Forthcoming, *International Journal of Health Care Finance and Economics*.

The final publication will be available at www.springerlink.com

Five Questions for Health Economists

Randall P. Ellis

Boston University

August 20, 2012

This paper is a written version of the Presidential Address of Randall Ellis, presented on June 11, 2012 at the fourth biennial conference of the American Society of Health Economists (ASHEcon) at the Carlson School of Management of the University of Minnesota in Minneapolis. The author is grateful to Tom McGuire, Arlene Ash, Keith Ericson and Michael Manove for their useful comments and discussion and Tim Layton for his excellent research assistance.

Keywords: health economics; behavioral economics; insurance; primary care; risk adjustment

JEL codes: I10, I11, I13

I had a lot of fun deciding what I wanted to talk about in front of this very distinguished audience of the American Society of Health Economists. I will be raising five questions for health economists that are intended to provoke discussion and perhaps foster future research. I propose these without claiming them to be exhaustive or the most important set of questions, only that they are an interesting set. The examples and facts presented are drawn largely from my own research, but also rely upon the research of others. Without further ado, let me get started.

My first question is

Q1: If the ceiling in this room fell down and permanently paralyzed both of your legs, what type of health insurance coverage would you want?

After recovering from the unpleasantness of this scenario, you might decide that this is a very silly question. You would of course say that you want to be in a health plan with very complete coverage, with low copayments and deductibles, and few, if any, restrictions on your choice of providers. Remarkably, a majority of US health economists and policy makers are working to ensure that you are not offered such a generous plan, because it costs too much.

This serious injury would also have many other effects. Some of you would change your careers or drop out of the labor force, jeopardizing your access to employer-sponsored insurance. The effects on your lifestyle would be dramatic as well. At a minimum this serious injury would require fundamental changes in your living arrangements, commuting, work, travel, and social activities. In almost every case your income would go down substantially, whether due to your changed occupation, your reduced ability to travel, or the huge amount of time that you would spend in treatment and traveling to physical therapy. The time cost and inconvenience of having to pay your share of health care costs and the anxiety of worrying about what services your health plan will cover will impose other burdens.

It is interesting to consider how a health economist would think about this traumatic lifestyle change. Many of you would probably say that your preferences for medical care would change. Would you agree?[1]

It might seem like your preferences would suddenly change, but that would be wrong, or at least misleading. What you would experience is primarily a change in the state of the world that you live in, not a change in preferences.[2] That is a very important difference. Conventional

[1] More than half of the audience raised their hands in agreement.

[2] After the lecture, one economist pointed out that in light of the accident preferences would almost certainly change dramatically in light of the new insights one obtains from the experience of being disabled. I do not

economic analysis would define preferences even across states of the world, and these preferences characterize the choices you will make in response to each state of the world. Chance took its move, and a very bad state of the world occurred.

Conflation between heterogeneity of preferences and heterogeneity of the states of the world in which one finds oneself is widespread among health economists. We often observe large differences in quantity of medical care consumed, and we attribute much of this to differences in taste or preferences. But most of the variation across consumers is due to differences in health status or the states of the world in which we find ourselves. There is a huge amount of randomness, luck, genetic variation, environmental exposure, and other things not under our control that cause most of the variation in which states of the world one finds oneself in. Of course health status and states of the world are not exogenous (prevention activities and risk-taking clearly enter in), but my own work using diagnoses as signals of health status for risk adjustment models shows that more than half of the variation in health care spending can be explained by age, gender and concurrent health conditions, with only a few additional percentage points of this variation explained by income, education, race, and other variables that we often consider as defining "tastes." Serious permanent injury, or the onset of new chronic conditions cause much more variation over time and across individuals in health spending than does variation in preferences.

How would most economists model the impact of this shock to your health?

