Sources of Advantageous Selection: Evidence from the Medigap Insurance Market*

Hanming Fang
Yale University

Michael P. Keane
Yale University

Dan Silverman
University of Michigan

February 21, 2006

Abstract

We find strong evidence for advantageous selection in Medigap insurance market. Using Medicare Current Beneficiary Survey (MCBS) data, we regress "Total Medical Expenditure" on medigap status and other controls. We find that (1) if we only use gender, age and State (three variables used in the pricing of Medigap), then those with Medigap incur about $4,000 less in total medical expenditure than those without Medigap; (2) however, if we add controls for observable health variables, then those with Medigap spend about $2,000 more than those without Medigap. The only way to rationalize these two results is that those with better health are more likely to purchase supplemental coverage. We interpret this as strong evidence of "advantageous selection." We then combine MCBS and Health and Retirement Study (HRS) to investigate the sources of advantageous selection. We find that, risk tolerance alone, even when interacted with measures of risks, is not enough to explain all of the advantageous selection we previously documented. Additional sources of advantageous selection include at least education, income, cognition, longevity expectations and financial planing horizon, among possibly other variables. We find that once we condition on all these factors, individuals with higher expected medical expenditure risks are indeed more likely to buy Medigap insurance.

Keywords: Asymmetric Information; Medigap Insurance; Adverse Selection; Advantageous Selection; Moral Hazard

JEL Classification Numbers: D82, G22, I11.

---

*We thank Pierre-André Chiappori, Amy Finkelstein, Alessandro Lizerri, Jim Poterba, Bernard Salanié, Aloysius Siow, and seminar participants at Toronto and Wharton for encouragement and helpful suggestions. We are grateful to Mike Chernew for the access to, and Tami Swenson for clarification of, the Medicare Current Beneficiary Survey data. All remaining errors are our own.

1 Introduction

Modern economics of insurance, pioneered by Arrow (1963) and Pauly (1974), emphasizes asymmetric information, including both *ex ante* adverse selection and *ex post* moral hazard. Information asymmetry leads to insurance markets that are inefficiently small, and in extreme cases (e.g. Akerlof 1970), may even cause the complete disappearance of markets that would provide welfare gains in the absence of information asymmetries.

The classic equilibrium models of insurance developed by Rothschild and Stiglitz (1976) and Wilson (1977) assume that potential buyers of insurance contracts are privately informed about their own risk type, and will choose from the offered menu of insurance contracts the one that best suits their risk type. The empirical prediction from these models is that there should be a positive correlation between coverage and *ex post* risk in the presence of asymmetric information, either in the form of adverse selection and/or moral hazard. Indeed this empirical prediction provide the basis of empirical test of asymmetric information in several recent papers.[1] Indeed, an important paper by Chiappori, Jullien, Salanié and Salanié (2005) show that, using a suitably modified notion of positive correlation between risk and coverage, this empirical prediction can be generalized to a large class of models when the insurance market is competitive and when the risk aversion of the insured is public knowledge. The intuition is simple: the incentive to purchase insurance is greater for those who know that they are relatively likely to suffer a loss. As a result, if adverse selection is important, there should be a positive correlation between insurance coverage and *ex post* risk; and adding moral hazard into the analysis will only further strengthen this positive correlation.

While this positive correlation between insurance coverage and *ex post* risk is a robust feature of classical insurance models, a recent and growing empirical literature has found little evidence of it in a variety of insurance markets. For example, Cawley and Philipson (1999) used four data sources including the Health and Retirement Study (HRS) and the Asset and Health Dynamics Among the Oldest Old (AHEAD) to examine whether there is a positive correlation between the self-perceived or actual mortality risk and the probability of purchasing life insurance. To the contrary, they found that the mortality rate of U.S. males purchasing life insurance is below that of the uninsured, even when controlling for many factors such as income that may be correlated with life expectancy. Chiappori and Salanié (2000) show that, after controlling for observable characteristics known to automobile insurers, the accident rate is lower for young French drivers who choose

[1] See, e.g., Chiappori and Salanié (2000), and the references in Chiappori (2000).

comprehensive automobile insurance than for those opting for the legal minimum coverage, though the difference is not statistically significant. Cardon and Hendel (2001) use a different approach to test for information asymmetries in health insurance markets. They estimate a structural model of health insurance and health care choices using data on single individuals from the National Medical Expenditure Survey (NMES), and found no link between coverage and *ex post* health care, thus no evidence of information asymmetries.[2] A natural conclusion to draw from these empirical studies is that either there are no asymmetries of information in many important insurance markets, or that the classical insurance models need amending.

A plausible explanation for the empirical puzzle of not finding asymmetric information in insurance markets is the notion of "advantageous selection."[3] The idea behind advantageous selection is that individuals have private information about both their risk type and other dimensions of their preferences for insurance. In this case, whether the association between insurance coverage and *ex post* risk is on net positive depends on whether the selection based on other dimensions of private information goes in the same or opposite direction as the selection based on risk. So far in the literature, risk aversion has been the focus of dimensions of preference that may affect the demand for insurance. For example, in de Meza and Webb (2001), they postulate that individuals have private information about both their risk and their risk aversion. When individuals differ both in risk and risk aversion, then given a menu of insurance contracts, those who choose a higher level of insurance will tend to be either riskier or more risk averse. Traditional adverse selection captures the positive correlation between risk and coverage. The selection based on risk aversion, however, may be advantageous if (1) more risk averse individuals buy more insurance coverage; *and* (2) if more risk averse individuals tend to have lower risks (i.e. if there is a negative correlation between risk and risk aversion). As a result of the possible presence of both adverse and advantageous selections, there may or may not be a positive correlation between insurance coverage and *ex post* risk, if one does not condition on risk aversion. More generally, we say that selection based on private information γ is *advantageous* if γ is positively correlated with insurance coverage but negatively correlated with risk (see Section 3 for details). In a related model, Jullien, Salanié and Salanié (2005) indeed showed that when insurance market is not competitive and when insurees'

[2]In particular, Cardon and Hendel (2001) postulated in their structural model that individuals have identical preferences.

[3]The first verbal description of this phenomenon in the economics literature appears to be Hemenway (1990, 1992) using the term "propitious selection." de Meza and Webb (2001) studies the competitive equilibrium of a stylized insurance market in which agents may have different risk preferences and risk.

risk aversion is private information, then the positive correlation property between coverage and *ex post* risk may not hold. On the empirical side, Finkelstein and McGarry (2005) directly examine the evidence for selection based on multidimensional private information in Long Term Care Insurance market;[4] Asinsky (2005) estimates a two-equation endogenous treatment model which explicitly includes indicators of latent attitudes for undertaking risky behavior using Markov Chain Monte Carlo methods and 2000 Medical Expenditure Panel Survey (MEPS) data. He obtains mixed results regarding the presence of private information regarding risk aversion.

Despite the focus on risk aversion as the additional source of selection in the theoretical literature on advantageous selection, there is no empirical evidence to support or refute its importance. Neither Finkelstein and McGarry (2005) nor Asinsky (2005) has direct measurements of risk aversion in their empirical analysis. This paper fills this gap and makes two main contributions: first, we present direct evidence of advantageous selection in Medigap insurance market; second, we investigate more broadly, beyond the narrow focus on risk aversion, the possible sources of advantageous selection.

A Medigap policy is health insurance sold by private insurance companies to fill the "gaps" in the original Medicare plan.[5] Several important features of the Medigap insurance market make it unusually well-suited for a study of the presence and sources of advantageous selection. The first key feature of the Medigap insurance market is that its supply-side is highly regulated by the government. Government regulations appear in several spheres. First, in most of the States insurance companies can only sell ten standardized Medigap policies. Second, within the six month Medigap open enrollment period, which starts on the first day of the first month in which an individual is both older than 65 and is enrolled in Medicare Part B, an insurance company cannot deny Medigap coverage, or place conditions on a policy, or charge more for pre-existing health conditions. Indeed Robst (2001) found that the average Medigap insurance premium an individual faces depends almost exclusively on his/her State of residence, age and gender. These special features of the supply-side of the Medigap insurance market allow us to focus mainly on the demand-side of the market. Thus, relative to other insurance markets, Medigap insurance market is less competitive both in the offerings of insurance menus and their pricings. As shown in the theoretical analysis of Julien, Salanié and Salanié (2005), the non-competitiveness of the Medigap insurance market is actually essential for the multidimensional private information to manifest itself in terms

[4]We discuss Finkelstein and McGarry (2005) in more details in Section 2.

[5]See Section 4 for more details about Medicare and Medigap insurance market.

of a non-positive correlation between *ex post* risk and coverage. The second key feature of the Medigap insurance market is that it is intimately linked to Medicare, and as such there exists detailed administrative data about diagnosis and expenditures. Indeed we exploit this link and use the Medicare Current Beneficiary Survey (MCBS) which combines survey data and administrative Medicare records as one of the key data sources for our empirical analysis. These supplemented administrative data from Medicare on medical expenditure provide perhaps the most accurate measure of health expenditure risk of any commonly available data. Though MCBS itself does not contain detailed information about risk aversion and other potential candidates for the source of advantageous selection, fortunately another large scale data set, Health and Retirement Study (HRS), which is a longitudinal data set covering a large sample of Medicare eligible population, has unique information about such variables. Our empirical strategy is to use MCBS and HRS jointly to examine the sources of advantageous selection.

Our first contribution – to present direct evidence of advantageous selection in Medigap insurance market – is achieved by using information from the MCBS alone. More specifically, the evidence is established via a series of regressions where "Total Medical Expenditure" is regressed on medigap status and a varying set of other conditioning variables. We find that: first, if we only condition on gender, age and state (three variables used in the pricing of Medigap), then those who purchase Medigap incur about $4,000 *less* in "Total Medical Expenditure" than those without Medigap second, if we add controls for observable health variables, then those with Medigap spend about $2,000 *more* in "Total Medical Expenditure" than those without Medigap.[6] The only way to rationalize these two set of regression results is that those with better observable health are more likely to purchase Medigap coverage. Indeed the conditional correlation of Medigap status on health factors also show a direct evidence that there is a positive correlation between good health factors and Medigap coverage. We interpret this as strong evidence of advantageous selection.

Our second contribution – to examine the sources of advantageous selection – can not be achieved using MCBS data alone, because it lacks direct information on most of the plausible candidates including risk aversion, cognition, longevity expectations etc. Such information, however, is available in HRS. But unfortunately HRS is not properly linked to Medicare administrative records at this moment. Our empirical strategy proposes a way to combine the information in MCBS and HRS. We first use MCBS data to estimate prediction equations for the mean and variance of total

[6]The second result does not depend on whether we directly use a long list of health measures in the MCBS data or use health factors extracted from these direct health measures.

medical expenditures and then use these estimates to impute a health expenditure risk for individuals in the HRS. We then regress, using the HRS sample, Medigap purchase on the imputed medical expenditure, controlling for gender, age and State of residence. We find that those with higher predicted medical expenditures are less likely to purchase Medigap insurance, and this negative correlation is statistically significant, confirming advantageous selection in Medigap previously documented for the MCBS data. We then gradually add variables that we suspect may contribute to the advantageous selection. Our key finding is that, controlling for risk aversion and risk aversion interacted with predicted riskiness of medical expenditures alone is not enough to overturn the negative coefficient estimate on predicted medical expenditure. Instead advantageous selection has multiple sources: besides the usual suspect of risk aversion, we find that cognition, income, education, and longevity expectation all contribute to advantageous selection. Once we condition on these factors, we find that those with higher predicted medical expenditure risks do buy more Medigap insurance.

The remainder of the paper is structured as follows. Section 2 reviews related literature; Section 3 provides a simple theoretical framework to illustrate the idea of advantageous selection; Section 4 provides some detailed background about Medicare and Medigap insurance markets; Section 5 describes the MCBS and HRS data sets used in our empirical analysis; Section 6 details the evidence of advantageous selection from MCBS data; Section 7 examines the sources of advantageous selection by combining MCBS and HRS data sets; Section 8 briefly discusses the conditions under which our estimates also provide a lower bound on the magnitude of moral hazard in Medigap insurance; and finally, Section 9 concludes.

2 Related Literature

Finkelstein and McGarry (2005). Our paper is most closely related to a recent paper by Finkelstein and McGarry (2005) that studies selection based on multi-dimensional private information in the long-term care (LTC) insurance market. Using the Aging and Health Dynamics Survey (AHEAD) they found a negative, though statistically insignificant, correlation between LTC coverage in 1995 and use of nursing home in the period between 1995-2000, even controlling for insurance companies' prediction of risk type. However, they also found that a self-reported probability assessment in the 1995 survey "What do you think are the chances that you will move to a nursing home in the next five years?" positively predicts both LTC coverage and nursing home use in 1995-2000, even after controlling for insurance companies' risk type assignment. They further find those who undertake potential preventive health care, which they use as proxy for risk preferences, are more likely to own LTC insurance and less likely to enter a nursing home.[7]

Our paper complements Finkelstein and McGarry (2005) by investigating advantageous selection in a different market with importantly different features. The LTC insurance market is relatively small – 10 percent of the elderly in the AHEAD had LTC insurance – and the finding of advantageous selection in LTC insurance presents a new puzzle: if, on net, selection is advantageous why is the market so small?[8] This puzzle motivates the examination of a market, like Medigap insurance, that is large, and yet has the potential for substantial asymmetries of information in favor of the insured. Specifically, at least half of our sample has a Medigap policy, and yet (as noted above) regulation effectively precludes price discrimination on the basis of health risk. Our finding of advantageous selection in the Medigap market reconciles these two otherwise competing facts.

Our paper also complements Finkelstein and McGarry (2005) by examining directly several possible sources of advantageous selection. Rather than using behavioral proxies for risk aversion, we exploit the rich and direct information in the HRS on objects of special theoretical interest such as risk attitudes, longevity expectations, planning horizons, financial cognition, etc. to shed light on the channels through which advantageous selection emerges in the Medigap market.

Finally, Medigap market has the virtue that the relevant source of selection, health expenditure risk, is relatively easily measured with available data. The MCBS, especially, offers comprehensive health expenditure data for the entire relevant age range (age 65 to death), and health expenditure risks are realized throughout this age range. LTC expenditure risk, on the other hand, usually occurs in just a single episode, often leading into the moment of death. For this reason, the best measure of *ex post* risk for LTC insurance is whether one *eventually* uses nursing home. The AHEAD data cover a rather old cohort, the sample represents the population born before 1924 and was thus at least 76 by the year 2000; but many are quite far from their time of death.[9] As a result the measure of *ex post* risk Finkelstein and McGarry (2005) use – use of nursing home between

[7]The potential preventive health care measures they include are: whether the individual had a flu shot, had a blood test for cholesterol, checked her breast for lumps monthly, had a mammogram or breast x-ray, had a Pap smear and had a prostate screen.

[8]Brown and Finkelstein (2004) argued that supply-side imperfection can not explain the limited size of LTC insurance market.

