\varepsilon_{2i}$ should not be

[42]Note that $\sigma_{12} < 0$ means that, ceteris paribus, people with higher expected expenditure $E_i^*$ tend to have lower demand for insurance.

[43]The posterior mean of $\sigma_{22}$ is equal to 0.31, while the posterior mean of $\sqrt{\sigma_{22}}$ is equal to 0.55.

Table 4: Marginal effects of individual characteristics on the probability of Medigap coverage.

| | (1) No SAS variables | | (2) Add hgc and inc | | (3) Add cogn, finpln, praliv75 | | (4) Add risktol and $\sigma^2_{s_i}$ | | (5) Add ethnicity and marst | |
|---|---|---|---|---|---|---|---|---|---|---|
| Variable | Post. mean | 5th-95th prct. | Post. mean | 5th-95th prct. | Post. mean | 5th-95th prct. | Post. mean | 5th-95th prct. | Post. mean | 5th-95th prct. |
| $E^*$ | -0.03 | -0.04 , -0.02 | 0.00 | -0.01 , 0.01 | 0.03 | 0.02 , 0.04 | 0.08 | 0.05 , 0.10 | 0.06 | 0.04 , 0.08 |
| $\sigma^2_{s_i}$, risk 50th pr | | | | | | | -0.05 | -0.08 , -0.03 | -0.04 | -0.07 , -0.02 |
| $\sigma^2_{s_i}$, risk 90th pr | | | | | | | -0.06 | -0.10 , -0.03 | -0.06 | -0.11 , -0.02 |
| female | 0.02 | 0.01 , 0.04 | 0.09 | 0.08 , 0.10 | 0.11 | 0.09 , 0.13 | 0.25 | 0.20 , 0.30 | 0.09 | 0.06 , 0.11 |
| age, male | 0.03 | 0.02 , 0.04 | 0.07 | 0.06 , 0.09 | 0.10 | 0.07 , 0.13 | 0.10 | 0.07 , 0.14 | 0.08 | 0.05 , 0.11 |
| New Eng | -0.10 | -0.13 , -0.06 | -0.08 | -0.12 , -0.04 | -0.09 | -0.14 , -0.04 | -0.09 | -0.14 , -0.04 | -0.07 | -0.11 , -0.03 |
| Mid Atl | 0.04 | 0.01 , 0.08 | 0.06 | 0.02 , 0.10 | 0.08 | 0.03 , 0.12 | 0.09 | 0.04 , 0.13 | 0.07 | 0.03 , 0.10 |
| East North Cent | 0.20 | 0.16 , 0.23 | 0.17 | 0.14 , 0.21 | 0.16 | 0.11 , 0.21 | 0.17 | 0.12 , 0.22 | 0.15 | 0.11 , 0.19 |
| West North Cent | -0.03 | -0.07 , 0.00 | 0.03 | -0.01 , 0.06 | 0.04 | -0.00 , 0.09 | 0.05 | 0.00 , 0.10 | 0.06 | 0.02 , 0.10 |
| South Atl | 0.05 | 0.01 , 0.09 | 0.14 | 0.10 , 0.18 | 0.16 | 0.12 , 0.21 | 0.17 | 0.12 , 0.22 | 0.15 | 0.11 , 0.19 |
| East South Cent | -0.02 | -0.06 , 0.01 | 0.03 | -0.01 , 0.06 | 0.05 | 0.00 , 0.10 | 0.06 | 0.01 , 0.11 | 0.07 | 0.03 , 0.11 |
| West South Cent | -0.08 | -0.12 , -0.04 | -0.11 | -0.15 , -0.07 | -0.10 | -0.16 , -0.04 | -0.10 | -0.15 , -0.04 | -0.10 | -0.14 , -0.05 |
| Mountain | -0.18 | -0.22 , -0.14 | -0.23 | -0.26 , -0.19 | -0.24 | -0.28 , -0.19 | -0.23 | -0.28 , -0.18 | -0.21 | -0.25 , -0.17 |
| Pacific | -0.24 | -0.30 , -0.18 | -0.04 | -0.11 , 0.03 | 0.09 | 0.02 , 0.16 | 0.10 | 0.03 , 0.16 | 0.16 | 0.10 , 0.21 |
| hgc: ls8th | | | 0.08 | 0.03 , 0.14 | -0.04 | -0.11 , 0.03 | -0.05 | -0.12 , 0.02 | -0.04 | -0.10 , 0.02 |
| hgc: somehs | | | 0.16 | 0.11 , 0.22 | -0.04 | -0.10 , 0.03 | -0.04 | -0.11 , 0.03 | -0.02 | -0.08 , 0.03 |
| hgc: hs | | | 0.25 | 0.19 , 0.30 | -0.03 | -0.10 , 0.04 | -0.04 | -0.11 , 0.03 | 0.01 | -0.05 , 0.06 |
| hgc: somecol | | | 0.26 | 0.21 , 0.32 | -0.05 | -0.12 , 0.02 | -0.06 | -0.13 , 0.02 | -0.00 | -0.07 , 0.06 |
| hgc: college | | | 0.28 | 0.23 , 0.34 | -0.08 | -0.15 , -0.01 | -0.09 | -0.16 , -0.01 | -0.01 | -0.07 , 0.06 |
| hgc: gradschl | | | 0.32 | 0.26 , 0.37 | -0.03 | -0.11 , 0.05 | -0.04 | -0.12 , 0.04 | 0.03 | -0.03 , 0.10 |
| hgc: nr | | | 0.12 | 0.02 , 0.21 | 0.15 | 0.07 , 0.23 | 0.15 | 0.07 , 0.22 | 0.11 | 0.03 , 0.18 |
| inc 5k-10k | | | -0.14 | -0.18 , -0.10 | -0.19 | -0.25 , -0.14 | -0.19 | -0.24 , -0.13 | -0.16 | -0.20 , -0.11 |
| inc 10k-15k | | | 0.10 | 0.06 , 0.13 | -0.01 | -0.06 , 0.04 | -0.01 | -0.06 , 0.05 | 0.02 | -0.02 , 0.06 |
| inc 15k-20k | | | 0.16 | 0.12 , 0.19 | 0.03 | -0.03 , 0.08 | 0.03 | -0.03 , 0.08 | 0.06 | 0.02 , 0.10 |
| inc 20k-25k | | | 0.19 | 0.15 , 0.23 | 0.07 | 0.02 , 0.13 | 0.07 | 0.02 , 0.13 | 0.08 | 0.04 , 0.13 |
| inc 25k-30k | | | 0.24 | 0.20 , 0.28 | 0.11 | 0.06 , 0.17 | 0.11 | 0.06 , 0.17 | 0.12 | 0.08 , 0.17 |
| inc 30k-35k | | | 0.24 | 0.20 , 0.29 | 0.10 | 0.05 , 0.16 | 0.11 | 0.05 , 0.17 | 0.12 | 0.08 , 0.17 |
| inc 35k-40k | | | 0.25 | 0.21 , 0.30 | 0.09 | 0.03 , 0.15 | 0.09 | 0.04 , 0.16 | 0.12 | 0.07 , 0.17 |
| inc 40k-45k | | | 0.30 | 0.26 , 0.35 | 0.14 | 0.08 , 0.20 | 0.14 | 0.08 , 0.20 | 0.16 | 0.11 , 0.22 |
| inc 45k-50k | | | 0.28 | 0.24 , 0.33 | 0.15 | 0.09 , 0.22 | 0.15 | 0.09 , 0.22 | 0.16 | 0.11 , 0.21 |
| inc 50plus | | | 0.33 | 0.29 , 0.37 | 0.17 | 0.12 , 0.23 | 0.18 | 0.13 , 0.24 | 0.19 | 0.14 , 0.24 |
| risk tolerance | | | | | | | -0.02 | -0.05 , -0.00 | -0.01 | -0.03 , 0.01 |
| cognition | | | | | 0.17 | 0.15 , 0.19 | 0.17 | 0.15 , 0.19 | 0.08 | 0.06 , 0.10 |
| finpln | | | | | 0.04 | 0.01 , 0.06 | 0.03 | 0.01 , 0.06 | 0.01 | -0.00 , 0.03 |
| praliv75 | | | | | -0.02 | -0.04 , 0.00 | -0.01 | -0.03 , 0.01 | 0.02 | 0.00 , 0.04 |
| black | | | | | | | | | -0.24 | -0.27 , -0.21 |
| Hispanic | | | | | | | | | -0.24 | -0.27 , -0.20 |
| married, male | | | | | | | | | 0.06 | 0.03 , 0.08 |
| age, female | | | | | | | | | 0.06 | 0.03 , 0.08 |
| married,female | | | | | | | | | 0.02 | -0.01 , 0.06 |
| $\rho$ | -0.92 | -0.96 , -0.87 | -0.90 | -0.93 , -0.86 | -0.95 | -0.96 , -0.92 | -0.94 | -0.96 , -0.91 | -0.95 | -0.96 , -0.93 |
| $\sigma_{22}$ | 0.21 | 0.15 , 0.29 | 0.28 | 0.21 , 0.35 | 0.30 | 0.23 , 0.38 | 0.28 | 0.22 , 0.36 | 0.31 | 0.23 , 0.39 |

* Note: The omitted categories for the dummy variables are non-response for census divisions, zero schooling for education and less than 5 thousand dollars for income. All effects are computed for a median individual. For example, the marginal effect of risk tolerance is computed taking account of both the linear term and the interaction with $\sigma^2_{s_i}$, holding $\sigma^2_{s_i}$ and all other variables fixed at their medians.

very different from the results reported in this paper.

In fact, we have also re-estimated our most general model with the covariance parameter $\sigma_{12}$ set to zero. The fit of this restricted model to the data was very similar to that of the unrestricted model, and the posterior distribution of the marginal effect of $E_i^*$ on the Medigap coverage probability was similar as well. This can be seen by comparing panels (e) and (f) of Figure 4. Note that the mean effect increases slightly from 0.055 to 0.077.

## 5.3 The Moral Hazard Effect

In this section we discuss our inferences about moral hazard. In Table 3 row 4 we present posterior means of the type-specific moral hazard effects of Medigap insurance on health care expenditure ($\gamma_{2j}$). The posterior mean of $\gamma_{2j}$ is positive for all types. Type 1, the healthiest type, has the smallest insurance effect of \$1,200, and type 4, an unhealthy type, has the largest effect of \$2,145.[44] The posterior mean of the moral hazard effect for type 5, the least healthy type, is \$1,640, but the posterior is quite diffuse.

Thus, in absolute terms the, moral hazard effect tends to be smaller for more healthy types. Interestingly, however, the moral hazard effect is much larger as a *proportion* of health care expenditure for more healthy types. For example, the individuals of type 1 who have Medigap insurance spend about 215% more than their counterparts with no Medigap, while individuals of type 5 who have Medigap spend only about 2.5% more.

The moral hazard effect of 215% for type 1 might seem very large, but note that this does not correspond to a large absolute expenditure increase (i.e., type 1 has average spending of \$600 when uninsured and \$1,810 when insured). Also note that for most individuals there is considerable posterior uncertainty about their type. In the data, low expenditure

[44] Analogously, Munkin and Trivedi (2010) find that the size of the moral hazard effect is higher for the high-expenditure latent type than for the low-expenditure type in their study of supplemental drug insurance. Of course, since Medigap plans may cover other aspects of costs besides drugs (e.g., co-pays), it is not necessarily the case that these patterns would be the same in both markets.

individuals have high posterior probabilities of being types 1-2 and low posterior probabilities of being types 3-5, while the opposite is true for high expenditure individuals. When uncertainty about type is taken into account, estimates of the individual-level moral hazard effect are averages over type specific effects. For example, for individuals whose posterior type probability is highest for type 1, the average moral hazard effect is \$1,514, which is a 44% increase over their average expected expenditure in the Medicare only state (\$3,450). For people whose posterior modal type is 5, the moral hazard effect is equal to \$1,827, which is 5.7% of their average expected expenditure in the Medicare only state (\$32,000).