As economists, most of you would probably think about this shock as increasing your demand for health care services. A traditional demand curve analysis of the accident might look similar to Figure 1. After your legs are paralyzed, it is extremely likely that your worsened health status will dominate the income effects of your reduced income, and your demand curve for medical care would shift outward. It is less clear whether your demand for medical care would become more or less price responsive, and hence whether the inefficiency of your overconsumption of medical care would decrease. Also unclear is whether your marginal utility of income would increase or decrease after such a terrible accident. Money could become more important to you, or you might simplify your life so much that income would become less important. Welfare calculations become problematic when the marginal utility of income changes with states of the world.

Most health economists would take the analysis a step further by adding a marginal cost curve to the figure, which enables them to talk about consumer welfare and the efficient level of care. Insurance is desirable because it reduces financial risk, but it increases inefficiency as

---

disagree. But the response solicited immediately after being asked this question do not yet reflect this "Aha!" realization, and hence reflects only a change in state of the world, not a change in preferences.

measured by consumer surplus welfare triangles. [3] Notice how the sicker person will plausibly generate much greater welfare losses and inefficiency than the healthy person.

Since choice of health plan is taken almost axiomatically as a good thing here in the US, most health economists would favor allowing a choice between more or less generous plans. After a serious injury, almost everyone would want to migrate into a more generous health plan, with the healthy doing the opposite, trying to avoid subsidizing the sick (through their premiums) by moving into a less generous plan. Allowing for health plan choice after the state of the world is known is illustrated in Figure 2, with the healthy imposing little welfare loss and the sick increasing the moral hazard problem and "overconsuming" health care by moving into a more generous plan.

Many health economists would bemoan how these high cost people are using up most of our health care, imposing costs on the healthy; a great deal of research has appropriately focused on how to change the behavior of these high cost individuals. Many would advocate higher cost shares and deductibles, encouraging both movement along a given demand curve as well as perhaps greater prevention effort ex ante to avoid becoming high cost.

At this point you can easily see the disconnect between our demand models and our own behavior as consumers, since as individuals we would almost unanimously agree that if we were to suffer a serious accident or misfortune, we would want to be in a generous plan, not a stingy plan that imposes serious financial burdens. Our paradigm for thinking about how to measure welfare losses is contrary to the way we behave in our own choices and health plan preferences, as well as our expectations of what is a reasonable response when someone faces a serious accident, a surprise attack of multiple sclerosis or some other major chronic illness.

I will admit that I am guilty of this same type of thinking in much of my work. Will Manning and I published a paper in the Journal of Health Economics a few years ago in which we calculate optimal cost sharing precisely using these linear demand curves (Ellis and Manning, 2008). We calculated optimal cost sharing and worried about moral hazard and financial risk exactly as shown in these figures. But we also introduced another consideration which can greatly lower the optimal cost sharing relative to the conventional analysis. Our insight was to recognize that there are many non-medical costs of poor health, such as the time costs of seeking treatment and the cost of missing work. For the well-insured, what determines whether you or I go to the doctor for most illnesses is not the price of treatment, but the time cost, the inconvenience, the

[3] As Nyman (2003) has argued, consumer surplus is a highly imperfect measure of welfare change when income effects are large, and the analysis is more precisely done using compensating variations. Yet compensating variations are themselves challenging to calculate when the marginal utility of income varies across states of the world.

stress, and the uncertainty. These are not compensated by existing health insurance plans. In the example I posited of the ceiling falling down on you, the biggest losses you face are not the increased out-of-pocket costs for medical expenses, but rather the fact that you may lose your job, change your housing, and change your whole life.

Uncompensated losses have received relatively little attention in our field, but give an efficiency rather than simply a fairness basis for preferring generous insurance coverage. If money is valued more highly in sick than in healthy states of the world, then we should give more, not fewer, resources to people who suffer major chronic illnesses, and generous insurance is an imperfect, though second-best, method for doing so. In Europe the healthy and wealthy are expected to subsidize the sick and the poor, while in the US much of our modeling and policy seems to try to punish the unfortunate, reducing their health care access, even if that is the opposite of what we would want to have happen to ourselves in the same situation. Europe and Canada rely very little on demand-side cost sharing to control costs, and yet somehow achieve costs that are a much lower percent of their GDP than in the US.