[9]According to life-table estimates, those alive at age 75 in 2000, had on average 11.4 more years to live. (Vital Statistics, 2004).

1995 and 2000 – may be a relatively noisy measure.

Literature on Adverse Selection in Medigap. Our paper is also related to a sizeable literature that looks for evidence of adverse selection in the Medigap insurance market. Wolfe and Goddeeris (1991) used data from Retirement History Survey (RHS) – a longitudinal survey conducted by Social Security Administration between 1969 and 1979 on recent retirees – to examine the moral hazard and adverse selection in Medigap insurance. The health expenditure variable they used was self-reported by survey respondents about their total medical bills for hospital, physician and prescription expenditures, including any amount paid by insurance. They found that despite the fact that respondents with better self-reported health were more likely to purchase supplemental private insurance, those with private insurance incurred higher, though statistically insignificant, expenditures from hospital stays, physician care and prescription drugs.

Hurd and McGarry (1997) used the first wave of AHEAD data to examine how health insurance influences the use of health care service by the elderly. They found that those who are the most heavily insured use the most health care services (mostly categorical answers of the number of times for hospital and doctor visits), after controlling for self-reported health indicators.[10] However, they also found little relationship between observable health measures and either the propensity to hold or to purchase health insurance, indicating little importance for adverse selection.

Ettner (1997), using MCBS 1991, found little evidence by health status in the probability of purchasing private insurance.[11] She also found that Medicare beneficiaries with individually purchased policies had higher total (and Part B) Medicare and physician expenditures than those with employer-provided policies, even after controlling for observable differences. However, Ettner's (1997) finding can be interpreted as evidence of adverse selection only under the assumption that "employment" for this age group is exogenous with respect to health, which we view as implausible. Moreover, those with employer-sponsored health insurance may have different rules from Medigap regarding whether Medicare is the primary coverage. Such issues are important because she examined only Medicare reimbursed expenditure.

Khandker and McCormack (1999), using MCBS 1991 and 1993, found that those with supplemental private insurance were more likely to use and incur higher level of *Medicare* spending, particularly Part B services. However, because of the nature of Medigap insurance (in that its

[10] They did not report results without controlling for measured health.

[11] Lillard and Rogowski (1995), using Panel Study of Income Dynamics (PSID), also found little evidence of adverse selection for supplemental insurance.

coverage is intimately related to the deductibles and co-payments of Medicare), the narrow focus on Medicare spending will mechanically lead to a positive correlation between Medigap coverage and Medicare spending. Indeed Khandker and McCormack (1999) argued that a better measure for the health expenditure to address adverse selection would have been total medical expenditure, not just medical expenditure reimbursed by Medicare. Unfortunately, the waves of the MCBS data used in their analysis only contained Medicare claim data, but did not contain information about total health cost, including beneficiary out-of-pocket cost as well as expenses paid by supplementary insurers.[12]

To summarize, while the empirical literature has demonstrated a correlation between insurance coverage and service use, it did not provide consistent evidence of the role for adverse selection in the purchase of supplemental insurance.

Other Related Literature on Adverse Selection in Insurance Markets. Despite the failure to find positive association between insurance coverage and *ex post* risk, there is a literature that has supported the presence of adverse selection when one focuses on the choice of contractual forms such as deductibles and co-payments etc. For example, Puelz and Snow (1994) argued, in the context of automobile collision insurance, that in equilibrium of models with adverse selection, one should expect to see that individuals with lower risk will choose a contract with a higher deductible, and contracts with higher deductibles should be associated with lower average prices for coverage. They found evidence in support of these predictions using individual data from an automobile insurer in Georgia.[13] More recently, Finkelstein and Poterba (2004) use a unique individual level data set of annuities from U.K. and find systematic relationships between *ex post* mortality and annuity characteristics, such as the timing of payments and the possibility of payments to the annuitants' estate. These patterns are consistent with the presence of asymmetric information, despite the lack of evidence of substantive mortality differences by annuity size.

[12]We are grateful to Tami Swenson for the clarification on these data issues.

[13]However, see Chiappori and Salanié (2000) and Dionne, Gouriéroux and Vanasse (2001) for discussions about the flaws in Puelz and Snow study. Both papers emphasized the importance to include flexible interaction terms in the empirical analysis.

3 Multi-dimensional Private Information and Advantageous Selection

The idea that if individuals select on multidimensional private information, then it is possible for the correlation between *ex post* risk occurrences and coverage may be negative, zero, or positive is now well understood from the work of Henmenway (1990, 1992), de Meza and Webb (2001), Araujo and Moreira (2001) and Jullien, Salanié and Salanié (2005). These papers all focused on private information of risk aversion as the source of advantageous selection. In this section, we first illustrated this idea using risk aversion as a partial equilibrium example; then, anticipating our main empirical focus, we will expand our definition of advantageous selection to include selection based on other unobservable private information.

Risk Aversion as the Source of Advantageous Selection Consider an individual with constant relative risk aversion utility function

$$u(y) = \frac{y^{1-\gamma}}{1-\gamma},$$

where γ is the relative risk aversion parameter. Suppose that she has wealth $Y > 0$, but faces a risk in health expenditures in that she has to incur a health expenditure (over and above what is covered by the Medicare) of $L > 0$ with probability $p \in [0,1]$.[14] The individual can choose to purchase Medigap insurance at a price m that will reduce the out-of-pocket expenditure to $\tilde{L} < L$. The individual decides whether to purchase Medigap. Her expected utilities from purchasing and not purchasing Medigap insurance are respectively given by

$$\begin{aligned} V_B(p,\gamma) &= pu\left(Y - m - \tilde{L}\right) + (1-p)\,u(Y-m) \\ V_N(p,\gamma) &= pu(Y-L) + (1-p)\,u(Y). \end{aligned}$$

We assume this individual will purchase Medigap according to the Logit probability

$$Q(p,\gamma) = \frac{\exp[V_B(p,\gamma)]}{\exp[V_B(p,\gamma) + V_N(p,\gamma)]} \tag{1}$$

Simple algebra shows that $Q(p,\gamma)$ is increasing in p and γ. That is, more risky and more risk averse individuals are more likely to purchase Medigap insurance.

[14] We assume away the price effect (also called moral hazard as in Cutler and Zeckhauser 1999) by assuming that the expenditure level L does not depend on the health insurance status.

Now suppose that in the population there is a joint distribution over individuals' private types (p, γ) given by F, and let the CDF of risk aversion conditional on risk type p be $F_{\gamma|p}(\cdot|\cdot)$.[15] If we do not control for risk aversion γ and look at only the relationship between risk-type p and the probability of purchasing Medigap, we will have

$$\tilde{Q}(p) = \int Q(p, \gamma)\, dF_{\gamma|p}(\gamma|p). \tag{2}$$

If p and γ are negatively correlated, then $\tilde{Q}(p)$ may or may not increase in p.

We can also compare the average accident rate of health risk for those with and without Medigap insurance. The average accident rate for those with Medigap insurance is given by

$$A_B = \frac{\int Q(p, \gamma)\, p dF(p, \gamma)}{\int Q(p, \gamma)\, dF(p, \gamma)}, \tag{3}$$

where the denominator is the measure of individuals who purchase Medigap, and the numerator is the expected number of health risks that occur to those who purchase Medigap. Similarly the average accident rate of health risk for those without Medigap insurance is

$$A_N = \frac{\int [1 - Q(p, \gamma)]\, p dF(p, \gamma)}{\int [1 - Q(p, \gamma)]\, dF(p, \gamma)}. \tag{4}$$

Chiappori and Salanié's (2000) test for asymmetric information is a test of whether $A_B > A_N$. However, if p and γ are negatively correlated, it is possible that $A_B \leq A_N$ despite the presence of asymmetric information.

Note the above example is merely illustrative, because we only analyzed individuals' insurance purchase decision given their types assuming a particular equilibrium, and did not analyze the full equilibrium in which insurance companies compete by offering insurance contracts.[16] However,

[15] Suppose that in equilibrium insurance companies offer a contract $\left\langle m, L - \tilde{L} \right\rangle$ where m is the insurance premium and $L - \tilde{L}$ is the payout in the event that the health risk occurs. Then the expected profit for the insurance company is given by

$$\Pi\left(m, \tilde{L}\right) = \int Q(p, \gamma)\left[m - p\left(L - \tilde{L}\right)\right] dF(p, \gamma).$$

For $\left\langle m, L - \tilde{L} \right\rangle$ to be a competitive equilibrium, it has to satisfy $\Pi\left(m, \tilde{L}\right) = 0$. Moreover, no insurance company can find a profitable deviation to another contract that yields positive profit. As we mentioned, this is not important for our empirical analysis.

[16] We refer the readers to Julien, Salanié and Salanié (2005) for a formal analysis of the equilibrium analysis when the insurance market is not competitive. Chiappori, Julien, Salanié and Salanié (2005) showed that, when the insurance market is competitive, then a suitable modified version of the positive correlation property still holds even with multi-dimensional private information.

this simple example captures the idea that, when individuals differ in both their risk type and risk aversion, and if there is a negative correlation between risk aversion and risk, then it is possible that in equilibrium, those who purchase more insurance coverage will on average have lower risk than those who purchase less coverage.

Sources of Advantageous Selection: General Discussion While the above illustration showed explicitly how heterogeneous risk aversion γ that is agents' private information can be a source of advantageous selection for the insurance companies, we will now expand this concept. For this purpose, we will again denote p as the probability of health risk, but we will interpret the variable γ as any other private information possessed by the agents that may affect their probability of purchasing Medigap $Q(p,\gamma)$. Now instead of deriving $Q(p,\gamma)$ explicitly as we did when γ is interpreted as risk aversion, we will take the probability of Medigap purchase $Q(p,\gamma)$ as the reduced form entity of focus. Viewed in this perspective, we can state the general properties for some private information γ to act as a source of advantageous selection as follows:

Property 1: γ is positively correlated with insurance coverage, i.e., $Q(p,\gamma)$ is increasing in γ;

Property 2: γ is negatively correlated with risk p.

Under these two conditions, the average probability of insurance purchase for a given risk type p, namely $\tilde{Q}(p)$ as defined in (2), may not be monotonic in p; and the ranking over A_B and A_N, defined respectively by (3) and (4), can go either way.

In our empirical analysis, we first provide, in Section 6, evidence that is akin to "$A_B < A_N$," that is, the health risk occurrence for those with Medigap insurance is lower than those without Medigap insurance, thus suggesting the existence of some sources of advantageous selection. Then in Section 7 we examine the sources of advantageous selection, that is, look for elements of γ that may be contributing to the earlier finding that $A_B < A_N$.

4 Background on Medicare and Medigap

4.1 Medicare

Medicare is the primary health insurance program for most seniors in the United States. All Americans age 65 and older who have, or whose spouses have, paid Medicare taxes for more than

40 quarters are eligible. The Original Medicare Plan consists of two programs.[17] Medicare Part A (Hospital Insurance Program) covers inpatient hospital, skilled nursing facility, and some home health care. For hospital stays in each benefit period Medicare pays all covered costs except the Medicare Part A deductible (which varies by year and is equal to $912 in 2005) during the first 60 days and coinsurance amounts for hospital stays that last beyond 60 days (in 2005, the coinsurance amount is equal to $228 per day for days 61-90, and $456 per day for days 91-150). Hospital stays beyond 150 days are not covered at all by Medicare Part A. For Skilled Nursing Facility Care, the coinsurance amount is about $114.00 per day for days 21 through 100 each benefit period; and no coverage is provided beyond the 100th day in the benefit period. Almost all retirees are automatically enrolled in Medicare Part A when turn 65 and there are no premiums paid for this coverage.

Medicare Part B (also called Medicare Insurance) covers Medicare eligible physician services, outpatient hospital services, certain home health services, and durable medical equipment. Part B enrollees have to pay a monthly premium ($66.60 in 2004). Almost all people choose to enroll in Part B when they turn 65; indeed they are automatically enrolled when the turn 65 if they have previously applied for Social Security Old Age Benefits. Under Part B, individuals are responsible for $110.00 deductible in 2005 and are responsible for a 20% coinsurance payment for all Medicare-approved services after exceeding the deductible.

4.2 Medigap

As is clear from above, Medicare leaves retirees at significant risk of health care expenditures. To insure Medicare beneficiaries against some of that risk, private insurance companies sell "Medigap" policies. Medigap policies cover some of the coinsurance, deductibles and uninsured expenses, i.e. the gaps, in the Original Medicare Plan.[18] Since 1990, by Federal Law, Medigap policies have been standardized into ten plans, "A" through "J," each representing a different constellation of benefits. The basic plan, Plan A, covers all coinsurance payments for hospital stays longer 60 days, and all

[17] For details, see Centers for Medicare and Medicaid Services (2005), page 55-64. The Original Plan is available everywhere in the country. Some areas also offer what are now called Medicare Advantage Plans, which are largely managed care, and preferred provider organization plans. In 2001, approximately 15% of Medicare beneficiaries were enrolled in what was then the equivalent of an Advantage Plan.

[18] Those who choose Medicare Advantage Plans receive similar Medicare gap coverage. For this reason, those with Medicare Advantage Plans are discouraged, though not precluded, from purchasing Medigap policies.

coinsurance payments from Medicare part B (except the deductible).[19] All other plans offer these basic benefits, and more. Plan B, for example, also covers the deductible ($912 in 2005) for hospital stays shorter than 60 days; Plan C, which is the most popular, adds skilled nursing coinsurance, Medicare Part B deductible and foreign travel emergency coverage. Plan J adds to this, among other things, extended drug benefits with a $3000 annual limit in 2004. While not all Medigap policies are offered in every state, almost every state has a provider which offers the basic plan.[20] If an insurer offers any Medigap policy, by law it must offer the basic plan.

In addition to being regulated with respect to quality, Medigap pricing and coverage are regulated in ways that tend to amplify the asymmetries of information favoring the insured. Most importantly, Medigap policies are required by law to have an open enrollment period. For six months after the first day of the first month an individual is both age 65 or older and enrolled in Medicare Part B, insurers cannot deny Medigap coverage, delay coverage, or price coverage based on pre-existing conditions.[21] Instead, during open enrollment, insurers effectively price only on age, gender and state of residence.[22] Moreover, these insurance policies are required by law to be guaranteed renewable. That is, beneficiaries may not be dropped from policies so long as they continue the timely payment of the contracted premiums. The combination of these pricing and offering regulations thus make the potential for asymmetries of information favoring would-be beneficiaries especially high.

[19]These basic benefits also pay for the first three pints of blood, which Medicare does not cover.

[20]The exceptions are Massachusetts, Minnesota and Wisconsin which have received waivers that allow them to offer somewhat different standardized plans.

[21]If the policy holder was previously uninsured, but received a diagnosis of, or treatment for, a condition during the six months prior to the Medigap policy's starting date, the insurer may refuse to cover expenditures associated with that condition for a waiting period of up to six months.