Our results suggest that the price elasticity of health care demand decreases as health status deteriorates. This seems intuitive. For instance, much of the health expenditure for healthy low expenditure individuals may go towards treatment of minor ailments - treatment that one may fairly easily forgo due to cost. In contrast, expenditures for unhealthy individuals are presumably more often for essential treatment of serious illness.

Next, we consider the mean of the moral hazard effect in the whole population. The moral hazard effect of Medigap for a person with observable characteristics $\mathbf{x}_i$ can be computed as: $E(MH_i|\mathbf{x}_i,\boldsymbol{\theta}) = \sum_{j=1}^{m} \gamma_{2j} \cdot P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)$.[45] The posterior mean (over the posterior of parameters) of $E(MH_i|\mathbf{x}_i,\boldsymbol{\theta})$ can be approximated as $10^{-3}\sum_{k=1}^{10^{-3}} \gamma^k_{2s^k_i}$, where $s^k_i$ are simulated as discussed in section 5.1. This posterior mean varies between \$1,267 and

[45] Alternatively, we could define the moral hazard effect as the difference between the expected actual expenditure $E_i$ of an individual with and without Medigap, i.e

$$E(MH_i|\mathbf{x}_i,\boldsymbol{\theta}) = \sum_{j=1}^{m}(E^1(E_i|\text{type}_i = j, \mathbf{xe}_i, \boldsymbol{\theta}) - E^0(E_i|\text{type}_i = j, \mathbf{xe}_i, \boldsymbol{\theta})) \cdot P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)),$$

where

$$E^I(E_i|\text{type}_i = j, \mathbf{xe}_i, \boldsymbol{\theta}) = \Phi(\frac{\mathbf{xe}'_i\boldsymbol{\beta}_j + \gamma_{1j}Y_i + \gamma_{2j}\cdot I}{(\sigma^2_j + \sigma_{22})^{0.5}})[\mathbf{xe}'_i\boldsymbol{\beta}_j + \gamma_{1j}Y_i + \gamma_{2j}\cdot I + (\sigma^2_j + \sigma_{22})^{0.5} \cdot \frac{\phi(\frac{\mathbf{xe}'_i\boldsymbol{\beta}_j + \gamma_{1j}Y_i + \gamma_{2j}\cdot I}{(\sigma^2_j + \sigma_{22})^{0.5}})}{\Phi(\frac{\mathbf{xe}'_i\boldsymbol{\beta}_j + \gamma_{1j}Y_i + \gamma_{2j}\cdot I}{(\sigma^2_j + \sigma_{22})^{0.5}})}]$$

for $I = 0, 1$. This last expression is due to the Tobit specification for the distribution of actual expenditure. The two definitions of the moral hazard effect produce similar results.

\$2,151 in our sample. The sample average of the moral hazard effect is \$1,615. This is a 24% increase from the average expenditure risk in the Medicare only state (\$6,789). This result is reasonably comparable to the effect of insurance found in the RAND Health Insurance Experiment. For example, Manning et al. (1987) report that a decrease in the co-insurance rate from 25% to 0 increased total health care expenditure by 23%. Such a drop in co-pays is similar to the consequences of adopting many typical Medigap plans that cover co-pays.

In contrast to the selection effect, restricting the covariance between unobservables $\varepsilon_1$ and $\varepsilon_2$, $\sigma_{12}$, to zero has a large impact on inferences about the moral hazard effect. In such a specification the posterior mean of the moral hazard effect $E(MH_i|\mathbf{x}_i, \boldsymbol{\theta})$ is only \$687, compared to \$1,615 in the full model. This drop in magnitude is not surprising given the large negative correlation between the unobervables shown in Table 4 (i.e., advantageous selection into Medigap). Once we control for this advantageous selection on unobservables, the moral hazard effect of insurance is revealed to be larger.

It is interesting to evaluate the potential effects on aggregate health expenditure of a policy which expands Medigap coverage by making it more affordable. Thus, we simulate a situation where the price of Medigap drops sufficiently so that Medigap coverage increases by 10% (or 5 percentage points). According to the estimated price elasticity of demand for health insurance in Buchmueller (2006), this would require approximately a \$25 drop in Medigap premiums.[46] The simulations suggest that the individuals who are attracted to Medigap insurance by this policy would on average spend \$8,300 when Medigap-insured. This compares to average expenditure of \$8,100 for those who were already covered before the policy was implemented. The newly insured spend (slightly) more because they have

[46]Buchmueller (2006) estimates that a \$5 increase in an insurance premium would decrease a plan's enrolment by 2% in his sample of retirees over the age of 65. His study relies on changes in demand for different plans caused by an exogenous change in the retiree health insurance contributions policy of a single employer. Another study that uses a natural experiment to estimate the elasticity of health insurance demand is Gruber and Washington (2005). We use the estimate of Buchmueller (2006) because the demographic characteristics of individuals in his sample are similar to those in our data.

higher expenditure risk - their average expenditure risk in the Medicare only state ($E_i^*$) is \$6,700, compared to \$6,400 for individuals who had Medigap coverage before the policy was implemented.[47] Thus, expanding Medigap coverage results in a somewhat higher cost per insured person mainly due to advantageous selection. But the increase in the average health expenditure of all insured individuals is very small (from \$8,060 to \$8,090). The policy increases per capita expenditure from \$7,650 to \$7,730.[48]

In contrast, expanding Medigap coverage universally would have a large effect on expenditure, increasing per capita expenditure from \$7,650 to \$8,390. This increase is primarily due to the moral hazard effect: the newly insured (who make up about 50% of the sample) increase their spending by \$1,500 dollars on average, which increases average expenditure by about \$740. Of course, the welfare consequences of expanding Medigap coverage cannot be evaluated using our model, but this is an important issue for future research.

Finally, it is interesting to see how much different health types contribute to the aggregate increase in spending resulting from universal Medigap coverage. Returning to Table 3, we see that the healthiest individuals (type 1) account for about 29% of the increase in spending. This is computed as the ratio of the type-specific moral hazard effect to the average moral hazard effect, weighted by the type probability: $\frac{1210 \cdot 0.39}{1615} = 0.29$. Similarly, the contributions of individuals of types 2-5 to the aggregate increase in spending are 37%, 18%, 12% and 3%, respectively. Thus, the two healthiest types, who make up 71% of the population, but who account for only 18% of total spending under the status quo, account for 66% of the total spending increase induced by universal coverage.

[47]The newly insured have a slightly lower moral hazard effect than those previously insured (\$1,600 vs \$1,614), and their household incomes are on average lower, but these effects are minor.

[48]These calculations are based on artificial data samples of $E_i^k$, $I_i^k$ and $E_i^{*k}$ simulated as discussed in section 5.1 for two situations: (i) before the policy (the original posterior distribution of the parameters is used); (ii) after the policy (the intercept term in the insurance equation is increased by a constant to achieve the average Medigap coverage of 0.55, all random terms (e.g. $\varepsilon_{i1}, \varepsilon_{i2}, s_i, \eta_i$) are the same as in (i)). The average expenditure risk and moral hazard effects of the two groups, (i) with Medigap before the policy and (ii) with no Medigap before the policy but with Medigap after the policy, are computed as discussed above.

## 5.4 Marginal Effects of Covariates on Aspects of Expenditure

Table 5 presents posterior means and the $5^{th}$ and $95^{th}$ percentiles of the posterior distributions of marginal effects of covariates on the following covariate-dependent functions of interest:

1. The expected expenditure risk (columns 1-3):[49]

$$E(E_i^*|\mathbf{xe}_i, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m) = \sum_{j=1}^{m} E(E_i^*|\mathbf{xe}_i, \text{type}_i = j, \boldsymbol{\beta}_j) \cdot P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m), \tag{21}$$

   where $E(E_i^*|\mathbf{xe}_i, \text{type}_i = j, \boldsymbol{\beta}_j) = \mathbf{xe}_i'\boldsymbol{\beta}_j$ and $P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)$ is given in equation (6);

2. The moral hazard effect of Medigap insurance on expenditure (columns 4-6):

$$E(MH_i|\mathbf{xw}_i, \gamma_{21}, ..., \gamma_{2m}, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m) = \sum_{j=1}^{m} \gamma_{2j} \cdot P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m); \tag{22}$$

3. The unconditional standard deviation of the forecast error (columns 7-9):

$$SD(\eta_i|\mathbf{xw}_i, \sigma_1^2, ..., \sigma_m^2, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m) = \left(\sum_{j=1}^{m} \sigma_j^2 \cdot P(\text{type}_i = j|\mathbf{xw}_i, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)\right)^{\frac{1}{2}}. \tag{23}$$

The marginal effects in Table 5 are computed for a median individual (i.e. an individual whose covariates $\mathbf{xe}_i$ are set to the sample median level), and are measured in thousands of dollars. For continuous covariates the effects are for a one standard deviation increase in the covariate from its sample median level. For categorical covariates the effect is from moving to the next category. The 5th and 95th percentiles reflect the uncertainty with respect to the posterior distribution of parameters, i.e. the effects of the covariates on the expressions

[49] Because in our analysis $\mathbf{xw}_i$ is a subset of $\mathbf{xe}_i$, as discussed in section (4.1), conditioning on $\mathbf{xe}_i$ is equivalent to conditioning on both $\mathbf{xw}_i$ and $\mathbf{xe}_i$

(21)-(23) were evaluated for 1000 draws from the posterior distribution of parameters, and the mean and 5th and 95th percentiles of these 1000 values are reported in Table 5.

The expenditure risk in the Medicare only state $E(E_i^*|\mathbf{xe}_i, \boldsymbol{\beta}_1, ..., \boldsymbol{\beta}_m, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)$ is lower for females, blacks and Hispanics. On average females are expected to spend \$370 less than males, while blacks and Hispanics spend \$740 and \$830 less than other ethnic groups, respectively.[50] Expenditure risk is insensitive to age, conditional on detailed measures of health status, and is lower for married individuals. Unhealthy individuals are expected to spend more than their more healthy counterparts - a one standard deviation increase in the unhealthy factor 2 raises expenditure risk by \$3,820, while a one standard deviation increase in the healthy factor 3 decreases risk by \$1,860.

Expenditure risk also varies by census division. In particular, residents of the New England region have the highest expenditure risk, while residents of the Pacific region have the lowest expenditure risk (conditional on other variables). This finding of large unexplained regional differences is consistent with the extensive literature on regional variation in health care costs and practices (see Zuckerman et al (2010), Fisher et al (2003), Welch et al (1993)). In fact, we contribute to this literature by showing that regional differences persist even given our more extensive controls for health status, and our SMT model of expenditure, which fits the shape of the health care cost distribution far better than models used in prior work.

The moral hazard effect of Medigap is lower for individuals in better health, and for blacks and Hispanics. Individuals living in the Pacific census division have the smallest moral hazard effect, while individuals living in New England have the largest. The standard deviation of the forecast error $SD(\eta_i|\mathbf{xw}_i, \sigma_1^2, ..., \sigma_m^2, \boldsymbol{\delta}_1, ..., \boldsymbol{\delta}_m)$ is lower for females and higher for less healthy individuals. The variance of the forecast error is highest for those living in

[50]It is notable that race and marital status were not significant predictors of health expenditure in the OLS regression in Table 2, but they are significant in the full model. This may be due to the more flexible functional form for the conditional expectation of expenditure in the full model, compared to OLS. It may also be due to bias in the OLS Medigap coefficient due to endogeneity of insurance.

New England and lowest for those living in the Pacific and West South Central census divisions. Thus, there are unexplained regional differences in the variance of expenditure, not just the level.

# 6 Conclusion

This paper studies selection and moral hazard in the US Medigap health insurance market. We develop a unified econometric model of insurance demand and health care expenditure, using data from the HRS and MCBS. We extend earlier work by Fang, Keane and Silverman (2008) in several ways. Most importantly: (1) we account for endogeneity of insurance coverage when we estimate the extent of selection, and (2) we estimate the extent of moral hazard. As instruments for insurance coverage we use a set of behavioral variables (e.g., cognitive ability) that FKS found were important determinants of demand for insurance but that can be plausibly excluded from the expenditure equation (conditional on health status). Notably, this paper is the first to estimate selection and moral hazard effects jointly in the Medigap market while accounting for endogeneity of insurance choice.