Imposing significant demand-side cost sharing on people with serious illness shocks also has major consequences on another measure of household wellbeing, namely personal bankruptcies. Himmelstein et al (2009) calculated that 62.1% of all personal bankruptcies in 2007 in the US involved high medical expenses, even though 80% of the people declaring bankruptcy had health insurance. At the time of bankruptcy, those with health insurance had average medical expenses of just under $18,000, while those without health insurance had average medical bills of nearly $27,000. Michelle Miller (2011) (one of my own Ph.D. graduates now at Rutgers) has a terrific working paper on bankruptcy filings which is also a chapter in her dissertation. She notes, “In 2010, over 1.5 million households filed for bankruptcy, discharging more than $459 billion in debt and rivaling Medicare as one of the country’s largest transfer of wealth programs.” That is an astonishingly large number. At this ASHEcon conference we have a large number of papers on how to control the costs of Medicare and Medicaid, but a quick scanning did not reveal any of them on how to reduce medical bankruptcy rates. Of course, these bankruptcy costs are largely hidden, since we pay for them through higher credit card fees, and higher interest rates on mortgages, loans and purchases. But these high bankruptcy costs are a reminder of the further, uncompensated costs of demand-side cost sharing in health care.

As many of you are aware, I am known especially for my work on supply-side cost-sharing, especially with my good friend and colleague, Tom McGuire. We wrote a paper in 1986 on provider behavior under prospective payment which has stood the test of time and still provides useful insights. Fundamentally, we recommend trying to reduce and control costs not

by imposing more risk on consumers, but rather by incenting doctors and hospitals to try to hold costs under control. We identified the concept of a "mixed payment system" in which you pay providers both a marginal reimbursement for a fraction of their costs as well as a risk adjusted lump-sum bundle.[4] I'm not going to present the Ellis and McGuire model today, but I recommend its approach to you as an alternative to demand-side cost sharing to control costs.

This leads me to my second question. It is:

Q2. What approach do you prefer to ensure that the individual premiums of older workers (age 65 and over) are not ten times the premium of young workers?

If you want to allow health plans to choose premiums for their insurance policies, then the natural thing is to allow them to price just like any private firm. If a customer is more costly, why not let the plan experience rate the premiums based on observable information such as age and gender? If they do, in a competitive environment they will want to charge considerably more for insuring an older than a younger individual.

Figure 3 shows the distribution of health care spending (as measured by allowed charges) by one-year age cohorts for men and women using the Thomson-Reuters MarketScan 2004 commercial claims and encounter data. The numbers shown are from a sample of 14.6 million people, primarily employed by large employers in generous health plans. I use the 2004 data rather than more recent data because it includes a sizable number of people over age 65. Note that these enrollees are all privately insured, not relying on Medicare, most likely because they are still employed or are the spouse of somebody who is still employed.

The figure reveals many interesting patterns, such as: newborns are more expensive than toddlers; children are very cheap to insure; women in their child-bearing years cost more than men; and just before age 65 men start costing more than women. Both men and women in this sample peak in their annual health costs in their early 80s; and there is a distinct discontinuity in the spending distribution at age 65, as many of the sicker people presumably switch from private insurance to Medicare. But the key point of the figure is that older enrollees cost somewhere in excess of $10,000 per person per year to cover, while children and male 20 year olds cost in the vicinity of $1,000 per year, so there is roughly a tenfold difference in average health costs using just age and gender alone.

[4] The essence of the Ellis and McGuire (1986) model is that if providers care about both profits and patient benefits but weight patient benefits by proportion $0 < \alpha \leq 1$, then the optimal marginal reimbursement is $r^*=(1-\alpha)\cdot$(marginal cost) while the optimal lump sum reimbursement is $R^*=\alpha\cdot$(Expected total cost). Much of my research on risk adjustment over the past 25 years can be seen as trying to calculate the expected total cost.