[22]Some Medigap insurance companies offer menus of policy options that may help to discriminate among those with varying health risks. To our knowledge, the pricing comes in only three forms: 1) "age-issued policies" that have a flat premium that depends only on inflation and the age at which the policy was purchased; 2) "age-attained policies" that have a premim that starts lower than the age-issued policies but rises on a predictable schedule as the beneficiary ages; and 3) "community rated," whose premiums do not depend either on age of purchase or age attained.

5 Data

5.1 Medicare Current Beneficiary Survey (MCBS)

Our analysis relies on two large data sets, the MCBS and HRS. The MCBS was first conducted from September through December 1991 and is a continuous, rotating panel survey of a nationally representative sample of the Medicare population, conducted by the Office of Strategic Planning of the Centers for Medicare & Medicaid Services (CMS).[23] The central goals of MCBS are to determine expenditures and sources of payment for all services used by Medicare beneficiaries, including co-payments, deductibles, and noncovered services; to ascertain all types of health insurance coverage and relate coverage to sources of payment; and to trace processes over time, such as changes in health status, spending down to Medicaid eligibility, and the impacts of program changes.

MCBS is unique in covering the entire Medicare population, whether aged or disabled, living in the community or in institutions; oversampling significant subpopulations; and following and reinterviewing the sample to obtain a continuous longitudinal picture. Other features include the collection of a wide variety of data on each sample person, including topical supplements; combining survey and administrative data; and being able to retrieve data to respond to urgent Medicare policy issues. Beneficiaries sampled from Medicare enrollment files (or appropriate proxies) are interviewed in person, three times a year using computer-assisted personal interviewing.

Important for our purposes, MCBS interview data include information about whether the respondent is also covered by medical insurance that they purchased themselves, and all of the survey data are linked to Medicare claims and other administrative data including buy-in status and capitated plan membership. The final file consists of survey, administrative, and claims data and thus provides a comprehensive view of respondents' heath care costs and use. In addition to information on the health and demographics of respondents, our focus is on the total health expenditure, i.e. the combined expenditures that were covered by Medicare, other public insurance, private insurance, or paid out-of-pocket. The Data Appendix provides detailed explanation about how the variables used in our analysis are constructed.

Following our basic empirical strategy, we use MCBS data to estimate the relationship between *realized* total health expenditure and demographic, health and economic, including health insurance, variables. To form *expected* health expenditures for the HRS sample, we estimate this equation using

[23]See `http://www.cms.hhs.gov/mcbs/` for more details.

MCBS data on individuals without Medigap,[24] export the estimated coefficients to the HRS data and from those coefficients generate a predicted expenditure value $\left(\hat{E}_i\right)$ for each HRS respondent.

5.2 Health and Retirement Study (HRS)

The HRS began in 1992 as a panel survey of a nationally representative sample of those age 51 to 61, and their spouses, with oversamples of blacks, hispanics and residents of Florida. This original cohort of 12,654 respondents has since been interviewed every other year to the present; and in 1998 the sample was supplemented with both somewhat older and younger cohorts. Our interest in those age 65 and older with information on risk attitudes and planning horizons leads us to limit our analysis to decisions of those from the original HRS cohort assessed in 2000 and 2002, the latest year for which a final version of the HRS data is available.

The HRS is particularly well-suited to a study of advantageous selection in Medigap insurance. It contains detailed information about the current and past health status of respondents along with rich data on their insurance choices and costs. The health information includes both self-reported health and more objective measures such as diseases diagnosed and activities that the respondent has difficulty performing.[25] The insurance data include information on where the insurance was acquired, its premiums, and its coverage. A detailed description of the health and insurance information we use is provided in the Data Appendix. The health and insurance data are combined with high quality information about economic and demographic variables, including education, income, wealth and cognition. In addition, the HRS is distinctive in its attention to variables central to economic theory. These variables include measures of risk preferences, expectations, and financial planning horizons. The following, describes these theoretically important measures in greater detail.

5.2.1 Measures of Risk Preferences

As described above, risk preferences are central to existing theories of advantageous selection. Beginning in the first wave, the HRS asked (subsamples of) respondents a series of questions

[24]To provide robustness checks we also experiment with other samples. See Section 7.1 for details about the pros and cons of the different imputation procedures.

[25]It is important that the health information components of HRS and MCBS contain very similar, though not identical, set of questions. This obviously plays an important role in the imputation method we will describe in Section 7.1.

regarding their risk attitudes. In Wave 1 (1992) all respondents are first asked the following question:

> "Suppose that you are the only income earner in the family, and you have a good job guaranteed to give you your current (family) income every year for life. You are given the opportunity to take a new and equally good job, with a 50-50 chance it will double your (family) income and a 50-50 chance that it will cut your (family) income by a third. Would you take the new job?"

If the answer to the first question is "yes", the interviewers will continue with the following question:

> "Suppose the chances were 50-50 that it would double your (family) income, and 50-50 that it would cut it in half. Would you still take the new job?"

If the answer to the first is "no," the interviewers continues with the following question:

> "Suppose the chances were 50-50 that it would double your (family) income and 50-50 that it would cut it by 20 percent. Would you then take the new job?"

The responses to these questions place respondents into four, ordered risk categories: I (unwilling to risk any income cuts) through IV (willing to risk a 50% cut in income). In Wave 2, a randomly selected sub-sample answered the same sequence of questions, now supplemented to include jobs with downside risks of 10% and 75%.

Assuming that an individual's responses to these hypothetical income gambles are error-prone reflections of his fixed, constant relative risk aversion preferences, Kimball, Sahm and Shapiro (2005) estimate the risk tolerance for each respondent in the HRS by maximum likelihood. In our analysis, we take as our measure of risk preferences Kimball et al.'s risk tolerance estimates (see their Table 6) which treat responses in 1992 and 1994 as multiple indicators of the same stable preference.

5.2.2 Expectations and Planning Horizons

Logically, longevity expectations should also play an role in determining health insurance choices, though the net effect of a higher expectation for longevity has a theoretically ambiguous effect on investments in health. Those who expect to live longer may want to spend more now on their health as such investment will pay dividends for longer.[26] On the other hand, the marginal value of a current health investment may be lower when a long life already seems likely. The HRS collects detailed information about longevity expectations. Our focus is on the response to the question, asked of all respondents age 65 and younger, and repeated in every HRS wave,

> "What is (percent chance) you will live to 75 or more?"

In our analysis, we use the most recent available response to this question as our measure of longevity expectations.[27]

Like expectations for longevity, the length of financial planning horizons (which presumably reflect both uncertainty and subjective time discounts) may influence insurance choices. Here, however, theoretical effect of a longer planning horizon seems unambiguous, those with longer horizons would be more willing to pay smaller immediate costs (premiums) to avoid larger expected future costs. The HRS also collects information on financial planning horizons. Specifically, respondents in Wave 1 were asked:[28]

> "In deciding how much of their (family) income to spend or save, people are likely to think about different financial planning periods. In planning your (family's) saving and spending, which of the time periods listed in the booklet is most important to you [and your (husband/wife/ partner)]? 1. Next few months, 2. Next year, 3. Next few years, 4. Next 5-10 years, 5. Longer than 10 years."

We use indicator variables for each of these four responses in Wave 1 as our measures of the respondent's financial planning horizon.

5.3 Medigap Insurance Status

Both the MCBS and the HRS contain detailed information about respondents' health insurance choices. Specifically, each data set indicates whether the respondent is covered by Medicare, parts

[26] This assumes that, conditional on health, those with insurance use more health services.

[27] Like the risk preference questions, this measure of longevity expectations is likely measured with error, and may reflect both beliefs about longevity and the degree of certainty about those beliefs. Evidence consistent with both error and uncertainty about beliefs is found in the heaping of responses around focal response such as zero, fifty and, to a lesser extent, one hundred. See Kezdi and Willis (2005) for a thorough discussion of these measures.

[28] This question was also asked, at random, of one out of ten respondents in Waves 4 and 5, but was not.not asked of anyone age 65 and older in Wave 6. Of the 11,626 respondents who answered this question in Wave 1, just 821 answered it again in Wave 4 and 941 in Wave 5.

A and B, and whether that coverage is provided by a Medicare Advantage Plan (HMO/PPO). Our goal is to identify those respondents who are covered Medicare parts A and B and who, if they did not buy a Medigap policy, would either not be covered by another private insurance plan, or would have to pay substantial premiums to obtain that coverage.

More specifically, we construct for the empirical analysis presented below two indicator variables for Medigap status. The first equals one if the respondent is covered by Medicare and has purchased additional private insurance that is secondary to Medicare. Only those who have no additional insurance beyond Medicare are coded as having no Medigap. In particular, those covered by any employer-provided health insurance, Medicaid or other government insurance, are treated as missing. The assumption is that these alternative forms of insurance are providing either higher quality or a lower price (or both) than that offered by Medigap. The second medigap status variable we construct equals one if the first variable equals one, *or* if the respondent is covered by Medicare, purchases private insurance from any source, and pays more than $500 per year in premiums for that insurance. As with the first medigap status variable, those covered by any employer-provided health insurance for which they pay less than $500 per year, and those covered by Medicaid or other government insurance are treated as missing. With this second definition of medigap status we seek to capture those who pay a substantial amount for employer-provided health insurance that is secondary to Medicare. In subsequent sections we separately report the results under the two different definitions of Medigap status, and it is important to preview that none of the results, either qualitatively or quantitatively, depends on which definition of Medigap status we use.

6 Evidence of Advantageous Selection

In this section we present results from a set of simple regressions to argue that they jointly provide strong evidence of advantageous selection in the Medigap insurance market. That is, those who purchase Medigap seems to be healthier than those without Medigap in terms of their realized total medical expenditure. We also present direct evidence that the more healthier people are more likely to purchase Medigap insurance, conditional on observables that determine the prices.

6.1 Basic Regression Results: Indirect Evidence of Advantageous Selection

Tables 1 reports regression results of "Total Medical Expenditure" on Medigap status, gender, a third-order polynomial of age, and controls for State and year. The dependent variable "Total

Medical Expenditure" we use corresponds to variable `pamttot` in the MCBS data. This variable is constructed by CMS from a variety of sources, including the Medicare administrative records and survey responses and sometimes uses imputation when missing values are encountered.[29] The key point is that, in calculating `pamttot`, CMS include, for any event, payments from 11 potential sources of payment: Medicare fee-for-service, Medicaid, Medicare managed care, private insurance managed care, Veterans' Administration, employment-based private health insurance, individually purchased private health insurance, private insurance with unknown sources, out-of-pocket, Uncollected liability, and other public insurance. Thus the variable `pamttot` is as close as possible to measure the total health expenditure from all sources.

[Table 1 About Here]

Panels A and B of Table 1 differ only in the definitions of "Medigap" as we explained in Section 5.3. Each panel contains three columns depending on whether the results are for the full sample, or separately for male and female samples. The column labeled "All Sample" in each panel shows that, those with "Medigap" on average spend more than $4,000 less than those without "Medigap" coverage, if we do not directly control for health. It is interesting to note that the negative correlation between "Medigap" and total medical expenditure is stronger for the female sample (about $6,000) than for the male sample (about $2,000). Total medical expenditure is higher for older individuals, as expected. It is important that, due to the government regulations of the pricing of the Medigap private insurance, the control variables Female, Age polynomial and State are meant to capture variation in the pricing of Medigap.[30] Of course, to the extent that gender and age may partly predict health, the regressions also partly control for health.

It is important to emphasize that in Table 1 (and subsequent tables as well) we reported our results separately for male and female samples; none of the qualitative statements differs by gender and though the quantitative magnitudes do.[31] Reporting estimation results separately is important in the light of the critiques raised by Dionne, Gouriéroux and Vanasse (2001) in their critique of Puelz and Snow (1994). Dionne, Gouriéroux and Vanasse (2001) argued that Puelz and Snow's finding of adverse selection in automobile insurance market resulted from over-simplistic

[29] See MCBS public use documentation on "Cost and Use" Sections 3 and 5 for more details. This documentation is available online at `http://www.cms.hhs.gov/apps/mcbs/`

[30] To repeat what we described in the introduction, Medigap insurance premium an individual faces depends almost exclusively on his/her state of residence, age and gender (see Robst 2001).

[31] Indeed, the evidence seems to suggest that advantageous selection is quantitatively larger among females.

functional form assumptions in their estimating equation, and once additional interaction terms of the variables used in Puelz and Snow's analysis are included, the finding of adverse selection disappears (this point was also raised in Chiappori and Salanié 2000, see Footnote 13 for more details). Reporting results separately for male and female samples is equivalent to interact all the terms in the regression with gender.[32]

[Table 2 About Here]

Table 2 reports of results from regressions analogous to those in Table 1, with the addition of controls for health variables. The difference between specifications (1)-(3) and (4)-(6) is whether demographics are included as conditioning variables. The health variables and demographic controls are detailed in the Data Appendix under the category "Health" and "Demographics" respectively. Note that, once controlling for health, those with "Medigap" spend about \$1,900 (\$1,700 respectively) *more* than those without "Medigap" if we do not include (respectively, include) other demographic controls. The positive association between "Medigap" and total medical expenditure seems to be stronger for males (about \$2,300) than for females (about \$1,500).

A legitimate concern for the results in Table 2 is whether the contemporaneous health variables we used in our analysis could be an *ex post* measure of health conditions that may be affected by the Medigap insurance status.[33] While most of health measures in the health survey instruments of both MCBS and HRS are probably not subject to this problem, nonetheless this concern can be addressed by exploiting the panel nature of the MCBS data. Indeed we can use the lagged health conditions, instead of the contemporaneous health conditions, as the explanatory variables. The quantitative results are almost identical, which indicate, not surprisingly, that there is a very strong correlation between lagged health and contemporaneous health.

[Table 3 About Here]

Table 3 reports results from regressions analogous to those in Table 2, except that we summarize the list of health variables by five health factors. Note that the coefficient estimates for "Medigap" are qualitatively not changed from those in Table 2 when we use only five health factors to summarize all the health control variables.

[32]Since a third order polynomial in age is already included, the only interaction we did not include is the interaction of States with age.

[33]For example, if those with Medigap are more likely to get diaganostic care, then the contemporaneously reported health problems may not be the relevant private information in agents' decision to purchase Medigap.

These factors have rather interesting interpretations. Looking at the factor loadings (not reported), we can give pretty clear interpretations of these factors.[34] Factor 1 can be interpreted as a "Not Reported" factor, which loads heavily on variables that are indicators of non-report. Factor 2 loads negatively on self-reported health and difficulties in instrumental activities of daily living (IADLs), thus is an important unhealthy factor. Factor 3 leads positively on self-reported health and negatively on measured medical conditions in the past two years, and thus is a healthy factor. Factor 4 loads positively on self reported health and self-reported health changes in the last year, which represents the part of self-reported health that are not captured in Factor 2 and 3. Factor 5 is more or less noise.