We also incorporate two technical innovations. First, to capture the complex shape of the health care expenditure distribution, we employ a smooth mixture of Tobits model that generalizes the smoothly mixing regressions (SMR) framework of Geweke and Keane (2007). Second, because neither the HRS nor the MCBS contains all the data necessary for our analysis, we develop a novel MCMC data augmentation algorithm to combine information from the two sources and obtain the posterior distribution of our model parameters.

Our results imply that the moral hazard (or price) effect is substantial. We find that individuals with Medigap insurance spend about $1,615 more on health care (on average) than similar individuals without Medigap ($6,789 vs $8,404). This is a 24% increase. As a result of the moral hazard effect, we find that a policy of expanding Medigap coverage to all

Table 5: Marginal effects of individual characteristics on selected functions of interest

| Variable | $E(E_i^*\|\mathbf{xe}_i,\boldsymbol{\beta}_1,...,\boldsymbol{\beta}_m,\boldsymbol{\delta}_1,...,\boldsymbol{\delta}_m)$ | | $E(MH_i\|\mathbf{xw}_i,\gamma_{21},...,\gamma_{2m},\boldsymbol{\delta}_1,...,\boldsymbol{\delta}_m)$ | | $SD(\eta_i\|\mathbf{xw}_i,\sigma_1^2,...,\sigma_m^2,\boldsymbol{\delta}_1,...,\boldsymbol{\delta}_m)$ | |
|---|---|---|---|---|---|---|
| | Post. mean | $(5^{th},95^{th})$ pct | Post. mean | $(5^{th},95^{th})$ pct | Post. mean | $(5^{th},95^{th})$ pct |
| | 1 | 2 | 3 | 4 | 5 | 6 |
| Female | -0.37 | -0.87, 0.09 | -0.00 | -0.07, 0.06 | -0.77 | -1.26, -0.32 |
| Age | -0.08 | -0.32, 0.13 | 0.01 | -0.01, 0.03 | -0.10 | -0.33, 0.16 |
| New Eng | 0.31 | -0.43, 1.02 | 0.06 | -0.04, 0.17 | 0.76 | -0.28, 1.69 |
| Mid Atl | -1.08 | -1.79, -0.35 | -0.02 | -0.10, 0.05 | -0.25 | -1.15, 0.55 |
| East North Cent | -1.17 | -1.97, -0.39 | -0.05 | -0.13, 0.04 | -0.30 | -1.31, 0.72 |
| West North Cent | -0.86 | -1.59, -0.19 | 0.00 | -0.07, 0.07 | -0.24 | -1.12, 0.60 |
| South Atl | -1.59 | -2.40, -0.81 | 0.01 | -0.09, 0.11 | -0.68 | -1.68, 0.17 |
| East South Cent | -0.98 | -1.75, -0.27 | -0.01 | -0.09, 0.06 | -0.05 | -0.97, 0.83 |
| West South Cent | -1.38 | -2.14, -0.65 | -0.05 | -0.14, 0.05 | -0.94 | -1.84, -0.15 |
| Mountain | -1.00 | -1.76, -0.36 | -0.01 | -0.09, 0.06 | -0.30 | -1.18, 0.51 |
| Pacific | -3.31 | -4.19, -2.44 | -0.14 | -0.32, 0.06 | -2.12 | -3.24, -0.86 |
| Health factor 1 | 3.82 | 3.04, 4.81 | 0.11 | -0.08, 0.31 | 1.84 | 1.20, 2.64 |
| Health factor 2 | -1.86 | -2.18, -1.57 | -0.08 | -0.16, 0.01 | -1.01 | -1.32, -0.73 |
| Health factor 3 | 0.87 | 0.60, 1.17 | 0.02 | -0.03, 0.08 | 0.41 | 0.06, 0.83 |
| Health factor 4 | 0.94 | 0.76, 1.14 | 0.03 | -0.00, 0.07 | 0.11 | -0.13, 0.37 |
| Health factor 5 | -0.36 | -0.67, -0.06 | -0.01 | -0.03, 0.02 | -0.33 | -0.64, -0.00 |
| Health factor 6 | -0.70 | -0.88, -0.53 | -0.03 | -0.07, 0.01 | -0.47 | -0.69, -0.23 |
| Health factor 7 | 1.05 | 0.65, 1.49 | 0.07 | 0.01, 0.14 | 0.44 | 0.10, 0.85 |
| Health factor 8 | 0.04 | -0.23, 0.28 | 0.01 | -0.01, 0.04 | -0.09 | -0.41, 0.30 |
| Health factor 9 | 0.46 | 0.26, 0.71 | 0.02 | -0.01, 0.04 | 0.04 | -0.22, 0.35 |
| Health factor 10 | 0.16 | -0.05, 0.42 | 0.01 | -0.02, 0.04 | 0.21 | -0.09, 0.58 |
| Black | -0.74 | -1.29, -0.15 | -0.09 | -0.18, 0.00 | -0.29 | -0.89, 0.54 |
| Hispanic | -0.83 | -1.37, -0.34 | -0.06 | -0.14, 0.01 | -0.75 | -1.31, -0.21 |
| Married | -0.45 | -0.84, -0.12 | 0.01 | -0.04, 0.05 | -0.46 | -0.86, -0.05 |
| Year | -1.17 | -1.47, -0.88 | -0.03 | -0.08, 0.01 | -0.60 | -0.94, -0.26 |

* Note: The marginal effects are measured in thousand dollars and are evaluated for the median individual. For continuous covariates the effects are for a one standard deviation increase in the covariate from its sample median level. The omitted category for census division is the group with missing census division information.

senior citizens would increase per capita health care expenditure by about $740.

An important contribution of this paper is the rich structure of heterogeneity that we build into the health care expenditure function by using the SMR technique. This lets the moral hazard effect of Medigap vary in a very flexible way across types of people. Thus, we can go beyond the mean effect and estimate the whole distribution of moral hazard effects across different types. In particular, we find that the price elasticity of demand for health care is much greater for people in better health.[51] The healthiest 70% of the senior population account for only 18% of costs in the current context, but they would account for 2/3 of the increase in health care spending that would accompany a universal extension of Medigap coverage.

We also find evidence of advantageous selection in the Medigap market. Conditional on characteristics that insurance companies can legally use for pricing, healthier individuals are more likely to buy Medigap. Specifically, a one standard deviation increase in expenditure risk *decreases* the probability of Medigap coverage by 2.7 percentage points. This contradicts the prediction of classic models of adverse selection with a single dimension of private information (risk type).

However, when we condition on additional types of private information that cannot be used in pricing (cognitive ability, risk attitudes, financial planning horizon, longevity expectations, education, race and marital status) we do find adverse selection on health status. But this effect is modest: a one standard deviation increase in expenditure risk increases the probability of insurance coverage by 5.5 percentage points, *ceteris paribus*.

These findings are qualitatively and quantitatively similar to the results of Fang, Keane and Silverman (2008). We do find that insurance status is correlated with unobserved aspects

[51] Intuitively, this is because healthy people tend to have relatively minor ailments where treatment can be easily forgone due to the price of health care, while people in poor health are more likely to have serious ailments where treatment cannot be forgone without severe consequences. Of course, we expect from basic demand theory that lowering the cost of services would induce a disproportionate increase in demand for treatment in "marginal" cases where treatment is less critical.

of health, a possibility that FKS ignore. Nevertheless, we find that the FKS results are robust to endogenizing insurance. Our explanation is that the observed controls for health (here and in FKS) are so extensive that the unexplained part of risk type is not quantitatively important for determining the nature of selection. On the other hand, we find that failure to account for endogeneity of insurance status leads to a substantial underestimate of the moral hazard effect (i.e., \$687 in the exogenous insurance model vs. \$1,615 when we allow for correlated unobservables between the insurance and expenditure equations).

The estimated effects of covariates on demand for insurance and expenditure risk are in some cases of considerable interest. We find that our measure of cognitive ability has a substantial impact on demand: a one standard deviation increase in measured cognitive ability increases the probability of having Medigap by 0.08 for a median individual. Somewhat surprisingly, the effect of financial risk tolerance on demand for insurance, while in the expected direction, is rather modest. As a result, heterogeneity in private information about risk aversion does little to help explain advantageous selection.

We also find that, *ceteris paribus*, blacks and Hispanics are much less likely to purchase Medigap insurance than whites, and have lower health care expenditure (conditional on health). As a result, race is an important source of adverse selection. An important avenue for future research is to determine why blacks and Hispanics have roughly half the rate of Medigap coverage of whites, even conditional on extensive controls for health, demographics and psychometric measures.

Finally, an important limitation of our analysis is that the welfare consequences of expanding Medigap coverage cannot be evaluated using our model. This is obviously an important issue for future research.

# References

[1] Arrow, K.J., 1963. Uncertainty and the Welfare Economics of Medical Care. American Economic Review 53 (5), 941-973.

[2] Bajari, P., Han Hong, Ahmed Khwaja. 2011. A Semiparametric Analysis of Adverse Selection and Moral Hazard in Health Insurance Contracts. Working Paper.

[3] Bajari, Patrick, Hong, H., Khwaja, A., Marsh, C., 2011. Moral Hazard, Adverse Selection and Health Expenditures: A Semiparametric Analysis. Working Paper.

[4] Blough, D.K., Madden C.W., Hornbrook M.C., 1999. Modeling Risk Using Generalized Linear Models. Journal of Health Economics 18 (2), 153-171.

[5] Buchmueller, T., 2006. Price and the Health Plan Choice of Retirees. Journal of Health Economics 25(1), 81-101.

[6] Beeuwkes Buntin, M., Zaslavsky, A.M., 2004. Too Much Ado About Two-Part Models and Transformations? Comparing Methods of Modeling Medicare Expenditures. Journal of Health Economics 23(3), 525-542.

[7] Cardon, J.H., Handel, I., 2001. Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey. The RAND Journal of Economics 32 (3), 408-427.

[8] Cawley, J., Philipson, T., 1999. An Empirical Examination of Information Barriers to Trade in Insurance. American Economic Review 89 (4), 827-46.

[9] Celeux, G., Hurn, M., Robert, C.P., 2000. Computational and Inferential Difficulties with Mixture Posterior Distributions. Journal of the American Statistical Association 95 (451), 957-970.

[10] Chib, S., 1992. Bayes inference in the Tobit censored regression model. Journal of Econometrics 51 (1-2), 79-99.

[11] Cohen, A., 2005. Asymmetric Infromation and Learning: Evidence from the Automobile Insurance Market. Review of Economics and Statistics 87 (2), 197-207.

[12] Cohen, A., Siegelman, P., 2010. Testing for Adverse Selection in Insurance Markets. The Journal of Risk and Insurance 77 (1), 39-84.

[13] Chiappori, P.-A., Salanie B., 2000. Testing for Asymmetric Information in Insurance Markets. Journal of Political Economy 108 (1), 56-78.

[14] Chiappori, P.-A., Durand F., Geoffard P.Y., 1998. Moral Hazard and the Demand for Physician Services: First Lessons from a French Natural Experiment. European Economic Review 42 (3-5), 499-511.

[15] Dardanoni, V., and Li Donni P., 2012. Incentive and selection effects of Medigap insurance on inpatient care. Journal of Health Economics 31(3), 457-470.

[16] Deb, P., Munkin M., Trivedi P.K., 2006. Bayesian Analysis of the Two-Part Model with Endogeneity: Application to Health Care Expenditure. Journal of Applied Econometrics 21 (7), 1081-1099.

[17] Deb, P., Manning, W., Norton E., 2010. Preconference Course: Modeling Health Care Costs and Counts. Lecture Slides Prepared for the ASHE-Cornell University Conference, 2010.