In the absence of government regulation or remarkable degrees of altruism, insurance companies will want to charge premiums for these large differences, charging older workers, either individually or through their employer, ten times the premium of younger workers, or charging 30 year-old women twice the premium of 30 year-old men. Are you comfortable with this? Is that something that you want insurers to do? And notice that I am not even allowing for the possibility that premiums are allowed to differ in other dimensions, such as surcharges for people with diabetes, discounts for people in fitness clubs, or differentials between smokers and non-smokers. Permitting premium differentiation in these other dimensions would allow differentiating people whose expected costs might vary by more than 100-fold, from $400 per year to $40,000 per year or more.

I personally don't think it's a good idea to have premiums that differ that much. Which leads to the question I have posed, what do you want to do about it? One approach is community rating, forcing insurers to charge a flat premium to all. This levels the premium playing field but creates very strong incentives for health plans to try to select healthy enrollees. The recent health reforms in Massachusetts and the Affordable Care Act (ACA) propose using premium rate bands of 2 to 1 or 3 to 1, which restrict health plans according to the ratio of the maximum to minimum premiums within a given plan design. This seems to be working relatively well in Massachusetts, although I worry that it will create some of the same strong incentive effects as community rating. Moreover, for the national reform, exempting existing plans from complying, and allowing plan design proliferation greatly weakens the leveling effect. Both reform efforts also propose to use risk adjustment by the sponsor to level the playing field, making plans more willing to offer generous coverage and reducing incentives to differentiate premiums.

Because risk adjustment is one of my specialties, I want to also highlight the importance of risk adjustment for leveling the playing field across different enrollee populations and even benefit plan designs.[5] Ellis (2008) used the 2004 MarketScan sample of privately insured people, to calculate the average age and average annual health spending of enrollees across four common health plan types: traditional indemnity, HMO (Health Maintenance Organization), PPO (preferred provider organizations) and POS (point of service) plans. The average age of HMO enrollees was eight years younger than the nearly extinct traditional indemnity plans, with the PPO and POS plans having mean ages that are intermediate between the two. Average health care spending per person also differed enormously in the four plans, with the HMO the lowest and the indemnity plan the highest. If health plan premiums were simply equal to each plan's average cost per person, the HMO could offer a premium of $2,668 per person versus the

[5] As Glazer and McGuire (2000) demonstrate, it is tricky to do well, but in principle if a sponsor is allowed to pool premiums and make appropriate risk-adjusted payments to competing health plans, then in health plans can be induced to charge uniform premiums to all of their enrollees.

indemnity plan's offer at $4,844 per person. Thus, the HMO appears to be 45 percent cheaper than the indemnity plan. Of course most of this difference is driven by age and gender differences, and much of this will be captured by family versus individual policies, but the key issue I am focusing on here is the challenge of pricing this cost heterogeneity appropriately.

Age and gender are not the only dimensions along which enrollees differ across plan types: enrollees also differ in their health status, and HMOs seem to be able to attract relatively healthy people within a given age and gender cohort (Glazer and McGuire, 2000). To understand these differences, in Ellis (2008) I created age profile charts by health plan type which are recreated as Figures 4 and 5. These figures show the distribution of average health spending only up through age 65, since there are few people over age 65 in the sample. Several things are evident in Figure 4. The indemnity plan is notably more expensive even within an age interval and mean health spending in the PPO is almost indistinguishable from spending in the indemnity plan after controlling for age. After controlling for just age, HMOs are about 15-20 percent cheaper across the age spectrum, down significantly from the 45 percent apparent sample mean difference. POS plans are in between the PPO and HMO average costs.

Age and gender adjustment alone cannot compensate for differences in how sick people are; for that more powerful risk adjustment is needed. I used the Verisk Health/DxCG Hierarchical Condition Category (HCC) concurrent risk adjustment model, which is a more elaborate and powerful version of the Centers for Medicare and Medicaid (CMS) HCC model used for risk adjusting the Part C and D payments in Medicare. The results of controlling for current year diagnoses are shown in Figure 5. Here we see that the HMO cost advantage after risk adjustment is less than ten percent compared to PPOs. Risk adjusted spending in the indemnity and PPO plans are essentially indistinguishable, and, if anything, the POS plans that are so popular have slightly higher risk adjusted spending than PPOs and indemnity plans. Given that diagnoses cannot fully risk adjust for health status differences between HMO and PPO/indemnity enrollees, it is plausible that the cost savings from HMOs in the MarketScan data are only on the order of 3 to 5 percent relative to PPOs, not even the 5-10 percent estimated here.