The only way to rationalize the different coefficient estimates of "Medigap" in Table 1 and Tables 2-3 is that those who purchase "Medigap" are actually healthier than those who do not purchase "Medigap," which is what we have called "advantageous selection."

6.2 More Direct Evidence of Advantageous Selection

Indeed, Table 4 reports the partial correlation between "Medigap" coverage and the health factors, conditional on gender and age. As before, Panel A and B respectively present results under the two alternative definitions of "Medigap." Columns labelled "EXP" simply report the regression coefficients for the factors from specification (4) in Table 3. These coefficient estimates inform us whether the factor is a "healthy" factor or an "unhealthy" factor: those factors with large positive and significant coefficient estimates are the unhealthy factors and those with large negative and significant coefficient estimates are the unhealthy factors. It is important to note that the health factors for the three samples "All", "Female" and "Male" are separately estimated, and as a result, the factors for the three samples are actually different factors. Columns labelled "PCORR" report the partial correlation. For most part, Table 4 reveals a striking pattern: important unhealthy (healthy, respectively) factors tend to have a negative (positive, respectively) and significant partial correlation with "Medigap" coverage. For example, in Panel A, Column "All" shows that factor 2 and 3 are respectively the most important unhealthy and healthy factors; and factor 2 has a sizeable negative correlation with "Medigap" with p-value of almost 0; while factor 3 has a sizeable positive correlation with "Medigap." The factors that have the "wrong" correlation sign with "Medigap" are typically of two kind: either the factor itself is not very important (with small and insignificant coefficient estimates); or the partial correlation is statistically insignificant (with large p-values).

[34]The factor loadings are not reported to save spaces, and they are available from the authors upon request.

[Table 4 About Here]

6.3 Summary

To summarize, Table 1 showed that there is a negative correlations between Medigap coverage and "Total Medical Expenditures," if we only control for factors that determine Medigap prices. Table 1 itself suggests multidimensional private information. The reason is simple: if there is no asymmetric information, then the insurance companies would be able to price the insurance contracts to capture the health risk differences with only functions of age, gender and State, and on average, Medigap coefficient should be zero; if there is only one dimensional private information about risk type, then Julien, Chiappori, Salanié and Salanié's (2005) analysis would have predicted a robust positive correlation between Medigap coverage and "Total Medical Expenditures." Moreover, the multidimensional private information must work in a way such that healthier people are more likely to buy Medigap coverage, and this advantageous selection must be strong enough to overcome the natural *ex post* "moral hazard" (or price effect) of medical care usage. Table 2 and 3 confirm this connection because they establish that, indeed, after controlling for observable health, we do find strong positive association between Medigap coverage and *ex post* health expenditure, that is, moral hazard in medical care usage does exist.[35] From the two sets of results, we can thus conclude that, there is multidimensional private information, and the other dimensions of private information, regardless of what it is, work to create a phenomenon that has been termed "advantageous selection," namely, healthier individuals are more likely to buy Medigap coverage.

Despite the apparent similarity between our conclusions so far and those in Finkelstein and McGarry (2005), it is useful to highlight the differences in the inferential process through which we reached these conclusions. As we mentioned in Section 2, Finkelstein and McGarry's (2005) inference about multidimensional private information runs in two steps. First, they demonstrate that individuals have private information about their risk type, based on the finding that a self-reported probability assessment in the 1995 survey "What do you think are the chances that you will move to a nursing home in the next five years?" positively predicts both LTC coverage and nursing home use in 1995-2000, even after controlling for insurance companies' risk type assignment. Second, they show that despite this private information, they found an almost-zero correlation between insurance coverage and the use of long-term care predicted by uni-dimensional models

[35] See Section 8 for more discussions about the implications of our findings for the magnitude of moral hazard in Medigap insurance market.

of asymmetric information. Thus they conclude that there must be a second form of unobserved heterogeneity. Because zero-correlation between insurance coverage and the use of long-term care is consistent with both no asymmetric information and multidimensional private information, their first step, the step that directly establishes the existence of private information, is crucial.[36]

7 Sources of Advantageous Selection

In Section 6, we used MCBS data to provide evidence of advantageous selection: healthier people are more likely to purchase "Medigap" coverage than those who are less healthy. Moreover, we also find evidence consistent with important moral hazard effects: after controlling for observable health, those with "Medigap" actually spend more than those without "Medigap" (see Section 8 for more discussions on moral hazard). Importantly, the magnitude of the advantageous selection into "Medigap" is strong enough to dominate the moral hazard in the use of medical service.

In this section we will investigate the sources of advantageous selection. We seek to identify other dimensions of individuals' private information that satisfy the two properties we mentioned in Section 3: on the one hand, make them more likely to purchase Medigap, and on the other hand, is negatively correlated with their health risk.

7.1 Empirical Strategy

The ideal data set for our analysis of the sources of advantageous selection (in fact, for the whole project) would be the HRS data augmented by the administrative total medical expenditures from the links with Medicare. Unfortunately at this moment HRS is not yet properly linked to the Medicare administrative records, and has poor information about out-of-pocket spending estimates than MCBS. On the other hand, MCBS does not contain information about many of the usual suspect sources of advantageous selection. We now describe an empirical strategy that combines MCBS and HRS to examine the sources of the advantageous selection.[37]

[36] Because LTC insurance pricing was essentially unregulated in the period covered by Finkelstein and McGarry's (2005) data, their first step relies crucially that the risk type assignments used by the insurance companies optimally utilized all information the insurance companies possess about the insured.

[37] There is a quite sizeable literature on the empirical methods to deal with the incomplete data problem where an ideal data set that contains all of the relevant variables is unavailable. Several methods have been proposed to combine multiple data sets. The proposed methods typically vary by the context of the research and the assumptions one is willing to make about the behavioral model and the joint distributions of the unobservables across data sets. For

To illustrate our empirical strategy, it is useful to describe the contents of MCBS and HRS relevant to our empirical analysis in the following way. The data in MCBS can be written as

$$\{E_i, M_i, \mathbf{H}_i, \mathbf{D}_i\}_{i \in \mathcal{I}_{MCBS}}, \tag{5}$$

and the data in HRS is

$$\{M_j, \mathbf{H}_j, \mathbf{D}_j, \mathbf{X}_j\}_{j \in \mathcal{I}_{HRS}} \tag{6}$$

where $\mathcal{I}_{MCBS}$ and $\mathcal{I}_{HRS}$ denote the MCBS and HRS sample respectively. Note that the variables $\{M, \mathbf{H}, \mathbf{D}\}$ are common to both data sets while E, the total medical expenditure, only appears in MCBS; and $\mathbf{X}$, the list of variables that we think are potential sources of advantageous selection only appear in HRS.

Our empirical strategy is very simple: we propose to use the MCBS data to estimate prediction equations for total medical expenditure risk, both its mean and variance (or some other measures of medical expenditure risk, as we will describe in Section 7.4), for the HRS sample. It is important to emphasize that the predicted mean and variance of medical expenditures are supposed to serve as a bi-dimensional summary measure of health risks for each individual in the HRS, before they make Medigap purchase decisions.

Even though it is a straightforward idea, in actually implementing it we have to decide which sample in MCBS to use in estimating the prediction equations. Should we use only those without Medigap, or should we use all the sample? Given that we are examining selection problems, whether it is adverse or advantageous selection, in Medigap purchase, it is not clear which sample will provide the better basis for our imputation given that the purpose of the imputations is to provide a bi-dimensional pre-Medigap-purchase measure of health risks. We follow a practical strategy of

example, Angrist and Krueger (1992) dealt with the sample combination problem in instrumental variable estimation $\hat{\beta}_{IV} = (X'Z)^{-1} Z'Y$. They note that $\hat{\beta}_{IV}$ can be calculated as long as the researchers have the two moments $X'Z$ and $Z'Y$, and the moment $X'Y$ is not needed. Thus if two comparable data sets exist such that one data set has X and Z, while the other data set has Z and Y, then the two incomplete data sets are sufficient to estimate $\hat{\beta}_{IV}$. Arellano and Meghir (1992) proposed a model of female labor supply with on-the-job search and estimated their model by combining two data sets: the U.K. Family Expenditure Survey which contains information on income and expenditure and the U.K. Labor Force Survey which has data on hours and job search behavior. Their proposed estimation method relies heavily on the structure of their model. Recently Ichimura and Martinez-Sanchis (2004) proposed GMM method to deal with the problem of incomplete data. Their method again heavily relies on parametric structural assumptions and the assumption on the joint variations on the variables in the auxiliary data set. Ichimura and Martinez-Sanchis (2004) is also a good source for other related work in this area. It suffices to say that these existing methods do not apply for our purposes.

estimating the prediction equations in two ways, and show that our results are qualitatively robust to the sample we use in arriving at the prediction equations.

Prediction Equations Using MCBS Subsample with No Medigap Coverage. In the first method, we only use the subsample in MCBS with no Medigap coverage to estimate the mean and variance of medical expenditures. Suppose that the estimated mean and variance prediction equations are

$$\begin{aligned} \hat{E}_{i1} &= \hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}_1 \mathbf{H}_i + \hat{\boldsymbol{\alpha}}_2 \mathbf{D}_i, & (7) \\ \widehat{VAR}_{i1} &= \left(E_i - \hat{E}_{i1}\right)^2 = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1 \mathbf{H}_i + \hat{\boldsymbol{\beta}}_2 \mathbf{D}_i. & (8) \end{aligned}$$

We can then impute the mean and variance of medical expenditures for the HRS sample as follows: for each $j \in \mathcal{I}_{HRS}$, the imputed mean medical expenditure is

$$\hat{E}_{j1} = \hat{\alpha}_0 + \hat{\boldsymbol{\alpha}}_1 \mathbf{H}_j + \hat{\boldsymbol{\alpha}}_2 \mathbf{D}_j, \tag{9}$$

and the imputed variance of medical expenditure is

$$\widehat{VAR}_{j1} = \hat{\beta}_0 + \hat{\boldsymbol{\beta}}_1 \mathbf{H}_j + \hat{\boldsymbol{\beta}}_2 \mathbf{D}_j. \tag{10}$$

Prediction Equations Using the Whole MCBS Sample with Medigap Status Indicator. In the second method, we use the whole MCBS sample to estimate the mean and variance of medical expenditure with Medigap status indicator M_i. That is,

$$\begin{aligned} \hat{E}_{i2} &= \hat{\gamma}_0 + \hat{\gamma}_1 M_i + \hat{\boldsymbol{\gamma}}_2 \mathbf{H}_i + \hat{\boldsymbol{\gamma}}_3 \mathbf{D}_i, & (11) \\ \widehat{VAR}_{i2} &= \left(E_i - \hat{E}_{i2}\right)^2 = \hat{\xi}_0 + \hat{\xi}_1 M_i + \hat{\boldsymbol{\xi}}_2 \mathbf{H}_i + \hat{\boldsymbol{\xi}}_3 \mathbf{D}_i. & (12) \end{aligned}$$

We then impute, for $j \in \mathcal{I}_{HRS}$, the mean and variance for the HRS sample

$$\begin{aligned} \hat{E}_{j2} &= \hat{\gamma}_0 + \hat{\boldsymbol{\gamma}}_2 \mathbf{H}_j + \hat{\boldsymbol{\gamma}}_3 \mathbf{D}_j, & (13) \\ \widehat{VAR}_{j2} &= \hat{\xi}_0 + \hat{\boldsymbol{\xi}}_2 \mathbf{H}_j + \hat{\boldsymbol{\xi}}_3 \mathbf{D}_j. & (14) \end{aligned}$$

It is important to note that in the imputation equations (13) and (14), we are not using the actual Medigap coverage status M_j for the HRS sample. Thus the predictions above are for the mean and variance of medical expenditures without Medigap coverage; and this is, of course, what we need in order for the mean and variance to serve as a bi-dimensional summary of health expenditure risks individuals face when deciding whether to purchase Medigap.

Pros and Cons of the Two Imputations. We have to acknowledge from the outset that neither of the above imputation methods is perfect. If selection into Medigap were not an issue (of course this is an ostensibly erroneous assumption given our findings in Section 6), imputation based on MCBS subsample with no Medigap coverage would have been the theoretically correct imputation. However, given the selection problems, the imputation $\hat{E}_{j1}$ does not correspond to the theoretical notion of pre-Medigap-purchase mean health expenditure risk. If those who do not have Medigap coverage are systematically healthier on unobservable health than those with Medigap, then $\hat{E}_{j1}$ will be an underestimate of the expected medical expenditure for those in HRS who actually had Medigap. On the other hand, if those who do not have Medigap coverage are systematically unhealthier on unobservable health than those with Medigap (just as they are on observable health factors), then $\hat{E}_{j1}$ will be an overestimate of the expected medical expenditure for those in HRS who actually had Medigap.

The second imputation method has its own problems. First, note that including Medigap status indicator M_i in prediction equations (11) and (12) is necessary because otherwise we would be attributing the positive moral hazard effect (or price effect) of Medigap to the pre-Medigap-purchase expenditure risk. However, including Medigap status indicator M_i in the prediction equations is problematic because M_i is endogenous and presumably there are unobservable components of health that are not captured by the variables in $\mathbf{H}_i$.

Despite these problems, we hope that showing that our qualitative results below are robust to either of the above imputation methods provides credibility to our findings.

Identifying the Sources of Advantageous Selections. With the mean of the medical expenditure risk for the HRS samples $\hat{E}_{jk}, k \in \{1, 2\}$ imputed according to the above methods, we can now describe our approach to identify the sources of advantageous selection.

With the imputed $\hat{E}_{jk}$ and $\widehat{VAR}_{jk}$, our augmented HRS data can now be represented as:

$$\left\{M_j, \mathbf{H}_j, \mathbf{D}_j, \mathbf{X}_j, \hat{E}_{jk}, \widehat{VAR}_{jk}\right\}_{j \in \mathcal{I}_{HRS}}. \tag{15}$$

For expositional simplicity, we suppress the subscript k in $\hat{E}_{jk}$ and $\widehat{VAR}_{jk}$ that denotes the method of imputation in the discussion below, and we will report results separately for $k \in \{1, 2\}$.

We first regress

$$M_j = \delta_0 + \delta_1 \hat{E}_j + \boldsymbol{\delta}_2 \mathbf{D}_j + \varepsilon_j, \tag{16}$$

where, as before, the variables in $\mathbf{D}_j$ include a third order polynomial in age, gender and State of residence, in order to capture the pricing differences in Medigap insurance. As we will report below in Tables 6 and 7, and consistent with our finding in Section 6, we find negative and statistically significant estimate for δ_1, indicating evidence of advantageous selection in the purchase of Medigap in HRS: individuals with higher predicted expenditure is less likely to purchase Medigap after controlling for demographics that are allowed to price Medigap (gender, age and State).