[18] De Meza, D., Webb, D.C., 2001. Advantageous Selection in Insurance Markets. RAND Journal of Economics 32(2), 249-62.

[19] Ettner, S., 1997. Adverse Selection and the Purchase of Medigap Insurance by the Elderly. Journal of Health Economics 16(5), 543-562.

[20] Fang, H., Keane M.P., Silverman D., 2008. Sources of Advanatageous Selection: Evidence from the Medigap Insurance Market. Journal of Political Economy 116 (2), 303-349.

[21] Fang, H., Nicholas L. Silverman D., 2010. Cognitive Ability and Retiree Health Care Expenditure. Michigan Retirement Research Center Working Paper 230.

[22] Finkelstein, A., McGarry K., 2006. Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market. American Economic Review 96 (4), 938-958.

[23] Fisher S. E., Wennberg D.E., Stukel T.A. , Gottlieb D.J., Lucas F.L., Pinder A.L., 2003. The Implications of Regional Variations in Medicare Spending. Part 1: The Content, Quality, and Accessibility of Care. Annals of Internal Medicine 138(4), 273-E311.

[24] Fruhwirth-Schnatter, S., (2001). Markov Chain Monte Carlo estimation of classical and dynamic switching and mixture models. Journal of the American Statistical Association 96 (453), 194-209.

[25] Gelman, A., Carlin J.B., Stern H.S., Rubin D.B., 1995. Bayesian Data Analysis. London: Chapman and Hall.

[26] Geweke, J., 1992. Evaluating the accuracy of sampling-based approaches to the calculation of posterior moments. In: Berger, J.O., Bernardo, J.M., Dawid, A.P., Smith, A.F.M. (Eds.), Bayesian Statistics, vol. 4. Oxford University Press, Oxford, pp. 169-194.

[27] Geweke, J., 2004. Getting it Right: Joint Distribution Tests of Posterior Simulators. Journal of the American Statistical Association 99 (467), 799-804.

[28] Geweke, J., 2005. Contemporary Bayesian Econometrics and Statistics. John Wiley & Sons, Inc., Hoboken, New Jersey.

[29] Geweke, J., 2006. Interpretation and Inference in Mixture Models: Simple MCMC Works. Computational Statistics & Data Analysis 51 (7), 3529-3550.

[30] Geweke, J., Keane, M.P., 2007. Smoothly Mixing Regressions. Journal of Econometrics 138 (1), 257-290.

[31] Gilleskie, D.B., Mroz T.A., 2004. A Flexible Approach for Estimating the Effects of Covariates on Health Expenditures. Journal of Health Economics 23 (2), 391-418.

[32] Goldberger, A.S., 1991. A Course in Econometrics. Harvard University Press; Cambridge, Massachusetts; London, England.

[33] Gruber, J., Washingtion, E., 2005. Subsidies to Employee Health Insurance Premiums and the Health Insurance Market. Journal of Health Economics 24(2), 253-176.

[34] Gulila Z., McCulloch R.E. Rossi P.E., 2006. A Direct Approach to Data Fusion. Journal of Marketing Research 43 (1): 73-83.

[35] Hurd, M.D., McGarry K., 1997. Medical Insurance and the Use of Health Care Services by the Elderly. Journal of Health Economics 16(2), 129-154.

[36] Kaiser Family Foundation. 2005. Medicare Chart Book. Washington, DC: Kaiser Family Found.

[37] Keane, M.P, Stavrunova, O., 2011. A Smooth Mixture of Tobits Model for Health Care Expenditure. Health Economics 20 (9), 1126-1153.

[38] Khwaja, A., 2001. Health Insurance, Habits and Health Outcomes: A Dynamic Stochastic Model of Investment in Health. PhD dissertation, University of Minnesota.

[39] Kimball, M.S., Sahm, C.R., Shapiro M.D., 2008. Imputing Risk Tolerance From Survey Responses. Journal of American Statistical Association 103(483), 1028-1038.

[40] Koopmans T.C., Rubin H., Leipnik, R.B., 1950. Measuring The Equation Systems of Dynamic Economics. In Statistical Inference in Dynamic Economics Models. Cowles Commision Monograph 10, New York, John Wiley & Sons.

[41] Manning, W.G., Newhouse J.P.,Duan N., Keeler E.B., Leibowitz A., 1987. Health Insurance and the Demand for Medical Care: Evidence from a Randomized Experiment. The American Economic Review 77 (3), 251-277.

[42] Manning, W.G., 1998. The Lagged Dependent Variables, Heteroskedasticity, and the Retransformation Problem. Journal of Health Economics 17 (3), 283-295.

[43] Manning, W.G., Mullahy J., 2001. Estimating log models: to transform or not to transform. Journal of Health Economics 20(4), 461-494.

[44] Manning, W.G., Basu A., Mullahy J., 2005. Generalized modeling approaches to risk adjustment of skewed outcome data. Journal of Health Economics 24 (3), 465-488.

[45] Mullahy, J., 1998. Much Ado About Two: Reconsidering Transformation and the Two-Part Model in Health Econometrics. Journal of Health Economics 17 (3), 247-281.

[46] Munkin, M., Trivedi P.K., 2008. Bayesian analysis of the ordered probit model with endogenous selection. Journal of Econometrics 143 (2), 334-348.

[47] Munkin, M., Trivedi P.K., 2010. Health Economics 19 (9), 1093-1108.

[48] O'Grady K.F., Manning W.G., Newhouse J.P., Brook R.H., 1985. The impact of cost sharing on emergency department use. The New England Journal of Medicine 313(8), 484-490.

[49] Pauly, M.V., 1968. The Economics of Moral Hazard: Comment. American Economic Review 58 (3), 531-537.

[50] Rothschild, M., Stiglitz J.E., 1976. Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information. Quarterly Journal of Economics 90 (4), 629-49.

[51] Rubin, D.B., 1986. Statistical Matching Using File Concatenation with Adjusted Weights and Multiple Imputations. Journal of Business and Economic Statistics 4(1), 87-94.

[52] Welch W.P., Miller M.E., Welch H.G., Fisher E.S., Wennberg J.E., 1993. Geographic variation in expenditures for physicians' services in the United States. New England Journal of Medicine. 328 (9), 621-7.

[53] Wolfe, J.R., Goddeeris J.H., 1991. Adverse Selection, Moral Hazard, and Wealth Effects in the Medigap Insurance Market. Journal of Health Economics 10 (4), 433-459.

[54] Wooldridge, J.M., 2002. Econometric Analysis of Cross Section and Panel Data. The MIT Press. Cambridge, Massachusetts.

[55] Zuckerman, S., Waidmann T., Berenson R., Hadley J., 2010. Clarifying Sources of Geographic Differences in Medicare Spending. New England Journal of Medicine. 363, 54-62.

# Appendix

## A-1. Prior Distributions

We specify the following prior distributions:

1. $\alpha_0 \sim N(\underline{\alpha}_0, \underline{h}_{\alpha_0}^{-1})$, where $\underline{\alpha}_0 = 0$ and the prior variance $\underline{h}_{\alpha_0}^{-1} = 0.4$. This specification implies that for an individual whose probability of Medigap coverage is equal to 0.5, the effect of a one sample standard deviation increase in expected expenditure (in Medicare only state) on this probability is centered at zero, while the 1st, 25th, 75th and 99th percentiles of this effect are -0.31, -0.10, 0.10, 0.31, respectively.[52] That is, this prior reflects the belief that the effect of $E_i^*$ is not very large, but still places a non-negligible probability on the event that a one standard deviation increase in $E_i^*$ can change the Medigap coverage substantially (e.g., from 0.5 to 0.80 or 0.20).

2. $\boldsymbol{\alpha} \sim N\left(\underline{\boldsymbol{\alpha}}, \underline{\mathbf{H}}_{\alpha}^{-1}\right)$, where $\underline{\boldsymbol{\alpha}} = \mathbf{0}$ and the variance-covariance $\underline{\mathbf{H}}_{\alpha}^{-1}$ is a diagonal matrix which allows for reasonable prior uncertainty about the effects of the variables on the probability of Medigap coverage. In Table A-1 we present the diagonal elements of $\underline{\mathbf{H}}_{\alpha_1}^{-1}$ for the continuous variables, as well as the implied effects of a one sample standard deviation increase in these variables on the probability of Medigap coverage at the Medigap probability of 0.5.[53] The diagonal elements of $\underline{\mathbf{H}}_{\alpha_1}^{-1}$ for the intercept and for the indicator variables (i.e., census division, gender, education and income categories) are set to 1, so the prior distribution of the effect of increasing the indicator variables from 0 to 1 (evaluated at a Medigap probability of 0.5) are -0.49, -0.23, 0.23, 0.49 at 1st, 25th and 75th and 99th percentiles respectively.

3. $\boldsymbol{\beta}_j \sim N\left(\underline{\boldsymbol{\beta}}, \underline{\mathbf{H}}_{\beta}^{-1}\right)$ for $j = 1, \ldots, m$. We specify that $\underline{\boldsymbol{\beta}} = [\overline{E}, \mathbf{0}_{K_E-1}]'$, where $\overline{E}$ is the sample average of expenditure in the MCBS subsample and $K_E$ is the size of $\mathbf{xe}_i$. The precision matrix $\underline{\mathbf{H}}_{\beta} = 0.1 \cdot \sum_{i \in MCBS} \mathbf{xe}_i \cdot \mathbf{xe}_i' / (N^M \cdot Var(E))$, where $Var(E)$ is the sample variance of expenditure in the MCBS subsample. This prior specification is based on the prior for the normal linear regression model proposed in Geweke (2005), Chapter 5. It specifies considerable prior uncertainty about the effects of a one standard

[52] Because the expected expenditure in the Medicare only state $E_i^*$ is a latent variable, its distribution across individuals is unknown until the estimation is completed. To set the prior variance of $\alpha_0$ we approximate $E_i^*$ by the health expenditure risk computed using the FKS imputation methodology. In particular, we compute the expenditure risk as $\mathbf{xe}_i\mathbf{b}$, where $\mathbf{b}$ is a vector of least squares coefficients on health status characteristics $\mathbf{xe}_i$ from the regression of health care expenditure on $\mathbf{xe}_i$ and the Medigap insurance status in the MCBS subsample. The sample standard deviation of the imputed expenditure is equal to 0.59 for the expenditure measured in tens thousands of dollars.

[53] Similarly, the distribution of $\sigma_{s_i}^2$ across the individuals in the sample is not known until the estimation is completed. To set the prior variance of $\alpha_1$ we approximate $\sigma_{s_i}^2$ by the imputed health expenditure variance computed using the FKS imputation methodology. In particular, we impute the expenditure variance as $\mathbf{xe}_i\mathbf{v}$, where $\mathbf{v}$ is a vector of least squares coefficients on heath status characteristics $\mathbf{xe}_i$ from the regression of $(E_i - \mathbf{xe}_i\mathbf{b} - I_i \cdot b_I)^2$ on $\mathbf{xe}_i$ and the Medigap insurance status in the MCBS subsample. The sample standard deviation of the imputed variance of expenditure is equal to 2.3 for the expenditure measures in tens thousands of dollars.

Table A-1: Prior distribution of $\boldsymbol{\alpha}_1$

| Variable | Prior Variance | 1st, 25th, 75th, 99th percentiles of marginal effect |
|---|---|---|
| $\sigma^2_{s_i}$ | 0.03 | -0.32, -0.10, 0.10, 0.32 |
| $\sigma^2_{s_i} \cdot$ risktol* | 0.015 | -0.096, -0.03, 0.03, 0.096 |
| Age♯ | 1.5 | -0.36, -0.13, 0.13, 0.36 |
| Age$^2$ | 3 | |
| Age$^3$ | 3 | |
| risktol | 10 | -0.35, -0.12, 0.12, 0.35 |
| cogn | 3 | -0.39, -0.14, 0.14, 0.39 |
| finpln | 0.15 | -0.39, -0.14, 0.14, 0.39 |
| praliv75 | 0.35 | -0.39, -0.14, 0.14, 0.39 |

* This marginal effect corresponds to the change in the effect of $\sigma^2_{s_i}$ brought about by one HRS sample standard deviation change in risk tolerance.