This analysis should make all health economists ask what we are getting from promoting plan choice, benefit variation and diversity in health plan contracting, when, fundamentally, after controlling for age and health status, there are relatively small differences in average allowed costs.[6] Careful risk adjustment is important because it can reveal and undo the consequences of

[6] Demand-side cost sharing does shift responsibility for out-of-pocket payment between employers and enrollees, but so does the share of premiums paid by the employee. More importantly, out-of-pocket costs shift burdens between the healthy and sick.

biased selection. Even though it does not solve all selection problems, it is still worth doing. The importance of risk adjustment is revealed in the developed countries that have multiple, competing health plans, which include Belgium, Germany, Israel, Netherlands and Switzerland. In each case risk adjustment is used to reallocate funds between competing health plans, but, more importantly, with the exception of Switzerland, these countries tightly regulate the allowed benefit designs so that all of the available plans offer virtually the same level of plan generosity. Moreover, European countries use very little demand-side cost sharing, relying instead on supply-side incentives and regulations on capacity investment to control costs. This observation is consistent with my first point, which is that if you all want to have generous coverage when you are really sick, then why are you offering less generous plans at all? I know this view is not popular in the US, but my view is we should be comparing ourselves to and learning from what goes on in Europe, Canada, and elsewhere, most of which restrict health plan choice variety and mainly use regulation and supply-side policies to address their cost and quality problems.

So the failure of today's benefit plan designs to meaningfully control costs, together with the relative success of some other countries at controlling costs, suggests that directly changing provider incentives rather than promoting plan level choice may be the most important tool.

This leads me to my third question, which I am going to pose in two phases.

Q3a: If you were to choose one class of providers to give incentives to reduce total health spending, which one would it be?

Just to help you out, consider the following important classes of providers: hospitals, specialists, labs/imaging, pharmacists, and primary care practitioners. Which of these groups do you think will be the most helpful to incentivize to control costs? Hospitals account for the largest share of total health spending, but in the US we've have already put a lot of effort into trying to change payment systems for hospitals, so I am skeptical that there is anything easy left to do there. Controlling costs by changing incentives facing laboratories and imaging providers is desirable, but difficult, since they themselves only want to generate high volumes of care. Specialists are closely tied to the hospitals, clinics and laboratories where they work, and will never wish to control their own costs, since that can only reduce their own incomes, which are already quite high. Pharmacists are another area where certainly incentives could work, but given the relatively competitive nature of this market, one would have to change pharmaceutical pricing and manufacturers more than pharmacists. What about primary care practitioners? That would be my answer. They are the most important group to change incentives, since through referrals, prescriptions, and prevention effort they affect the entire health care system. So the second part of my question 3 would be:

Q3b. What would be your preferred alternative to fee-for-service payment for primary care practitioners (PCPs)?

I should start by noting that I really like my own PCP, like almost every American. We all think our own doctors are great; it's everybody else's doctor who is causing problems with cost, quality, and pricing. I do think of PCPs as being different from other providers, more like a quarterback who provides referrals, guides the rest of the team of providers, tells you when to go for specialty care, what drugs to take, and when to do watchful waiting. Ideally, your PCP is supposed to also guide you in lifestyle choices like losing weight, stopping smoking, changing diet and making other smart choices. Among all health care providers, PCPs may be the least profit oriented, and, because they are not paid as highly as many other specialists, perhaps there is greater hope at changing the incentives facing them in a fundamental way so as to reduce spending on other services and providers.