Then we gradually add more controls from the list of variables contained in $\mathbf{X}_j$. We first add risk tolerance $risktol_j$ and risk tolerance interacted with $\widehat{VAR_j}$; then we add education, income, cognition, longevity expectation and financial planning horizon in order. While these variables may not have as a clear-cut theoretical link with the decisions to purchase Medigap as risk tolerance, they may reflect different parts of the decision process. For example, cognition will affect an individual's ability to think through the costs and benefits of Medigap insurance. Economists are not typically comfortable to think that cognition ability should be a determinant of such decisions, but such discomfort is more due to the fully-rational model that economists have been used to study.

In the end we will show that, if we estimate the decision rule for Medigap purchase controlling not only for $\mathbf{D}_j$ but also $\mathbf{X}_j$, the coefficient estimate for $\hat{E}_j$ will be positive. More specifically, if we estimate

$$M_j = \theta_0 + \theta_1 \hat{E}_j + \theta_2 risktol_j + \theta_3 \widehat{VAR}_j \times risktol_j + \boldsymbol{\theta}_4 \mathbf{X}_j + \boldsymbol{\theta}_5 \mathbf{D}_j + \varepsilon_j, \qquad (17)$$

$\hat{\theta}_1$ will be positive and significant. That is, once we control for $risktol_j$, $\widehat{VAR}_j \times risktol_j$ and $\mathbf{X}_j$, the decision to purchase Medigap insurance is indeed as predicted by insurance models with single-dimensional private information, namely, individuals with higher expected medical expenditure (or less healthy individuals) are more likely to purchase Medigap. It is in this sense we say that we have identified the sources of advantageous selection to include not only risk tolerance $risktol_j$, but also variables included in $\mathbf{X}_j$, namely education, income, cognition, longevity expectation and financial planning horizon.

Comments About Our Empirical Strategy. We would like to make three comments about our empirical approach. First, it is essential to remember that the imputation of expected medical expenditure for the HRS sample from a prediction equation using MCBS data is meant to serve as a uni-dimensional summary measure of health expenditure risk prior to their Medigap purchase decisions. An alternative strategy would have been to directly use the health variables in HRS $\mathbf{H}_j$ and use factor analysis to obtain a small number of factors for health. The advantage of this health-factor approach is that no imputations would be necessary and thus the problems associated with imputation would have been irrelevant. The disadvantage of this approach (and the reason we did not adopt it) is that typically there are at least about four independent and important health factors (as we saw in Section 6), thus it is difficult to interpret the coefficient estimates had we run regressions analogous to (16) and (17) with $\hat{E}_j$ being replaced by the health factors. Moreover, the existing empirical research we cited earlier all used a uni-dimensional summary such as claim probability, or health expenditure, or probability of using nursing homes etc. to measure the relevant risks.

Second, a seemingly plausible alternative (and in some sense symmetric) empirical strategy is to impute the missing variables in the list $\mathbf{X}$ for the MCBS sample, using some sort of prediction equations for $\mathbf{X}$ using HRS sample. It is not clear, however, what variables we can use to predict $\mathbf{X}$ except for the demographics variables. However, some of the $\mathbf{X}$ variables are themselves demographics variables, such as education and income. Moreover, there would be more necessary imputations to conduct than the approach we took.

Third, given our trepidations about the weakness of either imputation methods we described above, it seems reasonable to think about estimating a joint model of Medigap purchase and medicare usage with explicit modelling of selections. This is obviously an important direction for future research, but it would require more theoretical understanding of how to work with multi-dimensional selection models with incomplete data.

7.2 Comparison of MCBS and HRS Data

Before we proceed to describe our main results, we first show in this section that MCBS and HRS samples are in fact quite similar, and thus using MCBS to impute mean and variances of medical expenditure for HRS may be reasonable extrapolations.

[Table 5 About Here]

Panel A of Table 5 compares MCBS and HRS means for the common set of demographic variables under the first definition of "Medigap." 60.2 percent of the MCBS and 57.2 percent of the HRS sample are female. In both MCBS and HRS, the average age is higher for females than for males, with MCBS sample slightly older than HRS sample. The percentage of individuals with Medigap is also similar in the two samples: 45.9 percent in MCBS and 48.6 percent in HRS. This similarity is quite remarkable because the exact survey modules for health insurance status are quite different in the two surveys. In both MCBS and HRS, close to 95 percent of the samples are covered by both Medicare A and B, which is consistent to many previous findings. The marital status, number of children and educational attainment of the two surveys slightly differ, but the difference is very minor.

Panel B of Table 5 reports similar comparisons between MCBS and HRS where the sample uses second definition of "Medigap." Recall that the key difference between the two definitions of Medigap is regarding whether individuals with employer-sponsored health insurance to which they contributed more than $500 are coded as having Medigap or dropped. As expected, adding those with employer sponsored health insurance slightly lowers the female proportion because most of those working are men; slightly lowers the average age because those working tend to be slightly younger. The percentage with "Medigap" in both sample increased: 54.5 percent in MCBS and 56.8 percent in HRS, again a remarkable similarity given the different survey designs regarding health insurance. Overall it is quite obvious that the MCBS and HRS sample are remarkable similar for the means of the common set of demographic variables. Thus we are confident that using MCBS to impute the expenditures for HRS sample is reasonable.

7.3 Sources of Advantageous Selection: Main Findings

In this section we describe our main results regarding the sources of advantageous selection. Table 6 reports results for which the imputation is done using only observations in MCBS with no Medigap coverages (the first imputation method described above). Panel A and B respectively report the results for the two definitions of "Medigap" insurance status (see Section 5.3 for details). In each panel we report coefficient estimates on the predicted mean expenditure $\hat{E}_j/1000$ in the first three columns headlined by "Coefficient Estimate of Pred. Exp./1000," while the columns headlined by "Conditioning Variables" indicate the list of variables included in the regressions for each row, with "Y" (and "N" respectively) indicating that the variable listed in the head of the column is (is not, respectively) included. As we add more variables to the regressions, the number of observations, which we record in the last column of the table, drop due to missing values. For example, adding risk tolerance into the regression eliminates about 2/3 of the observations because HRS did not ask everyone the financial risk questions. Similarly when we include cognition variables (in particular cognition questions related to numeracy) we lose another half of the sample. In order to address the concern that our results are driven only by the changing samples, we rerun all previous regressions with smaller samples. It is important to note that smaller samples do affect the magnitude and p-values of the coefficient estimates, but the qualitative conclusions are not affected.

Rows (1) and (9) showed that, irrespective of the definitions of "Medigap" insurance status, if we do not control for any of the $\mathbf{X}_j$ variables, there is evidence that individuals with higher mean for the health expenditure risk are less likely to purchase Medigap. This, of course, is simply a confirmation (via a different angle of looking at the same data) of our finding of advantageous selection reported in Section 6.

Rows (2) and (10) added risk tolerance alone into the regression (17). This specification is important because the previous literature we have cited, both theoretical and empirical literature, has focused exclusively on risk aversion (which is simply the inverse of risk tolerance) as the source of advantageous selection. Note that the inclusion of risk tolerance into the regression only slightly reduce, if at all, the magnitude of the negative coefficient estimate for the predicted expenditures.

However, the inclusion of risk tolerance, predicted variance and their interactions, whose results are reported in rows (3) and (11), significantly affects the coefficient estimate for $\hat{E}_j$, reversing its sign from negative to positive, albeit still insignificant statistically.

As we include more variables from $\mathbf{X}_j$, reported in rows (4)-(8) for the first Medigap definition and in rows (12)-(16) for the second Medigap definition, the coefficient on $\hat{E}_j$ eventually becomes positive and statistically significant at 5 to 10 percent level.

[Tables 6-7 About Here]

Table 7 reports similar results using the second imputation method we described earlier where we use all observations in MCBS in estimating the prediction equations (11) and (12). We will not labour through the details of the results in Table 7 because they are qualitatively very similar to those in Table 6. We do note, however, the results in Table 7 when we use the second definition of Medigap status, though qualitatively in the same directions, are somewhat weaker in the statistical significance.

The messages in Table 6 and 7 are quite clear. Risk tolerance alone, even when interacted with measures of risks, is not enough to explain all of the advantageous selection we documented in Section 6. The sources of advantageous selection include at least education, income, cognition, longevity expectations and financial planing horizon, among possibly other variables. This conclusion is based on the fact that, once we condition on all these factors, we do find that individuals with higher expected medical expenditure risks are indeed more likely to buy Medigap insurance. Our findings seem to be quite consistent across the sample and methods of imputation for $\hat{E}_j$ and $\widehat{VAR}_j$. Also, We would like to add that, this is to the best of our knowledge the first to provide evidence that risk aversion itself is unlikely to be the sole source of advantageous selection, despite its purported importance in the prior theoretical and empirical literature; this is also the first to identify a set of variables that are sufficient to explain away the negative correlation between *ex post* health expenditure and Medigap coverage documented in Section 6.

Finally, we would like to briefly remark on a possible critique against our finding that cognition is a source of advantageous selection. One may argue that cognition, which is measured by composite scores from "Word Recall," "TICS Score," "Subtraction," and "Numeracy" in HRS, may itself be interpreted as a health variable. However we believe that these measures of cognition do not indicate illness that needs medical treatment, in fact many medical conditions that lower these above mentioned cognitive measures are not at all treated medically, and as such do not constitute part of the medical expenditure risks relevant to the insurance contracts.

7.4 Robustness of Results

In this section, we briefly discuss results from two robustness checks.

Alternative Measure of Health Expenditure Risk. In Section 7, we used the variance of medical expenditures as a measure of the risk that individuals care about when they decide whether to purchase Medigap insurance. One may argue that individual may care about some catastrophically large medical expenses, and the second moments may not be a good measure of the relevant risk. In Table 8 and 9, we report results from regressions similar to those reported in Tables 6 and 7 with the only exception that we use the ratio of the predicted 90th percentile medical expenditure over 10th percentile expenditure as a measure of the health expenditure risk (instead of the predicted variance). More specifically, we use the MCBS sample, again with two different imputation methods similar to those described above, to run quantile regressions for 90th and 10th quantile respectively; and then use these quantile regression coefficients to predict $\hat{Q}_{j90}$ and $\hat{Q}_{j10}$ for the HRS sample; then for each observation in HRS, we construct $\hat{Q}_{j90/10} = \hat{Q}_{j90}/\hat{Q}_{j10}$ as the measure of the medical expenditure risk, in place of $\widehat{VAR}_j$. From the coefficient estimates in Tables 8 and 9, it is quite clear that our qualitative results regarding the sources of advantageous selection is robust to this alternative measure of health expenditure risk.

[Tables 8-9 About Here]

Medicare HMO Coded as "Medigap". So far we have treated anyone with only Medicare coverage as without Medigap. However, in 2000 and 2001, approximately 15 percent of all Medicare beneficiaries chose to participate in a "Medicare Advantage Plan." These are HMOs that have contracted with Medicare to provide Medicare insurance. These Medicare Advantage Policies are not available everywhere, especially not available in rural areas. They require participation in Medicare A and B, and charge additional premiums. The benefit to participants of the extra premiums and the restrictions imposed by the HMO is that the HMO "generally" fills the same gaps that Medigap policies do. Thus, the U.S. government's "*Guide to Choosing a Medigap Policy"* tells those who have selected a Medicare Advantage Plan that they do not need a Medigap policy. Of course some of them still buy Medigap policy anyway. We experimented with coding those who have chosen a Medicare HMO either as having a Medigap status M_i of 1 or coding these respondents as missing. Without reporting all the results again, we note that the recoding of Medicare HMO participants actually somewhat strengthens our findings of advantageous selection in Section 6 (that is, the negative coefficients of M_i in Table 1 became even more negative); and does not change the qualitative findings about the sources of advantageous selection in Section 7.3.

8 Discussion: Selection versus Moral Hazard

A challenging question of central importance in economics of information is whether we can distinguish selection (either adverse or advantageous selection) from moral hazard. Chiappori and Salanié's (2000) test for asymmetric information via the positive correlation between *ex post* risk occurrence and insurance coverage, for example, is not designed to discriminate between adverse selection and moral hazard, since both can lead to the same qualitative relationship between *ex post* risk occurrence and insurance coverage. Abbring, Chiappori and Pinquet (2003) and Abbring, Chiappori, Heckman and Pinquet (2003) proposed to exploit dynamic insurance data to discriminate between selection and moral hazard. The idea is that, in dynamic automobile insurance contracts where experience rating is an important feature, adverse selection will lead to positive, while moral hazard will lead to negative, serial correlation in accident probabilities. Thus dynamic insurance data may enable researchers to qualitatively discriminate adverse selection from moral hazard. Recently, Bajari, Hong and Khwaja (2005) proposed a semiparametric method to test for the presence of moral hazard. The heart of their approach is the first order optimality condition of the insured, which relates the unobserved medical expenditure risk to "observable" consumption.

As such their approach is valid only if we literally believe in their static model. Moreover, in their model individuals are symmetrically uninformed about their health risk as the insurance company when they make insurance purchase decisions. Thus it is somewhat difficult to understand what adverse selection means in that context. In this section, we briefly discuss the relevant implications of our Tables 1-3 reported in Section 6 on the issue of selection versus moral hazard. We argue that, under some conditions, the coefficient estimates on "Medigap" presented in Table 2 and 3 also provide lower bound estimates of moral hazard (or simply, the price effect) of Medigap insurance.

To see this, suppose that the true medical expenditure equation is

$$E_i = \beta_0 + \beta_1 M_i + \boldsymbol{\beta}_2 \mathbf{H}_i^O + \boldsymbol{\beta}_3 \mathbf{D}_i + \beta_4 U_i + \varepsilon_i, \tag{18}$$

where E_i is total medical expenditure, M_i is the "Medigap" indicator, and $\mathbf{H}_i^O$ is a list of variable measuring the observable component of health, $\mathbf{D}_i$ is a list of demographic controls, and U_i is an index of unobservables including *both* unobservable components of health *and* preferences for medical service. Assume that the residual, ε_i, is uncorrelated with the independent variables.

In this equation coefficient β_1 will be the *true* measure of moral hazard, because we are assuming that the observable health $\mathbf{H}^O$ and the unobservable U jointly control for the true health and taste for medical services, thus eliminating selection issues. The problem, however, is that U_i, which contains unobservable components of health and preferences for medical service, is not observed.