♯ The marginal effect corresponds to the total effect when age changes by one sample standard deviation. Age, Age$^2$ and Age$^3$ interacted with gender indicator have the same prior variances.

deviation change in a covariate on the expenditure. In particular, the implied 1st and 99th percentiles of the effect on expenditure of a one standard deviation increase in any of the covariates is at least $\pm$ 18.5 thousand dollars, which is enough to take expenditure from the 50th to the 90th percentile of it's sample distribution.

4. $\boldsymbol{\gamma}_j \sim N(\underline{\gamma}, \underline{h}_\gamma^{-1})$, where $\underline{\gamma} = \mathbf{0}$ and $\underline{h}_\gamma$ is a diagonal matrix with the diagonal elements equal to 1.4 and 0.01 for $j = 1, ..., m$. This prior allows for substantial prior uncertainty about the effects of income and Medigap on health expenditure. The 1st and 99th percentiles of the prior effect of increasing income by one standard deviation is $\pm$98 thousand dollars. The prior effect of Medigap insurance is $\pm$232 thousand dollars.

5. $\underline{S} h_j \sim \chi^2(\underline{V})$, where $\underline{V} = 1$ and $\underline{S} = 0.59$ for $j = 1, ..., m$. This prior allows for substantial prior uncertainty about the type-specific variance of $\widehat{E}_i$. The interval constructed of the 1th and 99th percentiles of the prior distribution of $1/h_j$ is $[0.09, 3684.9]$, and it contains the variance of the observed expenditure (equal to 2.13 for expenditure measured in tens of thousands of dollars).

6. $\boldsymbol{\lambda}_k \sim N\left(\underline{\boldsymbol{\lambda}}_k, \underline{\mathbf{H}}_{\lambda_k}^{-1}\right)$ for $k = 1, ..., 4$, where $\underline{\boldsymbol{\lambda}}_k = [\bar{c}_k, \mathbf{0}_{K_c-1}]'$, $\bar{c}_k$ denotes sample average of $c_{ki}$ in the HRS subsample, and $K_c$ is the number of covariates in the vector $\mathbf{xc}_i$. The prior precision $\underline{\mathbf{H}}_{\lambda_k} = \sum_{i \in HRS} \mathbf{xc}_i \cdot \mathbf{xc}_i' / (N^H \cdot Var(c_k))$ for $k = 1, ..., K_c$, where $Var(c_k)$ is the sample variance of $c_k$ in the HRS subsample. The prior distributions of $\boldsymbol{\lambda}_k$ are independent.

7. $V_c^{-1} \equiv H_c \sim Wishart\left(\underline{V}_c, \underline{S}_c^{-1}\right)$, so that the expectation of $H_c$ is equal to $\underline{V}_c \cdot \underline{S}_c^{-1}$. We set $\underline{V}_c = 4$ and specify that $S_c$ is a diagonal matrix with diagonal elements equal to $\underline{V}_c \cdot 0.7 \cdot Var(c_k)$. This prior is based on that for the normal linear regression model proposed in Geweke (2005), Chapter 5, and specifies that for each $c_k$ the population multiple correlation coefficient $1 - \frac{1}{Var(c_k)H_{c,kk}}$ is equal to 0.3 at the prior expectation of $H_c$. The prior probability that this coefficient is greater than 50% is 23%.

8. $s_{22}h_{22} \sim \chi^2\left(\underline{V}_\sigma\right)$, where $\underline{V}_\sigma = 1$ and $s_{22} = 0.039$. This prior sets the population multiple correlation coefficient $1 - \frac{1}{Var(E_i^*)h_{22}}$ to 0.90 at the prior expectation of $h_{22}$. This reflects a prior belief that the fraction of the variance in expected expenditure $E_i^*$ due to unobserved determinants is much lower than that due to the wide array of observed health status characteristics that we use.[54] It seems plausible that expected expenditure is mostly due to observable health factors. However, our prior also allows for substantial prior variability in this coefficient. For example, the prior probability that it is less than 0.30 is 0.31. At the same time, the prior of $h_{22}$ is flexible enough that unobserved factors can account for all variability in health care expenditure: the prior probability that $1/h_{22}$ is greater than the sample variance of $E_i$ (equal to 2.13 for expenditure measured in tens thousands of dollars) is 0.11.

9. $\sigma_{12} \sim N\left(\underline{\sigma_{12}}, \underline{h}_{12}^{-1}\right)$, where $\underline{\sigma_{12}} = 0$ and $\underline{h}_{12}^{-1} = 50$. Together, the prior specifications of $h_{22}$ and $\sigma_{12}$ imply a significant uncertainty about the strength of the relationship between the unobservables, i.e. it specifies that the 1st, 25th, 75th and 99th percentiles of the prior distribution of the correlation coefficient between $\varepsilon_1$ and $\varepsilon_2$, $\frac{\sigma_{12}}{\sqrt{(1+h_{22}\cdot\sigma_{12}^2)h_{22}^{-1}}}$, are -0.91, -0.25, 0.25, 0.91, respectively.

10. $\boldsymbol{\delta}_j \sim N\left(\underline{\boldsymbol{\delta}}, \underline{\mathbf{H}}_\delta^{-1}\right)$ for $j = 1, \ldots, m-1$, where we specify $\underline{\boldsymbol{\delta}} = \mathbf{0}$ and $\underline{\mathbf{H}}_\delta = 0.1 \cdot \sum_i \mathbf{xw}_i \cdot \mathbf{xw}_i'/N$.

## A-2. Posterior Simulation Algorithm

To obtain the posterior distribution of parameters of the model we construct a Gibbs sampling algorithm. We split the parameters vector into several blocks introduced in section 4.3 so that it is relatively easy to sample from the conditional posterior distributions of each block. Let $\boldsymbol{\theta}_{-\theta_k}$ denote the vector of parameters $\boldsymbol{\theta}$ with the block of parameters $\theta_k$ removed. Define $V_{11} = 1 + \frac{\sigma_{12}^2}{\sigma_{22}} + 2\alpha_0\sigma_{12} + \alpha_0^2\sigma_{22}$, $V_{12} = \sigma_{12} + \alpha_0\sigma_{22}$, and $C_{ki} = c_{ki}^m \cdot S_i + c_{ki}^o \cdot (1 - S_i)$. The Gibbs sampler iteratively draws from the conditional posterior distributions of the following blocks of parameters and latent data:

1. The posterior conditional distribution of $\alpha_0$, $p(\alpha_0|\boldsymbol{\theta}_{-\alpha_0}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data})$, is proportional to the product of it's prior density $p(\alpha_0)$ given in Section A-1 and the density

[54] To set this prior distribution we approximate the sample distribution of $E_i^*$ by the imputed health care expenditure in Medicare only state, as described in the footnote to the discussion of the prior of $\alpha_0$ (bullet point 1). Hence, we set $Var(E^*)$ to $0.59^2$ for $E^*$ measured in tens thousands of dollars.

of observable and latent data $p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \alpha_0, \boldsymbol{\theta}_{-\alpha_0})$ given in equation (19). This distribution is not of any known form and is sampled using the random walk Metropolis-Hastings algorithm. In particular, on iteration $n$ the algorithm draws a proposal value $\widetilde{\alpha}_0$ from $N(\alpha_0^{n-1}, v_{\alpha_0})$, where the subscript $n-1$ indicates the value of $\alpha_0$ from a previous iteration of the Gibbs sampler. The proposal $\widetilde{\alpha}_0$ is accepted as the new draw $\alpha_0^n$ with probability

$$\rho_{\alpha_0} = \min\{1, \frac{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \widetilde{\alpha}_0, \boldsymbol{\theta}_{-\alpha_0})p(\widetilde{\alpha}_0)}{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \alpha_0^{n-1}, \boldsymbol{\theta}_{-\alpha_0})p(\alpha_0^{n-1})}\}.$$

The variance of the proposal distribution $v_{\alpha_0}$ was set so that 45% of the new draws were accepted.

2. The posterior conditional distribution of $\boldsymbol{\alpha}$, $p(\boldsymbol{\alpha}|\boldsymbol{\theta}_{-\alpha}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data})$, is proportional to the product of it's prior density $p(\boldsymbol{\alpha})$ given in Section A-1 and the density of observable and latent data $p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \boldsymbol{\theta})$ given in equation (19). To derive the posterior conditional distribution of $\boldsymbol{\alpha}$ we first need to establish some notation.

   Without loss of generality assume that the observations are arranged so that the first $N^M$ observations belong to MCBS subset, and the last $N^H$ belong to the HRS subset. Let $\widetilde{I}_i^* = I_i^* - \alpha_0 \boldsymbol{\beta}'_{s_i}\mathbf{xe}_i$, and $\widetilde{\widetilde{I}}_i^* = \widetilde{I}_i^* - \frac{V_{12}}{\sigma_{s_i}^2 + \sigma_{22}}(\widehat{E}_i - \boldsymbol{\beta}'_{s_i}\mathbf{xe}_i - \boldsymbol{\gamma}'_{s_i}\mathbf{yi}_i)$. Let $\widetilde{\mathbf{I}}_S^*$ denote the vector of elements $\widetilde{I}_i^*$ for $i = 1, ..., N^M$ (MCBS observations), while $\widetilde{\mathbf{I}}_{1-S}^*$ be the vector of elements $\widetilde{I}_i^*$ for $i = N^M + 1, ..., N$ (HRS observations). Let $\widetilde{\widetilde{\mathbf{I}}}_S^*$ be defined similarly for the elements $\widetilde{\widetilde{I}}_i^*$, $i = 1, ..., N^M$ (MCBS observations). Also, let $\mathbf{Z}_S$ be the matrix with the rows $[\sigma_{s_i}^2, \sigma_{s_i}^2 \cdot C_{1i}, \mathbf{xi}'_i, C_{1i}, ..., C_{4i}]$ for $i = 1, ..., N_M$, while $\mathbf{Z}_{1-S}$ be the matrix with these rows for $i = N^M + 1, ..., N$.

   Then, it can be shown that the posterior conditional distribution of $\boldsymbol{\alpha}$ is given by:

$$p(\boldsymbol{\alpha}|\boldsymbol{\theta}_{-\alpha}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data}) \sim N(\overline{\boldsymbol{\alpha}}, \overline{\mathbf{H}}_\alpha^{-1}),$$

   where

$$\overline{\mathbf{H}}_\alpha = \underline{\mathbf{H}}_\alpha + \frac{1}{V_{11}}\mathbf{Z}'_{1-S}\mathbf{Z}_{1-S} + \mathbf{Z}'_S\mathbf{Q}_\alpha\mathbf{Z}_S$$

   and

$$\overline{\boldsymbol{\alpha}} = \overline{\mathbf{H}}_\alpha^{-1}[\underline{\mathbf{H}}_\alpha\underline{\boldsymbol{\alpha}} + \frac{1}{V_{11}}\mathbf{Z}'_{1-S}\widetilde{\mathbf{I}}_{1-S}^* + \mathbf{Z}'_S\mathbf{Q}_\alpha\widetilde{\widetilde{\mathbf{I}}}_S^*],$$

   and where $\mathbf{Q}_\alpha$ is the $N^M \times N^M$ diagonal matrix with the $ii^{th}$ element given by $\frac{1}{V_{11} - \frac{V_{12}^2}{\sigma_{s_i}^2 + \sigma_{22}}}$.