I have been spending quite a bit of my time in recent years thinking about bundled payment for primary care, and moving away from fee-for-service payment, which is recommended not only by the Ellis and McGuire model, but also by a growing number of policymakers. Bundled primary care payment is very challenging to do well. Risk adjustment is important, because you can't just pay a fixed dollar amount per member per month and expect the doctor to be indifferent between managing the primary care needs of a twenty year old healthy person who comes in every three years and the fifty-five or seventy year old who requires monthly visits. Part of the challenge is that existing payments to primary care practitioners in the US do not give PCPs enough incentive to spend the time and effort at controlling costs and promoting healthy lifestyles that are so essential for a high quality health care system, so a different metric of spending is needed. Defining what should be included and excluded from the bundled payment is another challenge. Linking bundled payment with sizable performance rewards for moving care in the desired direction is also important. And there are certainly other challenges: administrative support, information systems, and inadequate data. While there are definitely a growing number of demonstrations going on in the US, we are still only beginning to learn what works. We can also learn from the UK, Netherlands and Denmark, where they have also been implementing changes. Denmark is one of my favorite reforms, because they have already been using a mixed payment system, with PCPs receiving capitation payments along with an additional low fee for each visit. Much can be learned from these countries.

I have worked a great deal on this topic with my colleague Arlene Ash, and with frequent conversations and inspiration from physician Allan Goroll and colleagues (2007, 2008). In Ash and Ellis (2012) we develop a framework for risk adjusting the bundled payment for primary care and introduce a new construct that we call the PCAL, the primary care activity level. (We

used to call this measure the patient primary care "burden level," but no patient wants to think of themselves as a burden.) We calculate the PCAL so as to approximate the level of resources needed by a PCP to provide high quality primary care, which is calibrated to exceed the existing level of payments to PCPs. An important contribution is that we do not try to bundle all kinds of services the patient receives, only those to be included in the bundled primary care payment. The metric we predict is a weighted sum of existing spending, not only including all spending on primary care services, but also using low weights on selected inpatient, specialist, laboratory, imaging, and pharmacy spending. We do this not because we want the PCP to bear financial responsibility for spending on all of these services, as other primary care capitation projects have attempted, but rather because patients who are likely to use a lot of these other services are also likely to require a great deal of primary care attention.

Developing a bundled primary care payment model is not easy. One challenge is to identify the specific weighted sum of services to use as a proxy for the PCAL. Other challenges, discussed in Ellis and Ash (2012) is that in many plans in the US it is difficult to assign people to a PCP in order to define the total volume of patient activity to be ascribed to the PCP, and calculate performance measures. Also, most physicians see patients from many different payers (private, Medicare, Medicaid, etc.) as well as many different benefit plans (HMOs, PPOs, FFS, POS); accommodating this diversity in a single bundled payment formula is a further challenge. Rather than let imperfect information and complexity prevent us from moving forward, in Ash and Ellis (2012) we used Thomson-Reuters MarketScan commercial claims and encounter data to calculate PCAL payments, assign patients to PCPs, and examine the PCAL model performance on PCPs from three primary care specialties: internal medicine, family medicine, and pediatrics. We are fortunate to have had to opportunity to assist the Capital District Physicians' Health Plan implement many of our ideas for the PCAL model as part of a patient-centered medical home demonstration in New York State in 2009.

In Ash and Ellis (2012) we evaluate the performance of the PCAL model at the PCP level, and demonstrate that for moderate size practices with 500-1500 patient enrollees we can achieve physician level R-squares of 78% in a sample of 436 PCP practices. This high R-square between the predicted PCAL and the PCAL proxy, and the correspondingly high correlation between our PCAL and current FFS revenue to PCPs is important because we are asking physicians to accept this bundled payment to largely replace their existing fee revenue. Most health economists are paid a salary, and if somebody cooked up a payment formula to replace that salary, you would want it to be highly correlated with your existing income (or higher) and relatively stable from year to year. Our PCAL model appears to have that property of high correlation and relative yearly stability. We also demonstrate in that paper that the payment formula is relatively fair

across provider specialties, health plan types, physician practice sizes, and patient age and gender.