The regressions we reported in Tables 2 and 3 are:

$$E_i = \tilde{\beta}_0 + \tilde{\beta}_1 M_i + \tilde{\boldsymbol{\beta}}_2 \mathbf{H}_i^O + \tilde{\beta}_3 \mathbf{D}_i + \tilde{\varepsilon}, \tag{19}$$

because we only control for the observable component of health $\mathbf{H}_i^O$ and not U_i. Thus, $\tilde{\beta}_1$ is biased from β_1 because of the omission of U_i. It is well-known that the degree of omitted variable bias can be calculated if we also run an auxiliary (and imaginary) regression:[38]

$$U_i = \pi_0 + \pi_1 M_i + \pi_2 \mathbf{H}_i^O + \boldsymbol{\pi}_3 \mathbf{D}_i + \mu_i. \tag{20}$$

We have

$$\tilde{\beta}_1 = \beta_1 + \beta_4 \pi_1.$$

If, without loss of generality, we define the unobservables in U_i to be "positives" (i.e., factors that will lead to less medical expenditure), we will have by definition $\beta_4 < 0$, then our estimate $\tilde{\beta}_1$ is a lower bound of β_1 – thus a lower bound estimate of the moral hazard – if $\pi_1 \geq 0$. Note that

[38]See, e.g., Wooldridge (2006, p. 120) for an expression of the omitted-variable bias.

parameter π_1 measures the partial correlation between U_i and M_i conditional on observable health H^O and $\mathbf{D}$ (which are controls for Medigap prices). The condition $\pi_1 \geq 0$ means that U and M are positively correlated conditional on H^O and Medigap pricing.

In practice, the regression (20) can never be actually implemented, thus it is impossible to examine whether the condition $\pi_1 \geq 0$ is actually satisfied. There is one case, however, we may expect this condition to be true. *Suppose* that conditional on health and demographics, tastes for medical service are unrelated to the taste for Medigap insurance purchase. This means that the variable U_i only contains unobserved health component H_i^U.[39] The sign of the partial correlation between H_i^U and M_i of course depends on the behavior of individuals' Medigap purchase. One scenario under which $\pi_1 \geq 0$ will be satisfied is as follows. If individuals make Medigap purchase decisions based only on observable health H_i^O – which is the only possibility if H_i^U is not observed not only by researchers but also unobserved by the individuals when they make Medigap purchase decisions[40]– then π_1 will be zero. If H_i^U is observed by individuals when they make Medigap purchase decisions, then the assumption that $\pi_1 \geq 0$ is that the overall selection based on unobservable health is also advantageous. While we can never explicitly verify whether or not this is true, we think it is plausible because we have established strong evidence of overall advantageous selection based on observable health. In this particular case, it is important to emphasize that whether or not the condition $\pi_1 \geq 0$ is satisfied is not constrained in any way by the covariance of H_i^O and H_i^U. In other words, even in situations where H_i^U and H_i^O are negatively correlated (which is implausible when we think of H_i^U as measurement error), the condition $\pi_1 \geq 0$ can still be true.

9 Conclusion

In this paper we use Medicare Current Beneficiary Survey (MCBS) data to document strong evidence of advantageous selection in Medigap insurance market. The indirect evidence for advantageous selection is established from the results from two sets of related regressions. In the first set of regressions, we regress "Total Medical Expenditure" on medigap status and only control for gender, age and State of residence – three variables that more or less determines the pricing of Medigap, we find that those with Medigap incur about \$4,000 less in total medical expenditure than those without Medigap. In the second set of regressions, we regress "Total Medical Expendi-

[39]In that case, of course, β_4 will be equal to β_2 because both H^O and H^U are measuring "health."

[40]Of course, H^U will affect the medical expenditures if it is revealed to them during the doctor visits.

ture" on age, gender, State of residence as well as controls for observable health variables. We find that those with Medigap spend about $2,000 more than those without Medigap. These two sets of results can be rationalized only if those with better health are more likely to purchase supplemental coverage, a phenomenon that has been termed as "advantageous selection." We also find direct evidence of advantageous selection by showing that "healthy" factors (i.e. health factors that tend to lead to lower medical expenditures) tend to be positively correlated with Medigap coverage, while "unhealthy" factors tend to be negatively correlated with Medigap coverage. These results are robust to different definitions of Medigap status, and whether we separate the sample into male and female subsamples. We do find that the magnitude of advantageous selection seems to be larger for the females than for males.

We then propose a simple empirical strategy to combine MCBS and HRS to empirically examine the sources of advantageous selection. We find that, risk tolerance (the usual suspect of the source of advantageous selection) alone, even when interacted with measures of risks, is not enough to explain all of the advantageous selection we previously documented. We find that additional sources of advantageous selection include at least education, income, cognition, longevity expectations and financial planing horizon, among possibly other variables. This conclusion is based on the fact that, once we condition on all these factors, we do find that individuals with higher expected medical expenditure risks are indeed more likely to buy Medigap insurance. Our findings seem to be quite consistent across the sample and methods of imputation.

Before we discuss some implications of our results, we would like to emphasize that our empirical findings are probably specific to the Medigap insurance market. The unique features due to the heavily regulated nature of the Medigap insurance market, as well as the natural link to the Medicare administrative data, make it as a perfect test ground for advantageous selection, but these unique features of Medigap insurance market at the same time also limit the generalizability of our quantitative findings to other markets, though we believe that as a qualitative finding, the evidence and sources of advantageous selection we uncover here should be more general.

One interesting implication of the theory of advantageous selection is that it offers a unified and plausible explanation about why different insurance markets vary so much in size.[41] For example, the markets for automobile insurance, health insurance, homeowner and life insurance are all enormous, while the markets for annuity insurance and long-term care insurance are quite

[41] The standard theory of insurance that features one-dimensional private information does not provide any convincing explanation.

small.[42] At first it may seem that institutional details could explain the size differences across insurance markets, at least for U.S. For example, automobile insurance is required by law for all vehicle owners; health insurance is frequently provided by employers; and social security is a mandatory annuity that crowds out private voluntary annuity, etc. However, these explanations are not adequate for several reasons. First, there is neither government regulation nor widespread employer provision for life insurance in the U.S. or in many other countries, and yet life insurance is a large and robust market. Second, the size differences among insurance markets are very similar across many developed countries that vary in institutional details. Third, for annuity markets theoretical results indicate that large welfare gains can be achieved by additional annuitization even in the presence of social security (see Yaari 1965 for an earlier result and Davidoff, Brown and Diamond 2005 for recent generalization).

To illustrate why selection based on multiple dimensions of private information provides a potential explanation for the size differences across insurance markets, it is useful to contrast the life insurance and the annuity insurance markets. The risks covered in these two insurance markets are exactly the opposite of each other, with the life insurance covering the risk of mortality and the annuity insurance covering the risk of longevity. In life insurance market, the "bad risks" from insurance company's viewpoint are those who have higher mortality probabilities. If selection based on other dimensions of private information leads to a negative correlation between mortality risk and life insurance coverage, i.e., if selection based on other dimensions of private information makes it more likely for healthier people to buy life insurance, it is possible that overall there is no positive correlation between life insurance coverage and *ex post* mortality risk, as empirically found by Cawley and Philipson (1999). In contrast, in annuity insurance market, the "bad risks" from insurance company's viewpoint are those who live long. Selection based on risk means that given any annuity premium, those with private information that they are relatively healthy will be more likely to purchase the annuity. Selection based on other private information, however, if it worked to alleviate the adverse selection based on risk in the life insurance market, must now work to actually exacerbates, rather than alleviates, the standard adverse selection based on risk! Thus the theory of advantageous selection provides a plausible explanation of the size difference between life and annuity insurance markets without relying on ad hoc assumptions on the differences in institutions.

[42] See Finklestein and Poterba (2004) for some discussion about the size of annuity markets. Finkelstein and McGarry (2005) report that about ten percent of elderly had long-term care insurance in their data.

Finally we would like to point out that the evidence of advantageous selection in certain insurance markets, and as a result the failure to find a positive correlation between *ex post* risk occurrence and insurance coverage, does not mean that there is no inefficiency in such markets. The policy implications of multidimensional selection models need further research.

References

[1] Abbring, Jaap H., Pierre-André Chiappori, James J. Heckman and Jean Pinquet (2003). "Adverse Selection and Moral Hazard in Insurance: Can Dynamic Data Help to Distinguish?" *Journal of European Economic Association Papers and Proceedings*, Vol. 1, 512-521.

[2] Abbring, Jaap H., Pierre-André Chiappori and Jean Pinquet (2003). "Moral Hazard and Dynamic Insurance Data." *Journal of the European Economic Association*, Vol. 1, 767–820.

[3] Akerlof, George (1970). "The Market for 'Lemons': Quality Uncertainty and the Market Mechanism," *Quarterly Journal of Economics*, Vol. 84, 488-500.

[4] Arellano, Manuel and Costas Meghir (1992). "Female Labor Supply and On-the-Job Search: An Empirical Model Estimated Using Complementary Data Sets." *Review of Economic Studies*, Vol. 59, 537-557.

[5] Angrist, Joshua D. and Alan B. Krueger (1992). "The Effect of Age at School Entry on Educational Attainment: An Application of Instrumental Variables with Moments From Two Samples." *Journal of the American Statistical Association*, Vol. 87,No. 418, 328-326.

[6] Araujo, Aloisio and Humberto Moreira (2001). "Non-Monotone Insurance Contracts and their Empirical Consequences." mimeo, IMPA, Brazil.

[7] Arrow, Kenneth (1963). "Uncertainty and the Welfare Economics of Medicare Care." *American Economic Review*, Vol. 53: 941-973.

[8] Asinski, Dzmitry (2005). "Health Insurance, Access to Care and Risk Aversion: Separating Selection and Incentive Effects." mimeo, Iowa State University.

[9] Bajari, Patrick, Han Hong and Ahmed Khwaja (2005). "A Semiparametric Analysis of Adverse Selection and Moral Hazard in Health Insurance Contracts." Preliminary Mimeo, University of Michigan and Duke University.

[10] Barsky, Robert B., F. Thomas Juster, Miles Kimball and Matthew D. Shapiro (1997). "Preference Parameters and Behavioral Heterogeneity: An Experimental Approach in the Health and Retirement Study." *Quarterly Journal of Economics*, 537-379.

[11] Beliveau, Barbara C. (1984). "Theoretical and Empirical Aspects of Implicit Information in the Market for Life Insurance." *Journal of Risk and Insurance.* Vol. 51, No. 2, 286-307.

[12] Brown, Jeffrey and Amy Finkelstein (2004). "Supply or Demand? Why is the Market for Long-Term Care Insurance so Small?" NBER Working Paper No. 10782.

[13] Cardon, James H. and Igal Hendel (2001). "Asymmetric Information in Health Insurance: Evidence from the National Medical Expenditure Survey." *Rand Journal of Economics*, Vol. 32, No. 3, 408-427.

[14] Cawley, John and Tomas Philipson (1999). "An Empirical Examination of Information Barriers to Trade in Insurance." *American Economic Review*, Vol. 89, No. 4, 827-846.

[15] Centers for Medicare and Medicaid Services (2004). *Choosing a Medigap Policy: A Guide to Health Insurance for People with Medicare.* U.S. Department of Health and Human Services, Baltimore, Maryland.

[16] Chiappori, Pierre-André (2000). "Econometric Models of Insurance under Asymmetric Information." in *Handbook of Insurance*, edited by Georges Dionne, p. 365-394.

[17] Chiappori, Pierre-André and Bernard Salanié (2000). "Testing for Asymmetric Information in Insurance Markets." *Journal of Political Economy*, Vol. 108, No. 1, 56-78.

[18] Chiappori, Pierre-André, Bruno Jullien, Benard Salanié and Francois Salanié (2005). "Asymmetric Information in Insurance: General Testable Implications." forthcoming, *Rand Journal of Economics.*

[19] Dahlby, Bev G. (1983). "Adverse Selection and Statistical Discrimination: An Analysis of Canadian Automobile Insurance." *Journal of Public Economics*, Vol. 20, 121-130.

[20] Davidoff, Thomas, Jeffery R. Brown and Peter A. Diamond (2005). "Annuities and Individual Welfare." forthcoming, *American Economic Review.*

[21] de Meza, David and David C. Webb (2001). "Advantageous Selection in Insurance Markets." *Rand Journal of Economics*, Vol. 32, No. 2, 249-262.

[22] Dionne, Georges, Christian Gouriéroux and Charles Vanasse (2001). "Testing for Evidence of Adverse Selection in Automobile Insurance Market: A Comment." *Journal of Political Economy*, Vol. 109, No. 2, 444-453.

[23] Ettner, Susan L. (1997). "Adverse Selection and the Purchase of Medigap Insurance by the Elderly." *Journal of Health Economics*, **Vol.** 16, 543-562.

[24] Finkelstein, Amy and Kathleen McGarry (2003). "Private Information and its Effect on Market Equilibrium: New Evidence from Long-Term Care Insurance." NBER Working Paper 9957.

[25] Finkelstein, Amy and Kathleen McGarry (2005). "Multiple Dimensions of Private Information: Evidence from the Long-Term Care Insurance Market." forthcoming, *American Economic Review*.

[26] Finkelstein, Amy (2004). "Minimum Standards, Insurance Regulation and Adverse Selection: Evidence from the Medigap Market." mimeo, Harvard University.

[27] Finkelstein, Amy and James Poterba (2004). "Adverse Selection in Insurance Markets: Policyholder Evidence from the U.K. Annuity Market." *Journal of Political Economy*, **Vol.** 112, 183-208.

[28] Hemenway, David (1990). "Propitious Selection." *Quarterly Journal of Economics*, Vol. 105, 1063-1069.

[29] Hemenway, David (1992). "Propitious Selection in Insurance." *Journal of Risk and Uncertainty*, Vol. 5, 247-251.

[30] Hurd, Michael D. and Kathleen McGarry (1997). "Medical Insurance and the Use of Health Care Services by the Elderly." *Journal of Health Economics*, Vol. 16, 129-154.

[31] Ichimura, Hidehiko and Elena Martinez-Sanchis (2004). "Identification and Estimation of GMM Models by Combining Two Data Sets." Working Paper, University College London.

[32] Jullien, Bruno, Benard Salanié and Francois Salanié (2005). "Screening Risk Averse Agents Under Moral Hazard: Single-crossing and the CARA Case." forthcoming, *Economic Theory*.

[33] Juster, F. Thomas and Richard Suzman (1995). "An Overview of the Health and Retirement Study." *Journal of Human Resources*, Vol. 30, S7-S56.

[34] Khandker Rezaul K. and Lauren A. McCormack (1999). "Medical Spending by Beneficiaries with Various Types of Supplemental Insurance." *Medical Care Research and Review*, Vol. 56, No. 2, 137-155.

[35] Kezdi, G. and R.J. Willis (2005) "Who Becomes a Stockholder? Expectations, Subjective Uncertainty, and Asset Allocation." Mimeo, University of Michigan.

[36] Kimball, Miles S., Claudia R. Sahm, and Matthew D. Shapiro (2005). "Using Survey-Based Risk Tolerance," mimeo, University of Michigan.

[37] Lillard, Lee A. and Jeannettte Rogowski (1995). "Does Supplemental Private Insurance Increase Medicare Costs?" Rand Corporation Working Paper Series 95-16.

[38] Pauly, Mark V. (1974). "Overinsurance and Public Provision of Insurance: The Roles of Moral Hazard and Adverse Selection." *Quarterly Journal of Economics*, Vol. 88, 44-62.