3. The posterior conditional distribution of $\boldsymbol{\beta}_j$, $p(\boldsymbol{\beta}_j|\boldsymbol{\theta}_{-\beta_j}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data})$, is proportional to the product of it's prior density $p(\boldsymbol{\beta}_j)$ given in Section A-1 and the density

of observable and latent data given in equation (19). To derive this distribution for $j = 1, ..., m$ we need to establish the following notation:

Let $\widehat{I}_i^* = I_i^* - [\sigma_{s_i}^2, \sigma_{s_i}^2 \cdot C_{1i}, \mathbf{xi}_i', C_{1i}, ..., C_{4i}]\boldsymbol{\alpha}$, $\widetilde{E}_i = \widehat{E}_i - \boldsymbol{\gamma}_{s_i}'\mathbf{yi}_i$, $\widetilde{\mathbf{xe}}_i = \alpha_0 \cdot \mathbf{xe}_i$. Define $\mathbf{XE}_S^j$ and $\widetilde{\mathbf{XE}}_S^j$ as the matrices with the rows $\mathbf{xe}_i'$ and $\widetilde{\mathbf{xe}'}_i$ respectively for observations $i$ such that $s_i = j$ and $i \in MCBS$. Let $\mathbf{XE}_{1-S}^j$ and $\widetilde{\mathbf{XE}}_{1-S}^j$ be similarly constructed matrices for observations $i \in HRS$. Similarly, let $\widehat{\mathbf{I}}_S^{*j}$ and $\widehat{\mathbf{I}}_{1-S}^{*j}$ denote the vectors of $\widehat{I}_i^*$ for $i$ with $s_i = j$ and $i \in MCBS$, or $s_i = j$ and $i \in HRS$, respectively. Let $\widetilde{\mathbf{E}}_S^j$ denote the vector of $\widetilde{E}_i$ for $i$ such that $s_i = j$ and $i \in MCBS$. Also, let the matrix $F^j$ with the elements $f_{kl}^j$ be defined as

$$F^j = \begin{bmatrix} V_{11} & V_{12} \\ V_{12} & \sigma_{22} + \sigma_j^2 \end{bmatrix}^{-1}.$$

Then for $j = 1, ..., m$ the posterior conditional distribution of $\boldsymbol{\beta}_j$ is independent of $\boldsymbol{\beta}_l$ for $l \neq j$ and is given by:

$$\boldsymbol{\beta}_j | (\boldsymbol{\theta}_{-\beta_j}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{I}^*, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data}) \sim N(\overline{\boldsymbol{\beta}}_j, \overline{\mathbf{H}}_{\beta_j}^{-1}),$$

where

$$\overline{\mathbf{H}}_{\beta_j} = \underline{\mathbf{H}}_\beta + f_{11}^j \widetilde{\mathbf{XE}}_S^{j'} \widetilde{\mathbf{XE}}_S^j + 2f_{12}^j \widetilde{\mathbf{XE}}_S^{j'} \mathbf{XE}_S^j + f_{22}^j \mathbf{XE}_S^{j'} \mathbf{XE}_S^j + \frac{1}{V_{11}} \widetilde{\mathbf{XE}}_{1-S}^{j'} \widetilde{\mathbf{XE}}_{1-S}^j$$

and

$$\overline{\boldsymbol{\beta}}_j = \overline{\mathbf{H}}_{\beta_j}^{-1} [\underline{\mathbf{H}}_\beta \underline{\boldsymbol{\beta}} + f_{11}^j \widetilde{\mathbf{XE}}_S^{j'} \widehat{\mathbf{I}}_S^{*j} + f_{12}^j \widetilde{\mathbf{XE}}_S^{j'} \widetilde{\mathbf{E}}_S^j + f_{12}^j \mathbf{XE}_S^{j'} \widehat{\mathbf{I}}_S^{*j} + f_{22}^j \mathbf{XE}_S^{j'} \widetilde{\mathbf{E}}_S^j + \frac{1}{V_{11}} \widetilde{\mathbf{XE}}_{1-S}^{j'} \widehat{\mathbf{I}}_{1-S}^{*j}].$$

4. The posterior conditional distribution of $\boldsymbol{\gamma}_j$, $p(\boldsymbol{\gamma}_j | \boldsymbol{\theta}_{-\boldsymbol{\gamma}_j}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data})$, is proportional to the product of it's prior density $p(\boldsymbol{\gamma}_j)$ given in Section A-1 and the density of observable and latent data given in equation (19). To derive the posterior conditional distribution of $\boldsymbol{\gamma}_j$ we need to establish the following notation. Let $\breve{I}_i^* = I_i^* - \alpha_0 \mathbf{xe}_i \boldsymbol{\beta}_{s_i} - [\sigma_{s_i}^2, \sigma_{s_i}^2 \cdot C_{1i}, \mathbf{xi}_i', C_{1i}, ..., C_{4i}]\boldsymbol{\alpha}$, and $\breve{\mathbf{I}}_S^{*j}$ denote a vector of $\breve{I}_i^*$ for $i$ with $s_i = j$ and $i \in MCBS$. Also, let $\breve{E}_i = \widehat{E}_i - \mathbf{xe}_i \boldsymbol{\beta}_{s_i}$, and $\breve{\mathbf{E}}_S^j$ denote a vector of $\breve{E}_i$ for $i$ with $s_i = j$ and $i \in MCBS$. Let $\mathbf{I}_S^j$ denote the matrix with raws $[Y_i, I_i]$ for $i$ with $s_i = j$ and $i \in MCBS$.

   Then, the posterior conditional distribution of $\boldsymbol{\gamma}_j$ for $j = 1, ...m$ is independent of $l \neq j$ and is given by:

$$\boldsymbol{\gamma}_j | (\boldsymbol{\theta}_{\boldsymbol{\gamma}_j}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data}) \sim N(\overline{\boldsymbol{\gamma}}_j, \overline{h}_{\boldsymbol{\gamma}_j}^{-1}),$$

where

$$\overline{h}_{\boldsymbol{\gamma}_j} = \underline{h}_\gamma + f_{22}^j \mathbf{I}_S^{j'} \mathbf{I}_S^{j'}$$

and

$$\overline{\boldsymbol{\gamma}_j} = \overline{h}_{\gamma_j}^{-1}(\underline{h}_\gamma \underline{\gamma} + \mathbf{I}_S^{j'}(f_{22}^j \breve{\mathbf{E}}_S^j + f_{12}^j \breve{\mathbf{I}}_S^{*j})).$$

5. The posterior conditional distributions of $h_j$ for $j = 1, ...m$ are proportional to the product of prior density of $h_j$, $p(h_j)$ given in Section A-1 and the density of observable and latent data as defined in (19). It is easy to see that the posterior conditional distributions of $h_j$ for $j = 1, ...m$ are independent of those of $h_l$ for $l \neq j$. For all $j$ the posterior conditional distribution of $h_j$ is not of any known form and is sampled using the Metropolis-Hastings algorithm. In particular, on iteration $n$ we draw the proposal value $\widetilde{h_j}$ from gamma distribution with the parameters $(\frac{v_j}{2}, \frac{2h_j^{(n-1)}}{v_j})$. Note, that the expected value of this distribution is equal to $h_j^{(n-1)}$. We set the parameters $v_j$ for $j = 1, ..., m$ so that the new draws are accepted with a probability of 0.45. Denote the probability density of this proposal gamma distribution as $g(\widetilde{h_j}|h_j^{(n-1)})$. We accept $\widetilde{h_j}$ as the new draw $h_j^n$ with probability

$$\rho_{h_j} = \min\{1, \frac{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \widetilde{h_j}, \boldsymbol{\theta}_{-h_j})p(\widetilde{h_j})g(h_j^{(n-1)}|\widetilde{h_j})}{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, h_j^{(n-1)}, \boldsymbol{\theta}_{-h_j})p(h_j^{(n-1)})g(\widetilde{h_j}|h_j^{(n-1)})}\}$$

6. The posterior conditional distribution of $\Lambda$, $p(\Lambda|\boldsymbol{\theta}_{-\Lambda}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data})$, is proportional to the product of it's prior density $p(\Lambda)$ given in Section A-1 and the density of observable and latent data given in equation (19). To obtain the conditional posterior distribution of $\boldsymbol{\Lambda}$ we need to establish the following notation. Let $\mathbf{XC}$ denote the matrix of covariates $\mathbf{xc}_i$ for observations $i = 1, ..., N$, i.e. $\mathbf{XC} = [\mathbf{xc}_1, ..., \mathbf{xc}_N]'$. Let $D_K$ denote the identity matrix of size $K$ and let $\mathbf{Z}_\Lambda = D_4 \otimes \mathbf{XC}$, where $\otimes$ denotes the Kroneker product.

   Then, the posterior conditional distribution of $\boldsymbol{\Lambda}$ is given by:

$$\boldsymbol{\Lambda}|(\boldsymbol{\theta}_{-\Lambda}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{W}, \mathbf{data}) \sim N(\overline{\boldsymbol{\Lambda}}, \overline{\mathbf{H}}_\Lambda^{-1}),$$

   where

$$\overline{\mathbf{H}}_\Lambda = \underline{\mathbf{H}}_\Lambda + \mathbf{Z}_\Lambda'(V_c \otimes D_N)^{-1}\mathbf{Z}_\Lambda$$

   and

$$\overline{\boldsymbol{\Lambda}} = \overline{\mathbf{H}}_\Lambda^{-1}[\underline{\mathbf{H}}_\Lambda \underline{\boldsymbol{\Lambda}} + \mathbf{Z}_\Lambda'(V_c \otimes D_N)^{-1}\mathbf{C}].$$

   In the above expressions $\underline{\mathbf{H}}_\Lambda$ is a block-diagonal matrix with the diagonal blocks $\underline{\mathbf{H}}_{\lambda_k}$, $k = 1, ..., 4$, $\underline{\boldsymbol{\Lambda}} = [\underline{\boldsymbol{\lambda}}_1', \underline{\boldsymbol{\lambda}}_2', \underline{\boldsymbol{\lambda}}_3', \underline{\boldsymbol{\lambda}}_4']'$, $\mathbf{C} = [\mathbf{C}_1', ..., \mathbf{C}_4']'$, and for $k = 1, ..., 4$ the $N \times 1$ vectors $\mathbf{C}_k$ consist of the elements $C_{ki}$.

7. The posterior conditional distribution of the inverse of the variance-covariance matrix

of the SAS variables missing from the MCBS data, $H_c \equiv V_c^{-1}$, $p(H_c|\boldsymbol{\theta}_{-H_c}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{W}, \mathbf{s}, \mathbf{data})$, is proportional to the product of its prior probability given in Section A-1 and the density of observable and latent data as defined in (19), and is given by:

$$H_c|(\boldsymbol{\theta}_{-H_c}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{I}^*, \mathbf{C}^m, \mathbf{C}^o, \mathbf{W}, \mathbf{s}, \mathbf{data}) \ \sim \ W((\underline{S}_c + S_c)^{-1}, \underline{V}_c + N),$$

where

$$S_c = \begin{bmatrix} (\mathbf{C}_1 - \mathbf{XC}\boldsymbol{\lambda}_1)'(\mathbf{C}_1 - \mathbf{XC}\boldsymbol{\lambda}_1) & \cdots & (\mathbf{C}_1 - \mathbf{XC}\boldsymbol{\lambda}_1)'(\mathbf{C}_4 - \mathbf{XC}\boldsymbol{\lambda}_4) \\ \vdots & \ddots & \vdots \\ (\mathbf{C}_4 - \mathbf{XC}\boldsymbol{\lambda}_{4x})'(\mathbf{C}_1 - \mathbf{XC}\boldsymbol{\lambda}_1) & \cdots & (\mathbf{C}_4 - \mathbf{XC}\boldsymbol{\lambda}_4)'(\mathbf{C}_4 - \mathbf{XC}\boldsymbol{\lambda}_4) \end{bmatrix}.$$

8. The posterior conditional distribution of $h_{22}$ is proportional to the product of it's prior density $p(h_{22})$ given in Section A-1 and the density of observable and latent data given in equation 19. This distribution is not of any known form and is sampled using the Metropolis-Hastings algorithm.