Paying PCPs a bundled payment may free them up to have time for improved patient care, but it may not reduce total spending or improve patient satisfaction. To do that, it may be desirable to also assess PCP performance and implement meaningful bonuses for improved quality, cost and patient experience. Here again, our innovation (building on Goroll et al) is to emphasize that all bonus payments should be risk adjusted, and PCPs should be rewarded for doing "better than expected" on a wide array of performance measures. As a proof of concept, in Ash and Ellis we also develop risk adjustment models for five utilization measures including hospital admissions, emergency department visits, and spending on antibiotics of concern, whether high cost or dangerous. What we found is that individually tailoring risk adjustment models using HCCs to predict each of these outcomes can greatly improve model fit, and that when aggregated to the practice level can explain a substantial amount of the practice level variation in these outcomes in most cases. The one major exception is emergency department visits, where the individual level R-square is only 3 percent and the practice level R-square is only 17%: counts of ED visits are not well explained by the health status of the enrollee, but rather seem to reflect demographic, provider, and environmental factors. Incorporating broader sets of variables in the risk adjustment models used for primary care payments and performance assessment is on our agenda for the future, and an important limitation on the work we have done so far.

This leads me to my fourth question.

Q4 How well is competition working at keeping down the costs of health care in the market where you live?

I believe in competition, and would like to see it work at holding down health care costs. But is it working? When I look at the Boston area, where I work and use health care services, I see a market dominated by two very strong hospital networks that own large numbers of private practices and specialty clinics, together with a small number of health plans. I see costs that are well above the national average. I tend to think that I may be getting a lot of quality, but not a lot of cost containment. Competition is not really working well. Perhaps instead of asking how we can make the market more competitive so as to hold down costs better, we should be asking whether there are alternatives to competition that could do an even better job at holding down costs, even if it may mean less choice or more reliance on regulation.

When I review the current US efforts at promoting plan-level competition, and making consumers more price conscious at both the plan and provider levels, I see many of the existing

arguments as having origins in the work on managed competition laid out by Alain Enthoven and Rick Kronick in their two articles in the *New England Journal of Medicine* in 1989. We certainly have not implemented their vision for managed competition, but we have definitely been influenced by it. Can managed competition really work in the US? One answer to this question was provided long ago by Enthoven's coauthor and former graduate student Rick Kronick along with David Goodman, John Wennberg, and Edward Wagner in a 1993 *NEJM* reply. They estimated that there are ten entire states, and altogether 37 percent of the US population living in markets where the number of people in the market is not sufficient (>360,000) to support at least three networks of doctors and hospitals, especially specialists. With the increased specialization of clinics and hospitals in the intervening twenty years, it seems likely that even fewer people today would live in sufficiently concentrated areas where competition could potentially work. So using their 1993 framework, perhaps 60 percent of the US population currently lives in large markets that could potentially rely on managed competition in today's world. Are these large markets workably competitive?

Recent evidence is provided by Leemore Dafny, Mark Duggan, and Subramaniam Ramanarayanan in the *American Economic Review* 2012. They note that the American Medical Association reports that 94% of 314 US market areas were found to have "highly concentrated" insurance markets. Dafny et al.'s own analysis used a large employer health insurer dataset from 1998-2006 spanning 139 market areas. For health plans, they found the fraction of markets falling into the top "highly concentrated" category (defined as a HHI > 1800) increased from 69 to 99%, the median four-firm concentration ratio increased from 79 to 90%, and the proportion of large firms self-insuring increased from 55 to 80%. This last point is of special interest, because if we think that insurance companies are going to control costs, then they have to have an incentive to do so. But if insurance companies are merely serving as claims processors charging a percent markup over claims as their fee, they have no incentive to reduce health care payments. Thinking about how competition between diffuse claims processors will work to control costs is problematic.

Further discouraging evidence about the prospects for competition is in Robert Kocher, and Nikhil Sahni, *NEJM*, 2011 who note that the past decade has seen a 75% increase in hospital-owned doctor practices, with more than half of all physicians practices now owned by hospitals. Ownership of primary care practices by hospitals and multi-specialty physician networks is particularly problematic for the bundled primary care payment models that I favor since it's hard to imagine PCPs competing for patients or striving to reduce hospitalizations and specialist referrals independently from the hospitals and provider groups who own them.