[39] Puelz, Robert and Arthur Snow (1994). "Evidence on Adverse Selection: Equilibrium Signaling and Cross-Subsidization in the Insurance Market." *Journal of Political Economy*, Vol. 102, No. 2, 236-257.

[40] Robst, John (2001). "Estimation of a Hedonic Pricing Model for Medigap Insurance." Unpublished Working Paper, Centers for Medicare and Medicaid Services.

[41] Rothschild, Michael and Joseph E. Stiglitz (1976). "Equilibrium in Competitive Insurance Markets: An Essay on the Economics of Imperfect Information," *Quarterly Journal of Economics*, **Vol.** 90, 629-649.

[42] Wilson, Charles (1977). "A Model of Insurance Markets with Incomplete Information." *Journal of Economic Theory*, Vol. 16, 167-207.

[43] Wolfe, John R. and John H. Goddeeris (1991). "Adverse Selection, Moral Hazard and Wealth Effects in the Medigap Insurance Market," *Journal of Health Economics*, Vol. 10, 433-459.

[44] Wooldridge, Jeffrey M. (2006). *Introductory Econometrics: A Modern Approach.* Third Edition, Thomson.

[45] Yarri, Menahem E. (1965). "Uncertain Lifetime, Life Insurance and the Theory of the Consumer." *Review of Economic Studies*, **Vol.** 32, 137-150.

Table 1: Correlation between "Medigap" Coverage and Total Medical Expenditure in MCBS, with No Health Controls

	Panel A: First "Medigap" Definition			Panel B: Second "Medigap" Definition		
Variables	All Sample	Female Sample	Male Sample	All Sample	Female Sample	Male Sample
medigap	-4392.7***	-6037.4***	-1863.4***	-4142.6***	-5883.1***	-1629.3***
	(347.0)	(456.6)	(540.8)	(323.4)	(432.2)	(494.9)
female	270.0			-35.4		
	(356.7)			(313.1)		
(age-65)	387.5***	460.6***	292.9	425.2***	470.4***	383.9*
	(138.2)	(176.0)	(229.3)	(122.2)	(156.3)	(198.1)
(age-65)^2	1.94	-1.79	5.58	-4.12	-5.85	-3.37
	(10.65)	(13.20)	(18,84)	(9.55)	(11.90)	(16.62)
(age-65)^3	.12	.17	.07	.25	.26	.23
	(.22)	(.27)	(.43)	(.21)	(.25)	(.39)
State Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
# of Observations	15,945	9,725	6,220	18,708	11,218	7,490
Adjusted R^2	.0702	.0873	.0531	.0638	.0830	.0442

Note: The Dependent variable is "Total Medical Expenditure." See text and Data Appendix for the two definitions of Medigap. Standard errors in parenthesis are clustered at the individual level.
All regressions used cross-section weights.
*, ** and *** denote significance at 10%, 5% and 1% respectively.

Table 2: Correlation between "Medigap" Coverage and Total Medical Expenditure in MCBS, with Direct Health Controls

	Panel A: First "Medigap" Definition					
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	All	Female	Male	All	Female	Male
medigap	1937.0***	1677.3***	2420.9***	1732.8***	1426.2***	2210.1***
	(257.6)	(349.0)	(397.4)	(272.4)	(358.4)	(418.9)
female	-751.6***			-754.1***		
	(283.7)			(294.0)		
(age-65)	394.5***	417.5***	355.4*	419.6***	444.2***	392.1**
	(117.4)	(145.0)	(197.6)	(113.3)	(137.4)	(198.9)
(age-65)^2	-27.5***	-32.0***	-22.8	-28.3***	-32.7***	-25.2
	(9.3)	(11.4)	(16.3)	(9.0)	(11.1)	(16.4)
(age-65)^3	.474**	.548**	.466	.491**	.562**	.520
	(.207)	(.254)	(.380)	(.202)	(.247)	(.382)
# of Observations	14,129	8,371	5,758	14,105	8,365	5,740
Adjusted R^2	.2087	.1915	.2462	.2135	.2007	.2484
	Panel B: Second "Medigap" Definition					
medigap	1967.3***	1638.5***	2529.7***	1760.2***	1372.5***	2353.1***
	(238.7)	(311.5)	(377.4)	(255.9)	(330.0)	(398.3)
female	-926.1***			-911.9***		
	(264.0)			(275.7)		
(age-65)	371.6***	404.5***	371.8**	384.3***	417.5***	392.6**
	(104.1)	(129.2)	(171.1)	(101.8)	(124.0)	(172.0)
(age-65)^2	-25.6***	-30.2***	-24.8*	-25.6***	-30.3***	-26.3*
	(8.3)	(10.2)	(14.1)	(8.1)	(9.9)	(14.1)
(age-65)^3	.418***	.504**	.479	.420**	.506**	.520
	(.185)	(.227)	(.330)	(.182)	(.222)	(.331)
# of Observations	16,885	9,860	7,025	16,853	9,852	7,001
Adjusted R^2	.2001	.1906	.2342	.2042	.1991	.2362
State Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Other Demographic Controls	No	No	No	Yes	Yes	Yes

Note: The Dependent variable is "Total Medical Expenditure." All regressions are weighted by the cross section sample weight. See text and Data Appendix for the two definitions of Medigap.
The variables included as direct health controls are detailed in Data Appendix. The other demographics included are race, education, marital status, income, working and number of children.
Standard errors in parenthesis are clustered at individual level.
*, ** and *** denote significance at 10%, 5% and 1% respectively.

Table 3: Correlation between "Medigap" Coverage and Total Medical Expenditure in MCBS, with Controls for Health Factors

	Panel A: First "Medigap" Definition					
	(1)	(2)	(3)	(4)	(5)	(6)
Variables	All	Female	Male	All	Female	Male
medigap	2083.3***	1601.1***	2775.5***	1838.7***	1347.5***	2450.0***
	(280.6)	(353.8)	(426.3)	(293.0)	(361.7)	(453.5)
female	-1311.2***			-1274.4***		
	(271.5)			(285.4)		
(age-65)	421.8***	447.6***	436.0**	433.5***	467.8***	451.6**
	(118.7)	(151.5)	(194.2)	(116.9)	(144.2)	(196.6)
(age-65)^2	-28.7***	-32.6***	-27.3*	-28.6***	-32.3***	-26.7*
	(9.2)	(11.9)	(16.0)	(9.1)	(11.6)	(16.2)
(age-65)^3	.479**	.534**	.543	.481**	.513**	.520
	(.205)	(.261)	(.375)	(.201)	(.255)	(.381)
Factor1	321.0	-410.7***	1097.1*	324.9	-415.5***	1094.5*
	(498.0)	(89.0)	(658.7)	(495.1)	(90.3)	(662.0)
Factor2	4917.5***	4902.2***	4880.6***	4928.9***	4976.5***	4855.9***
	(268.3)	(343.2)	(368.4)	(269.3)	(349.4)	(357.9)
Factor3	-2979.4***	-2500.8***	-4048.4***	-3055.1***	-2449.8***	-4211.7***
	(306.2)	(300.7)	(623.9)	(308.0)	(307.7)	(605.5)
Factor4	-652.1**	-684.3*	-1073.5	-746.3**	-794.9**	-1141.8
	(311.1)	(384.3)	(911.2)	(309.7)	(381.2)	(895.8)
Factor5	75.8	436.4*	-2278.9*	28.7*	459.3*	-2294.7*
	(305.0)	(253.8)	(1340.1)	(309.8)	(252.8)	(1308.9)
# of Observations	14,129	8,371	5,758	14,105	8,337	5,731
Adjusted R^2	.1398	.1345	.1792	.1448	.1450	.1857
	Panel B: Second "Medigap" Definition					
medigap	2154.5***	1577.6***	2944.7***	1914.2***	1312.0***	2628.6***
	(256.7)	(321.1)	(412.7)	(207.8)	(338.0)	(426.4)
female	-1468.5***			-1435.7***		
	(249.1)			(261.3)		
(age-65)	433.1***	442.3***	421.8***	434.9***	452.4***	414.9***
	(105.0)	(134.7)	(168.2)	(103.9)	(129.6)	(170.6)
(age-65)^2	-29.2***	-31.6***	-27.1**	-28.5***	-31.1***	-25.2*
	(8.2)	(10.6)	(13.7)	(8.1)	(10.4)	(13.8)
(age-65)^3	.468***	.508**	.484	.461*	.489**	.437
	(.184)	(.234)	(.322)	(.181)	(.230)	(.325)
Factor1	147.6	-843.9***	847.2	144.3	-865.8***	878.8
	(422.3)	(88.3)	(767.3)	(420.1)	(89.0)	(808.3)
Factor2	4908.8***	4743.5***	4982.6***	4918.5***	4820.0***	4970.3***
	(246.2)	(310.0)	(360.9)	(247.6)	(316.7)	(356.9)
Factor3	-2949.4***	-2578.1***	-3560.8***	-3008.7***	-2537.1***	-3701.8***
	(265.3)	(264.4)	(515.9)	(266.5)	(269.2)	(518.7)
Factor4	-435.3	-689.7**	181.3	-510.9*	-771.5**	216.9
	(266.3)	(333.7)	(453.0)	(265.8)	(331.3)	(458.2)
Factor5	182.9	408.2*	-153.0	140.4	425.2*	-184.0
	(267.5)	(229.2)	(409.4)	(271.7)	(227.9)	(408.4)
# of Observations	16,885	9,860	7,025	16,853	9,822	6991
Adjusted R^2	.1401	.1357	.1548	.1447	.1451	.1605
State Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Year Dummy	Yes	Yes	Yes	Yes	Yes	Yes
Other Demographic Controls	No	No	No	Yes	Yes	Yes

Note: The Dependent variable is "Total Medical Expenditure." All regressions are weighted by the cross section sample weight. See text and Data Appendix for the two definitions of Medigap.

The variables included as direct health controls are detailed in Data Appendix. The other demographics included are race, education, marital status, income, working and number of children.

Robust standard error in parenthesis. *, ** and *** denote significance at 10%, 5% and 1% respectively.

Table 4: Partial Correlation between "Medigap" Coverage and Health Factors in MCBS, Conditional on Gender and Age

Factors	Panel A: First "Medigap" Definition						Panel B: Second "Medigap" Definition					
	All		Female		Male		All		Female		Male	
	EXP	**PCORR**	**EXP**	**PCORR**	**EXP**	**PCORR**	**EXP**	**PCORR**	**EXP**	**PCORR**	**EXP**	**PCORR**
Factor1	324.9	.0296	-415.5***	.0274	1094.5*	.0342	144.3	.0296	-865.8***	.0373	878.8	.0267
		(.000)		(.012)		(.009)		(.000)		(.000)		(.025)
Factor2	4928.9***	-.1166	4976.5***	.1290	4855.9***	-.0978	4918.5***	-.1166	4820.0***	-.1193	4970.3***	-.0953
		(.000)		(.000)		(.000)		(.000)		(.000)		(.000)
Factor3	-3055.1***	.0319	-2449.8***	.0425	-4211.7***	.0200	-3008.7***	.0319	-2537.1***	.0434	-3701.8***	.0138
		(.000)		(.000)		(.129)		(.000)		(.000)		(.247)
Factor4	-746.3**	-.0177	-794.9**	-.0171	-1141.8	.0229	-510.9*	-.0177	-771.5**	-.0157	216.9	-.0389
		(.035)		(.117)		(.083)		(.035)		(.120)		(.001)
Factor5	28.7	.0207	459.3*	.0143	-2294.7*	-.0204	140.4	.0207	425.2*	.0258	-184.0	.0262
		(.014)		(.190)		(.122)		(.014)		(.010)		(.028)
# of Obs.		14,131		8,373		5,758		14,131		9,862		7,025

Note: The columns labelled with "EXP" are the regression coefficients from Table 3 for the specification with other demographic controls. They are included in the table for the interpretation of the factors.

The columns labelled with "PORR" lists the partial correlation of "medigap" with the corresponding factors. The number in parenthesis is the significance level of the correlation.

Table 5: Descripitive Statistics of MCBS and HRS Samples

	Panel A: First Definition of Medigap						Panel B: Second Definition of Medigap					
	MCBS			HRS			MCBS			HRS		
	All	Female	Male	All	Female	Male	All	Female	Male	All	Female	Male
Female	0.602	1.000	.000	.572	1.000	.000	.593	1.000	.000	.558	1.000	.000
	(.489)	(.000)	(.000)	(.495)	(.000)	(.000)	(.491)	(.000)	(.000)	(.497)	(.000)	(.000)
Age	75.808	76.546	74.690	75.301	75.847	74.415	75.602	76.336	74.534	75.112	75.672	74.263
	(7.729)	(8.047)	(7.076)	(7.078)	(7.353)	(6.611)	(7.612)	(7.933)	(6.983)	(7.010)	(7.295)	(6.547)
Medigap	.459	.462	.455	.486	.495	.473	.545	.540	.552	.568	.565	.573
	(.498)	(.495)	(.498)	(.500)	(.500)	(.499)	(.498)	(.498)	(.497)	(.495)	(.496)	(.495)
Medicare_AB	.958	.966	.945	.949	.947	.953	.955	.965	.940	.948	.947	.949
	(.201)	(.181)	(.228)	(.219)	(.223)	(.213)	(.208)	(.185)	(.237)	(.223)	(.224)	(.220)
Black	.091	.094	.087	.076	.082	.069	.087	.090	.082	.072	.078	.065
	(.288)	(.293)	(.282)	(.265)	(.274)	(.253)	(.281)	(.286)	(.275)	(.259)	(.269)	(.246)
Hispanic	.077	.073	.082	.042	.040	.046	.070	.067	.074	.039	.036	.042
	(.266)	(.261)	(.275)	(.202)	(.195)	(.211)	(.255)	(.251)	(.261)	(.193)	(.188)	(.200)
Married	.485	.344	.698	.532	.391	.743	.512	.371	.717	.543	.390	.755
	(.500)	(.475)	(.459)	(.499)	(.488)	(.437)	(.500)	(.483)	(.451)	(.498)	(.488)	(.430)
Widowed	.379	.523	.162	.350	.481	.151	.361	.504	.153	.342	.478	.146
	(.485)	(.499)	(.368)	(.477)	(.500)	(.358)	(.480)	(.500)	(.360)	(.474)	(.500)	(.353)
Divorced	.079	.084	.071	.079	.092	.063	.074	.079	.068	.077	.092	.060
	(.270)	(.287)	(.258)	(.269)	(.289)	(.244)	(.262)	(.269)	(.252)	(.267)	(.290)	(.238)
# of children	2.998	2.912	3.127	3.193	3.134	3.283	2.978	2.890	3.106	3.163	3.093	3.263
	(2.238)	(2.224)	(2.252)	(2.228)	(2.227)	(2.238)	(2.187)	(2.177)	(2.194)	(2.193)	(2.198)	(2.193)
Working	.124	.088	.179	.137	.118	.165	.127	.091	.180	.137	.118	.163
	(.330)	(.284)	(.383)	(.344)	(.323)	(.371)	(.333)	(.287)	(.385)	(.344)	(.322)	(.369)
Less than HS	.343	.342	.345	.306	.306	.307	.323	.322	.323	.285	.287	.280
	(.377)	(.375)	(.384)	(.360)	(.362)	(.361)	(.368)	(.367)	(.368)	(.348)	(.350)	(.347)
High School	.276	.300	.240	.363	.387	.329	.276	.300	.240	.357	.382	.324
	(.447)	(.458)	(.427)	(.481)	(.487)	(.470)	(.447)	(.458)	(.427)	(.479)	(.486)	(.468)
Some College	.210	.217	.198	.177	.183	.169	.220	.227	.209	.179	.186	.170
	(.407)	(.413)	(.399)	(.381)	(.386)	(.375)	(.414)	(.419)	(.406)	(.383)	(.389)	(.376)
College	.081	.068	.101	.080	.065	.099	.088	.074	.108	.086	.070	.107
	(.273)	(.252)	(.301)	(.271)	(.247)	(.299)	(.283)	(.261)	(.310)	(.281)	(.255)	(.309)

Note: Statistics are calculated using cross section sample weights. Standard deviations are in parenthesis. Number of observations vary by variable and sample.