   In particular, on iteration $n$ we draw the proposal value $\widetilde{h_{22}}$ from gamma distribution with the parameters $(\frac{v_{s_2}}{2}, \frac{2h_{22}^{n-1}}{v_{s_2}})$. Note, that the expected value of this distribution is equal to $h_{22}^{n-1}$. We set the parameter $v_{s_2}$ so that the acceptance rate is about 45%. Denote the probability density of this proposal gamma distribution as $g(\widetilde{h_{22}}|h_{22}^{n-1})$. We accept $\widetilde{h_{22}}$ as the new draw $h_{22}^n$ with probability

$$\rho_{\sigma_{22}} = \min\{1, \frac{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \widetilde{h_{22}}, \boldsymbol{\theta}_{-h_{22}})p(\widetilde{h_{22}})g(h_{22}^{n-1}|\widetilde{h_{22}})}{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, h_{22}^{n-1}, \boldsymbol{\theta}_{-h_{22}})p(h_{22}^{n-1})g(\widetilde{h_{22}}|h_{22}^{n-1})}\}$$

9. The posterior conditional distribution of $\sigma_{12}$ is proportional to the product of it's prior density $p(\sigma_{12})$ given in Section A-1 and the density of observable and latent data given in equation (19). This distribution is not of any known form and is sampled using the random walk Metropolis-Hastings algorithm.

   In particular, on iteration $n$ draw the proposal value $\widetilde{\sigma}_{12}$ from $N(\sigma_{12}^{n-1}, v_{\sigma_{12}})$. Accept $\widetilde{\sigma}_{12}$ as the new draw $\sigma_{12}^n$ with probability

$$\rho_{\sigma_{12}} = \min\{1, \frac{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \widetilde{\sigma}_{12}, \boldsymbol{\theta}_{-\sigma_{12}})p(\widetilde{\sigma_{12}})}{p(\mathbf{I}^*, \mathbf{I}, \widehat{\mathbf{E}}, \mathbf{E}^o, \mathbf{C}^o, \mathbf{C}^m, \mathbf{s}, \mathbf{W}|\mathbf{S}, \mathbf{X}, \sigma_{12}^{n-1}, \boldsymbol{\theta}_{-\sigma_{12}})p(\sigma_{12}^{n-1})}\}.$$

   The variance of the proposal distribution $v_{\sigma_{12}}$ was set so that 20% of the new draws are accepted.

10. The posterior conditional distribution of the vector of coefficients $\boldsymbol{\delta}_j$ which determine the latent type propensities $\widetilde{W}_{ij}$ is proportional to the product of the prior density of $\boldsymbol{\delta}_j$ given in Section A-1 and (19). It is easy to see that the posterior conditional

distributions of $\boldsymbol{\delta}_j$ are independent across $j$ and are given by:

$$\boldsymbol{\delta}_j|(\boldsymbol{\theta}_{-\delta_j}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{W}, \mathbf{s}, \mathbf{data}) \;\sim\; N(\overline{\boldsymbol{\delta}}_j, \overline{\mathbf{H}}_\delta^{-1})$$

where

$$\begin{aligned}
\overline{\mathbf{H}}_\delta &= \underline{\mathbf{H}}_\delta + \mathbf{XW}'\mathbf{XW}, \\
\overline{\boldsymbol{\delta}}_j &= \overline{\mathbf{H}}_\delta^{-1}[\underline{\mathbf{H}}_\delta \underline{\boldsymbol{\delta}} + \mathbf{XW}'\widetilde{\mathbf{w}}_j] \text{ for } j = 1, ..., m-1.
\end{aligned}$$

11. Latent utility of health insurance $I_i^* \sim p(I_i^*|\boldsymbol{\theta}, \widehat{\mathbf{E}}, \mathbf{I}_{-i}^*, \mathbf{C}^m, \mathbf{s}, \mathbf{data})$. From (19) the kernel of this posterior distribution for $i \in MCBS$ is given by

$$\begin{aligned}
& \exp(-(I_i^* - \alpha_0 \mathbf{xe}_i \boldsymbol{\beta}_{s_i} - \alpha_1 \sigma_{s_i}^2 - \alpha_2 \sigma_{s_i}^2 c_{1i}^m - \boldsymbol{\alpha}_3' \mathbf{xi}_i - \mathbf{c}_i^{m'} \boldsymbol{\alpha}_4 \\
- & \frac{V_{12}}{\sigma_{s_i}^2 + \sigma_{22}} (\widehat{E}_i - \boldsymbol{\beta}_{s_i}' \mathbf{xe}_i - \boldsymbol{\gamma}_{s_i}' \mathbf{yi}_i))^2 / (2(V_{11} - \frac{V_{12}^2}{\sigma_{s_i}^2 + \sigma_{22}}))) \\
\cdot & (\iota(I_i^* \geq 0) \cdot \iota(I_i = 1) + \iota(I_i^* < 0) \cdot \iota(I_i = 0)),
\end{aligned}$$

while for $i \in HRS$ it is given by

$$\begin{aligned}
& \exp(-(I_i^* - \alpha_0 \mathbf{xe}_i \boldsymbol{\beta}_{s_i} - \alpha_1 \sigma_{s_i}^2 - \alpha_2 \sigma_{s_i}^2 c_{1i}^o - \boldsymbol{\alpha}_3' \mathbf{xi}_i - \mathbf{c}_i^{o'} \boldsymbol{\alpha}_4)^2 / (2V_{11})) \\
\cdot & (\iota(I_i^* \geq 0) \cdot \iota(I_i = 1) + \iota(I_i^* < 0) \cdot \iota(I_i = 0)).
\end{aligned}$$

These can be recognized as kernels of truncated normal distributions. Thus,

$$I_i^*|(\boldsymbol{\theta}, \widehat{\mathbf{E}}, \mathbf{I}_{-i}^*, \mathbf{C}^m, \mathbf{W}, \mathbf{s}, \mathbf{data}) \sim TN_{R(I_i)}(\overline{I}_i^*, V_{I^*}),$$

where $TN_{R(I)}(a, b)$ denotes normal distribution with mean $a$ and variance $b$ truncated to interval $R(I)$, $R(0) = (-\infty, 0]$, $R(1) = (0, \infty)$. For $i \in MCBS$ we have

$$\overline{I}_i^* = \alpha_0 \mathbf{xe}_i \boldsymbol{\beta}_{s_i} + \alpha_1 \cdot \sigma_{s_i}^2 + \alpha_2 \cdot \sigma_{s_i}^2 c_{1i}^m + \boldsymbol{\alpha}_3' \mathbf{xi}_i + \boldsymbol{\alpha}_4' \mathbf{c}_i^m + \frac{V_{12}}{\sigma_{s_i}^2 + \sigma_{22}} (E_i^* - \boldsymbol{\beta}_{s_i}' \mathbf{xe}_i - \boldsymbol{\gamma}_{s_i}' \mathbf{yi}_i),$$

$$V_{I^*} = V_{11} - \frac{V_{12}^2}{\sigma_{s_i}^2 + \sigma_{22}},$$

while for $i \in HRS$ we have

$$\overline{I}_i^* = \alpha_0 \mathbf{xe}_i \boldsymbol{\beta}_{s_i} + \alpha_1 \cdot \sigma_{s_i}^2 + \alpha_2 \cdot \sigma_{s_i}^2 c_{1i}^o + \boldsymbol{\alpha}_3' \mathbf{xi}_i + \boldsymbol{\alpha}_4' \mathbf{c}_i^o$$

and

$$V_{I^*} = V_{11}.$$

12. Notional expenditure $\widehat{E}_i \sim p(\widehat{E}_i|\boldsymbol{\theta}, \widehat{\mathbf{E}}_{-i}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{W}, \mathbf{s}, \mathbf{data}, S_i = 1)$. The kernel of this

posterior distribution is given by

$$\exp(-\frac{(\widehat{E}_i - \mathbf{xe}_i\boldsymbol{\beta}_{s_i} - \mathbf{yi}_i\boldsymbol{\gamma}_{s_i} - \frac{V_{12}}{V_{11}}(I_i^* - \alpha_0\mathbf{xe}_i\boldsymbol{\beta}_{s_i} - \alpha_1\cdot\sigma_{s_i}^2 - \alpha_2\cdot\sigma_{s_i}^2 c_{1i}^m - \boldsymbol{\alpha}_3'\mathbf{xi}_i - \boldsymbol{\alpha}_4'\mathbf{c}_i^m))^2}{2(\sigma_{s_i}^2 + \sigma_{22} - \frac{V_{12}^2}{V_{11}})})$$
$$\cdot \quad (\iota(E_i^o = \widehat{E}_i)\cdot\iota(\widehat{E}_i \geq 0) + \iota(E_i^o = 0)\cdot\iota(\widehat{E}_i < 0)). \tag{24}$$

Thus, if $E_i = 0$ we draw notional expenditure from:

$$\widehat{E}_i|(\boldsymbol{\theta}, \widehat{\mathbf{E}}_{-i}, \mathbf{I}^*, \mathbf{C}^m, \mathbf{W}, \mathbf{s}, \mathbf{data}, S_i = 1) \ \sim \ TN_{(-\infty,0]}(\overline{E}_i, \sigma_{s_i}^2 + \sigma_{22} - \frac{V_{12}^2}{V_{11}})$$

where $\overline{E}_i = \mathbf{xe}_i\boldsymbol{\beta}_{s_i} + \mathbf{yi}_i\boldsymbol{\gamma}_{s_i} + \frac{V_{12}}{V_{11}}(I_i^* - \alpha_0\mathbf{xe}_i\boldsymbol{\beta}_{s_i} - \alpha_1\cdot\sigma_{s_i}^2 - \alpha_2\cdot\sigma_{s_i}^2 c_{1i}^m - \boldsymbol{\alpha}_3'\mathbf{xi}_i - \boldsymbol{\alpha}_4'\mathbf{c}_i^m)$, while if $E_i > 0$ we simply set $\widehat{E}_i = E_i$.

13. SAS variables missing from the MCBS: $\mathbf{c}_i^m \sim p(\mathbf{c}_i^m|\boldsymbol{\theta}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}_{-i}^m, \mathbf{W}, \mathbf{s}, \mathbf{data}, S_i = 1)$. The kernel of this posterior distribution is given by

$$\begin{aligned} & \exp(-(I_i^* - \alpha_0\mathbf{xe}_i\boldsymbol{\beta}_{s_i} - \alpha_1\sigma_{s_i}^2 - \alpha_2\sigma_{s_i}^2 c_{1i}^m - \boldsymbol{\alpha}_3'\mathbf{xi}_i - \boldsymbol{\alpha}_4'\mathbf{c}_i^m \\ - \quad & \frac{V_{12}}{\sigma_{s_i}^2 + \sigma_{22}}(\widehat{E}_i - \boldsymbol{\beta}_{s_i}'\mathbf{xe}_i - \boldsymbol{\gamma}_{s_i}'\mathbf{yi}_i)^2/(2(V_{11} - \frac{V_{12}^2}{\sigma_{s_i}^2 + \sigma_{22}}))) \\ \cdot \quad & \exp(-(\mathbf{c}_{1i}^m - XC_i\Lambda)'V_c^{-1}(\mathbf{c}_{1i}^m - XC_i\Lambda)/2) \end{aligned}$$

This kernel can be recognized as that of the conditional distribution $p(\mathbf{c}_i^m|\boldsymbol{\theta}, \widehat{E}_i, \mathbf{xi}_i, \mathbf{xe}_i; I_i^*)$, where the joint distribution $p(I_i^*, \mathbf{c}_i^m|\boldsymbol{\theta}, \widehat{E}_i, \mathbf{xi}_i, \mathbf{xe}_i)$ is multivariate normal with mean

$$\begin{bmatrix} \overline{I}_i^c \\ \overline{\mathbf{c}}_i \end{bmatrix} \equiv \begin{bmatrix} \alpha_0\mathbf{xe}_i\boldsymbol{\beta}_{s_i} + \alpha_1\sigma_{s_i}^2 + \alpha_2\sigma_{s_i}^2\cdot\mathbf{xc}_i'\boldsymbol{\lambda}_1 + \boldsymbol{\alpha}_3'\mathbf{xi}_i + \boldsymbol{\alpha}_4'XC_i\Lambda + \frac{V_{12}}{\sigma_{s_i}^2+\sigma_{22}}(\widehat{E}_i - \boldsymbol{\beta}_{s_i}'\mathbf{xe}_i - \boldsymbol{\gamma}_{s_i}'\mathbf{yi}_i) \\ XC_i\Lambda \end{bmatrix}$$

and variance matrix:

$$\mathbf{V}_{s_i}^c = \begin{pmatrix} v_{11s_i}^c & \boldsymbol{\alpha}_4'V_c + \alpha_2\sigma_{s_i}^2\mathrm{v}_c^{1\cdot} \\ V_c\boldsymbol{\alpha}_4 + \alpha_2\sigma_{s_i}^2\mathrm{v}_c^{\cdot 1} & V_c \end{pmatrix} \equiv \begin{pmatrix} \mathbf{V}_{s_i11}^c & \mathbf{V}_{s_i12}^c \\ \mathbf{V}_{s_i21}^c & V_c \end{pmatrix}$$

where

$$v_{11s_i}^c = V_{11} - \frac{V_{12}^2}{\sigma_{22} + \sigma_{s_i}^2} + \boldsymbol{\alpha}_4'V_c\boldsymbol{\alpha}_4 + \alpha_2^2\sigma_{s_i}^4\cdot\mathrm{v}_c^{11} + 2\alpha_2\sigma_{s_i}^2\sum_{l=1}^{4}\cdot\alpha_{4l}\cdot\mathrm{v}_c^{1l},$$

and where $\mathrm{v}_c^{kl}$ denotes $kl^{th}$ element of $V_c$, while $\mathrm{v}_c^{k\cdot}$ and $\mathrm{v}_c^{\cdot k}$ denote $k^{th}$ row and $k^{th}$ column of $V_c$, respectively. Using the results for the multivariate normal distribution the posterior conditional distribution of $\mathbf{c}_i^m$ is given by

$$\mathbf{c}_i^m|\boldsymbol{\theta}, \widehat{\mathbf{E}}, \mathbf{I}^*, \mathbf{C}_{-i}^m, \mathbf{W}, \mathbf{s}, \mathbf{data}, S_i = 1 \sim N(\overline{\mathbf{c}}_i + \mathbf{V}_{s_i12}^{c'}\mathbf{V}_{s_i11}^{c^{-1}}(I_i^* - \overline{I}_i^c), V_c - \mathbf{V}_{s_i12}^{c'}\mathbf{V}_{s_i11}^{c^{-1}}\mathbf{V}_{s_i12}^c).$$

14. The conditional posterior density kernel of latent type propensities $\widetilde{\mathbf{W}}_i$ is given by:

$$\exp(-\widetilde{W}_{im}^2/2 - \sum_{j=1}^{m-1}(\widetilde{W}_{ij} - \mathbf{x}\mathbf{w}_i'\boldsymbol{\delta}_j)^2/2) \tag{25}$$

$$\cdot \sum_{j=1}^{m}\left(\prod_{l=1}^{m}\iota(\widetilde{W}_{il} \in (-\infty, \widetilde{W}_{ij}])\right) \tag{26}$$

$$\cdot \; g_w(j),$$

where

$$\begin{aligned} g_w(j) &= \left\{\exp\left(-\frac{(I_i^* - \alpha_0\mathbf{x}\mathbf{e}_i\boldsymbol{\beta}_j - \alpha_1\sigma_j^2 - \alpha_2\sigma_j^2 c_{1i}^o - \boldsymbol{\alpha}_3'\mathbf{x}\mathbf{i}_i - \mathbf{c}_i^{o'}\boldsymbol{\alpha}_4)^2}{2V_{11}}\right)\right\}^{S_i=0} \\ &\cdot \left\{(V_{11} - \frac{V_{12}^2}{\sigma_j^2+\sigma_{22}})^{-\frac{1}{2}} \cdot \exp\left(-\left[I_i^* - \alpha_0\mathbf{x}\mathbf{e}_i\boldsymbol{\beta}_j - \alpha_1\sigma_j^2 - \alpha_2\sigma_j^2 c_{1i}^m - \boldsymbol{\alpha}_3'\mathbf{x}\mathbf{i}_i\right.\right.\right. \\ &- \left.\left.\mathbf{c}_i^{m'}\boldsymbol{\alpha}_4 - \frac{V_{12}}{\sigma_j^2+\sigma_{22}}(\widehat{E}_i - \alpha_0\mathbf{x}\mathbf{e}_i\boldsymbol{\beta}_j - \boldsymbol{\gamma}_j'\mathbf{y}\mathbf{i}_i)\right]^2 / (2(V_{11} - \frac{V_{12}^2}{\sigma_j^2+\sigma_{22}}))\right) \\ &\cdot \left.(\sigma_j^2+\sigma_{22})^{-\frac{1}{2}} \cdot \exp\left(-\frac{(\widehat{E}_i - \alpha_0\mathbf{x}\mathbf{e}_i\boldsymbol{\beta}_j - \boldsymbol{\gamma}_j\mathbf{y}\mathbf{i}_i)^2}{2(\sigma_j^2+\sigma_{22})}\right)\right\}^{S_i=1} \end{aligned}$$

Draws from this distribution are obtained by the Metropolis within Gibbs step suggested in Geweke and Keane (2007). The candidate draw $\widetilde{\mathbf{W}}_i^*$ is obtained from the normal density with the kernel given by (25). The function (26) then determines the candidate type $j^* : \widetilde{W}_{ij^*} \geq \widetilde{W}_{il}$ for all $l = 1, ..., m$. The candidate values are then accepted as new draws $\widetilde{\mathbf{W}}_i^n$ and $s_i^n$ with probability

$$\min\left\{\frac{g_w(j^*)}{g_w(j^{(n-1)})}, 1\right\},$$

where $j^{(n-1)}$ denotes observation's $i$ type from the previous iteration, i.e. $j^{(n-1)} = s_i^{n-1}$.

We checked that this algorithm was correctly implemented using the joint distribution tests of Geweke (2004).

## A-3. Inclusion of Exogenous Variables in the Equations of the Model

Table A-2 shows specification of the equations of the model in terms of exogenous covariates included in each equation. As discussed in section 4.1.A, to identify selection and moral hazard effects we use cross-equation exclusion restrictions. In particular, we assume that (i) health status variables (i.e. health factors 2-23) and survey year indicator affect insurance status only indirectly (i.e. through their effect on expenditure risk $E_i^*$), and (ii) SAS variables, such as education, income, risk tolerance, cognitive ability, longevity expectations and financial planning horizon, enter the insurance equation but not the expenditure risk equation, once we condition on health status variables. That is, these SAS variables may affect one's health indirectly by shifting investment in health, but once we condition on health itself, they have no direct effect on one's health expenditure risk.

The demographic characteristics (i.e. marital status, ethnicity and interactions of gender with marital status and age) are included in both the expenditure and the insurance equations (in the full model). These variables are included in the expenditure equation to capture differences in health status and tastes for medical care between different demographic groups. Similarly, these variables are included in the final specification of the insurance equation to capture heterogeneity in tastes for insurance. We do not include these variables in the baseline model because insurers cannot legally price Medigap policies based on race or marital status.

The specification of the insurance equation ($I_i^*$) is the same as in FKS. In particular, in addition to expenditure risk $E_i^*$, the benchmark model includes only insurance pricing variables (polynomial in age, gender and location of residence). The potential SAS variables (education, income, risk tolerance, cognitive ability, longevity expectations, financial planning horizon, race and marital status) are progressively added to the insurance equation in extended specifications. Hence, variables indicated by "SAS" in column 3 of Table A-2 correspond to the vector $[\mathbf{xi}_i', \mathbf{c}_i]$ (see equation (3)) in the full specification of the insurance equation, and $\mathbf{c}_i$ consists of variables indicated in the last four rows of column 3 (risktol, cogn, finpln and praliv75).

The variables marked by "Yes" in column 4 of Table A-2 correspond to the vector $\mathbf{xe}_i$ of characteristics included in the specification of the expenditure risk $E_i^*$ (see equation (4)). Income in thousands of dollars is included in the expenditure equation (1) but is not included in the expenditure risk equation (4)).

The variables marked by "Yes" in column 5 of Table A-2 correspond to the vector $\mathbf{xw}_i$ of variables affecting type propensities $\widetilde{W}$ (see equation (5)). Note, that the equations for type propensities include most of the variables included in $\mathbf{xe}_i$, with the exception of the polynomial terms in age and the interactions of age with gender and gender with marital status. We omit these variables to reduce the number of parameters, as the specification for conditional mean of expenditure is already very flexible.

Finally, the model for missing SAS variable (SAS) includes most of the exogenous variables used in the analysis to maximize predictive power. The variables marked by "Yes" in column 6 of Table A-2 correspond to the vector $\mathbf{xc}_i$ of exogenous variables included in the prediction equation (7).

Table A-2: Exogenous variables included in equations for insurance status, expenditure risk, type probabilities and the prediction model for the SAS variables.

| Variable | Description | $I^*$ | $E^*$ | $\widetilde{W}$ | SAS |
|---|---|---|---|---|---|
| 1 | 2 | 3 | 4 | 5 | 6 |
| Female | Indicator for female | Yes | Yes | Yes | Yes |
| Age | Age, years | Yes | Yes | Yes | Yes |
| Age$^2$ | Age squared | Yes | Yes | | Yes |
| Age$^3$ | Age cubed | Yes | Yes | | Yes |
| Married | Indicator for being married | SAS | Yes | Yes | Yes |
| Age*Female | Interaction of age polynomial with Female | SAS | Yes | | |
| Married*Female | Interaction of Married and Female | SAS | Yes | | |
| Health factor 1 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 3 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 7 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 8 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 10 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 11 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 17 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 20 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 22 | Health Status Factor | | Yes | Yes | Yes |
| Health factor 23 | Health Status Factor | | Yes | Yes | Yes |
| Black | Indicator for race black | SAS | Yes | Yes | Yes |
| Hispanic | Indicator for Hispanic | SAS | Yes | Yes | Yes |
| Survey year | Year | | Yes | Yes | |
| hgc: ls8th | Education: less than high school | SAS | | | Yes |
| hgc: somehs | Education: some high school | SAS | | | Yes |
| hgc: hs | Education: high school | SAS | | | Yes |
| hgc: somecol | Education: some college | SAS | | | Yes |
| hgc: college | Education: college | SAS | | | Yes |
| hgc: gradschl | Education: grad. school | SAS | | | Yes |
| hgc: nr | Education non-response | SAS | | | Yes |
| inc 5k-10k | Income: $5-10 thousand | SAS | | | Yes |
| inc 10k-15k | Income: $10-15 thousand | SAS | | | Yes |
| inc 15k-20k | Income: $15-20 thousand | SAS | | | Yes |
| inc 20k-25k | Income: $20-25 thousand | SAS | | | Yes |
| inc 25k-30k | Income: $25-30 thousand | SAS | | | Yes |
| inc 30k-35k | Income: $30-35 thousand | SAS | | | Yes |
| inc 35k-40k | Income: $35-40 thousand | SAS | | | Yes |
| inc 40k-45k | Income: $40-45 thousand | SAS | | | Yes |
| inc 45k-50k | Income: $45-50 thousand | SAS | | | Yes |
| inc 50plus | Income: $50+ thousand | SAS | | | Yes |
| risktol | Risk tolerance | SAS | | | |
| cogn | Congnition factor | SAS | | | |
| finpln | Financial planning horizon | SAS | | | |
| praliv75 | Subjective probability to live to be 75 or more | SAS | | | |

* Note: All equations include indicators for census divisions. The variables labelled "SAS" are not included in the baseline specification of the insurance equation ($I_i^*$). They are added later as potential sources of adverse/advantageous selection. The baseline insurance equation only includes pricing variables and expenditure risk.