Table 6: Sources of Advantageous Selection: Predicting Medical Expenditure Using Only MCBS No Medigap Observations

	Coefficient Estimate of Pred. Exp./1000			Conditioning Variables								
				risk tol.	pred. var.	risk_tol* pred. variance	educ.	inc.	cogn.	long. expec.	financial planning horizon	# obs.
Panel A: First Definition of Medigap												
(1)	-.0039071 (.000)	-.0055806 (.001)	-.0057424 (.116)	N	N	N	N	N	N	N	N	9973
(2)	…	-.0055818 (.001)	-.0056979 (.118)	Y	N	N	N	N	N	N	N	3467
(3)	…	-.0030057 (.162)	.0039308 (.121)	Y	Y	Y	N	N	N	N	N	3467
(4)	…	-.0023415 (.281)	.0050846 (.063)	Y	Y	Y	Y	N	N	N	N	3467
(5)	…	-.0003857 (.843)	.006364 (.060)	Y	Y	Y	Y	Y	N	N	N	3467
(6)	…	…	.0075755 (.049)	Y	Y	Y	Y	Y	Y	N	N	1696
(7)	…	…	.0078087 (.055)	Y	Y	Y	Y	Y	Y	Y	N	1695
(8)	…	…	.0078258 (.061)	Y	Y	Y	Y	Y	Y	Y	Y	1659
Panel B: Second Definition of Medigap												
(9)	-.0053412 (.000)	-.0075388 (.000)	-.0077557 (.022)	N	N	N	N	N	N	N	N	11866
(10)	…	-.0075402 (.000)	-.0077136 (.022)	Y	N	N	N	N	N	N	N	4295
(11)	…	-.004046 (.060)	.0022409 (.398)	Y	Y	Y	N	N	N	N	N	4295
(12)	…	-.0027332 (.212)	.004437 (.130)	Y	Y	Y	Y	N	N	N	N	4295
(13)	…	-.0006852 (.726)	.005602 (.121)	Y	Y	Y	Y	Y	N	N	N	4295
(14)	…	…	.0068328 (.087)	Y	Y	Y	Y	Y	Y	N	N	2146
(15)	…	…	.006936 (.089)	Y	Y	Y	Y	Y	Y	Y	N	2143
(16)	…	…	.007086 (.093)	Y	Y	Y	Y	Y	Y	Y	Y	2103

Note: p-value in parenthesis. All regressions include controls for female, a third order polynomial in age-65 and State.

Table 7: Sources of Advantageous Selection: Predicting Medical Expenditure Using All Observations in MCBS

	Coefficient Estimate of Pred. Exp./1000			Conditioning Variables: risk tol.	pred. var.	risk_tol* pred. variance	educ.	inc.	cogn.	long. expec.	financial planning horizon	# obs.
Panel A: First Definition of Medigap												
(1)	-.0027535 (.004)	-.0051885 (.002)	-.0046501 (.167)	N	N	N	N	N	N	N	N	9973
(2)	…	-.0051882 (.002)	-.0046106 (.169)	Y	N	N	N	N	N	N	N	3467
(3)	…	-.0029584 (.173)	.0026287 (.307)	Y	Y	Y	N	N	N	N	N	3467
(4)	…	-.0022188 (.313)	.0037951 (.133)	Y	Y	Y	Y	N	N	N	N	3467
(5)	…	-.0003383 (.862)	.0049004 (.091)	Y	Y	Y	Y	Y	N	N	N	3467
(6)	…	…	.0056235 (.068)	Y	Y	Y	Y	Y	Y	N	N	1696
(7)	…	…	.0057792 (.074)	Y	Y	Y	Y	Y	Y	Y	N	1695
(8)	…	…	.0058701 (.080)	Y	Y	Y	Y	Y	Y	Y	Y	1659
Panel B: Second Definition of Medigap												
(9)	-.004292 (.000)	-.0077586 (.000)	-.0083194 (.005)	N	N	N	N	N	N	N	N	11866
(10)	…	-.0077704 (.000)	-.0082673 (.005)	Y	N	N	N	N	N	N	N	4295
(11)	…	-.0002559 (.956)	-.0028388 (.705)	Y	Y	Y	N	N	N	N	N	4295
(12)	…	.0017564 (.695)	.0013984 (.853)	Y	Y	Y	Y	N	N	N	N	4295
(13)	…	.0048991 (.262)	.0048324 (.518)	Y	Y	Y	Y	Y	N	N	N	4295
(14)	…	…	.0056695 (.438)	Y	Y	Y	Y	Y	Y	N	N	2146
(15)	…	…	.006178 (.402)	Y	Y	Y	Y	Y	Y	Y	N	2143
(16)	…	…	.0069497 (.346)	Y	Y	Y	Y	Y	Y	Y	Y	2103

Note: p-value in parenthesis. All regressions include controls for female, a third order polynomial in age-65 and State.

Table 8: Sources of Advantageous Selection: Q90/Q10 as a Measure of Medical Expenditure Risk, Using Only MCBS No Medigap Observations

	Coefficient Estimate of Pred. Exp./1000			Conditioning Variables: risk tol.	Q90/Q10	risk_tol* (Q90/Q10)	educ.	inc.	cogn.	long. expec.	financial planning horizon	# obs.
Panel A: First Definition of Medigap												
(1)	-.0039071 (.000)	-.0055806 (.001)	-.0057424 (.116)	N	N	N	N	N	N	N	N	9973
(2)	…	-.0055818 (.001)	-.0056979 (.118)	Y	N	N	N	N	N	N	N	3467
(3)	…	-.0038107 (.196)	.0033842 (.133)	Y	Y	Y	N	N	N	N	N	3467
(4)	…	-.0027534 (.297)	.0052466 (.067)	Y	Y	Y	Y	N	N	N	N	3467
(5)	…	-.0004784 (.784)	.0061864 (.061)	Y	Y	Y	Y	Y	N	N	N	3467
(6)	…	…	.0074877 (.051)	Y	Y	Y	Y	Y	Y	N	N	1696
(7)	…	…	.0076887 (.054)	Y	Y	Y	Y	Y	Y	Y	N	1695
(8)	…	…	.0076875 (.059)	Y	Y	Y	Y	Y	Y	Y	Y	1659
Panel B: Second Definition of Medigap												
(9)	-.0053412 (.000)	-.0075388 (.000)	-.0077557 (.022)	N	N	N	N	N	N	N	N	11866
(10)	…	-.0075402 (.000)	-.0077136 (.022)	Y	N	N	N	N	N	N	N	4295
(11)	…	-.0041860 (.064)	.0023856 (.401)	Y	Y	Y	N	N	N	N	N	4295
(12)	…	-.0028634 (.241)	.0046441 (.129)	Y	Y	Y	Y	N	N	N	N	4295
(13)	…	-.0006968 (.704)	.0057314 (.115)	Y	Y	Y	Y	Y	N	N	N	4295
(14)	…	…	.0067893 (.088)	Y	Y	Y	Y	Y	Y	N	N	2146
(15)	…	…	.0070446 (.090)	Y	Y	Y	Y	Y	Y	Y	N	2143
(16)	…	…	.007177 (.092)	Y	Y	Y	Y	Y	Y	Y	Y	2103

Note: p-value in parenthesis. All regressions include controls for female, a third order polynomial in age-65 and State.

Table 9: Sources of Advantageous Selection: Q90/Q10 as a Measure of Medical Expenditure Risk, Using All Observations in MCBS

	Coefficient Estimate of Pred. Exp./1000			risk tol.	Q90/Q10	risk_tol* (Q90/Q10)	educ.	inc.	cogn.	long. expec.	financial planning horizon	# obs.
				Conditioning Variables								
Panel A: First Definition of Medigap												
(1)	-.0027535 (.004)	-.0051885 (.002)	-.0046501 (.167)	N	N	N	N	N	N	N	N	9973
(2)	…	-.0051882 (.002)	-.0046106 (.169)	Y	N	N	N	N	N	N	N	3467
(3)	…	-.0031677 (.186)	.0024898 (.324)	Y	Y	Y	N	N	N	N	N	3467
(4)	…	-.0023675 (.308)	.0037658 (.142)	Y	Y	Y	Y	N	N	N	N	3467
(5)	…	-.0003478 (.856)	.0048917 (.094)	Y	Y	Y	Y	Y	N	N	N	3467
(6)	…	…	.0056176 (.069)	Y	Y	Y	Y	Y	Y	N	N	1696
(7)	…	…	.0057898 (.077)	Y	Y	Y	Y	Y	Y	Y	N	1695
(8)	…	…	.0058420 (.085)	Y	Y	Y	Y	Y	Y	Y	Y	1659
Panel B: Second Definition of Medigap												
(9)	-.004292 (.000)	-.0077586 (.000)	-.0083194 (.005)	N	N	N	N	N	N	N	N	11866
(10)	…	-.0077704 (.000)	-.0082673 (.005)	Y	N	N	N	N	N	N	N	4295
(11)	…	-.0002588 (.894)	-.0029143 (.697)	Y	Y	Y	N	N	N	N	N	4295
(12)	…	.0018676 (.704)	.0013877 (.838)	Y	Y	Y	Y	N	N	N	N	4295
(13)	…	.0049121 (.248)	.0048286 (.533)	Y	Y	Y	Y	Y	N	N	N	4295
(14)	…	…	.0055988 (.449)	Y	Y	Y	Y	Y	Y	N	N	2146
(15)	…	…	.006099 (.422)	Y	Y	Y	Y	Y	Y	Y	N	2143
(16)	…	…	.0068750 (.367)	Y	Y	Y	Y	Y	Y	Y	Y	2103

Note: p-value in parenthesis. All regressions include controls for female, a third order polynomial in age-65 and State.

DATA APPENDIX

Category	Variable	Data	Description
Health expenditure	Total expendidure	MCBS	Total annual health care expenditure for 12 months of the survey year. Expenditure includes data from Medicare administrative files and survey responses for out-of-pocket and otherwise insured expenditures.
Insurance			
	medicare		Indicators for whether the respondent is covered by medicare part A and part B.
	medigap	MCBS	Indicator for whether repsondent, with medicare coverage also has self-purchased private health insurance. Those covered by employer provided health inurance, medicaid or VA Champus (Tri-care) are treated as missing.
		HRS	Indicator for whether respondent, with Medicare coverage also has private health insurance that is secondary to Medicare, and not purchased from a (spouse's) employer or union. Those covered by employer provided health inurance, medicaid or VA Champus (Tri-care) are treated as missing.
	medigap2		Variable equal to one if medigap is equal to one, or if covered by employer sponsored health insurance and pay more than $500 per year in premiums.
Demographics	race		Indicators for self-reported black, other and non-response
	hispanic		Indicators for self-reported hispanic and non-response
	education		Indicators for highest grade completed less than 8th grade, some high school, high school graduate, some college, college graduate, at least some grad school and non-response
	marital status		Indicators for maried, widowed, divorced, separated and non-response
	number of children		The number of children the respondent has ever had.
	income	MCBS	Indicators for self-reported total household income in $5,000 intervals from $5,000 to $50,000, and $50,000 plus
		HRS	Same indicators as above, except in this case we use reported as well as imputed values for total household income. Imputations are those generated by RAND. See http://hrsonline.isr.umich.edu/meta/rand/randhrse/randhrse.pdf for details of the imputation.
	work status		Indicators if currently working for pay and for non-response.

Category	Variable	Data	Description
Health	Self-reported		Indicators for self-reported health excellent, very good, good, and fair.
	Height		Self-reported height, in inches, and height squared.
	Body mass index		Self-reported (weight (kg)) / (height (m) squared)
	Ever a smoker		Indicator if respondent has "ever smoked" tobacco
	Current smoker		Indicator if respondent now smokes tobacco, and for non-response
	Diagnoses		Indicators for if a doctor has ever told the respondent he/she has: arthritis, high blood pressure, diabetes, (non-skin) cancer, lung disease, heart attack, chronic heart disease, stroke, psychiatric illness, Alzheimer's disease, broken hip and for each diagnosis, non-response.
	Treatments		Indicators for respondent ever having cataract surgery or a hearing aid
	(Instrumental) activities of daily living		Indicators for if a respondent has at least some difficulty walking 2-3 blocks, stooping, reaching overhead, lifting 10lbs, dressing, walking at all, bathing, eating, getting out of a chair, using the toilet, preparing meals, shopping, using the telephone, managing money and bills, and for non-response.
	Help with IADS		Indicators for if a respondent receives help dressing, walking at all, bathing, eating, getting out of a chair, using the toilet, preparing a meal shopping, using the telephone or managing money and bills and for non-response.
Risk attitudes	Risk Tolerance	HRS	Estimate of risk tolerance from Kimball, Salm and Shapiro (2004), using responses to hypothetical income gambles from 1992 and 1994.
Cognition	Word recall	HRS	Variables recording the number of words recalled from a list of 10, both immediately after the list was read and several minutes later.
	TICS Score	HRS	Telephone Interview for Cognitive Status. Number of correct answers on a test of knowledge, language and orientation. Questions include naming objects, vocabulary questions, and basic knowledge such as the U.S. President's name.
	Subtraction	HRS	Number of times respondent can subtract the number 7 sequentially, starting from 100.
	Numeracy	HRS	Number of correct answers to "word problems" of division and multiplication on topics of probability, compound interest, and division of assets. Asked only in 2002.

Category	Variable	Data	Description
Expectations	Longevity	HRS	Most recent answer to the question "What is the percent chance you will live to 75 or more"
Planning Horizon	Financial	HRS	Indicators for whether the respondents most important period for planning saving and spending is the next few months, the next year, the next few years, the next five to ten years, or more than 10 years.