If the structural evidence for effective competition is concerning, the literature on consumer behavior also highlights the challenges of getting consumers to make effective, cost-motivated choices. There is growing evidence that consumers do not switch plans or providers easily and often make what appear to be irrational choices. Recent examples in the literature of this include Schlesinger (2010), and Sinaiko and Hirth (2011), Abaluck and Gruber (2012), and Baicker, Mullainathan, Schwartzstein (2012).

I am going to take this opportunity to highlight my own study done back in 1989 that also suggests that inertia can be extreme, decisions irrational, and biased selection severe. Ellis (1989) documented an extreme case of biased selection in one natural experiment when employees of a large New York City financial institution were given a relatively simple choice of plans that differed only in their deductibles and stoplosses. Partly as an example of how times have changed, it is interesting to look at a few of the numbers. The employer was experimenting by offering a flexible benefits plan in which employees were encouraged to move out of a low deductible plan and choose higher deductibles. At the time, the bank had a $100 deductible plan with a $1000 stoploss, and the innovation was to offer $100, $200 and $300 deductible plans. For the higher deductible plans the stoploss remained at $1000, while for the $100 deductible plan the stoploss was raised to $2000, hence all of the plans were worse than the initial plan. In 1982, the year before the choice was offered the plan premium was $85, and in 1983 after the reform, the three plan premiums were $156, $96, and $80 respectively. (It is striking to note that these were annual, not monthly, individual premiums.) The heart of my paper was to conduct simulations for various types of enrollees that demonstrated that remaining in the $100 deductible plan was never an optimal choice, because of its higher premium and greater exposure to high out-of-pocket costs. Nonetheless 31% of the individual enrollees either actively chose this inferior plan (17%) or defaulted into it (14%) when they refused to fill out the form and indicate an active choice. A further result of that analysis was that patterns of choices by people with different levels of prior year health care spending suggested a non-parametric loss function with risk loving behavior toward large losses, a feature often found in the behavioral economics literature.

The health plan enrollees in this 1983 experiment sorted themselves remarkably well when offered a simple choice that required only a comparison of deductibles and stoplosses and their own health spending experience. The sample mean total covered costs across plans in 1982 displayed an 8 to1 ratio before the plan choice and a 5 to 1 ratio in the year after the choices were made. This is a reminder that enrollees can anticipate their spending the following year pretty well for some purposes. We tend to confuse this fact because plan choices in a modern setting are now complicated by numerous benefit design differences (in ten dimensions rather than two), and choices often include HMO, PPO, POS, or health savings accounts that differ so

much in their complexity that they result in less perfect sorting. Fundamentally, most consumers do have a good sense of what their spending the following year will be, primarily because most health care spending is for chronic rather than acute conditions.

This leads me to my fifth and final question.

Q5 Why do economists spend so much time looking under lampposts?

I suspect you've all heard the joke about the drunk who loses his keys. He's out looking for them under a lamppost, when a policeman comes along and says, "What are you doing?"

"I'm looking for my keys."

The policeman joins him in looking for a while and then asks, "Are you sure you lost your keys just here?"

And the drunk replies, "Oh, no, I lost them across the street, but the light is so much better over here under the lamppost."

I think that most health economists, including me, are guilty of spending a lot of their time building beautiful, elegant models of perfect foresight, hyperbolic discounting and expected utility maximization even though we know about the existence of abundant evidence that the world does not really work that way. We find it inconvenient to deal with all the behavioral and non-conventional ways of thinking about the problem and stick with the beauty of the neoclassical economist's world.

My last question is intended to encourage you to spend serious time thinking about the wonderful insights from behavioral economics. I'm delighted to see so many of the papers in the ASHEcon program that touch on behavioral economics, but thought I would perhaps tweak the interest of a few more people by mentioning three of my favorite books, even if many of you are already aware of them.

There's a terrific book, "Thinking, Fast and Slow" by Daniel Kahneman, that came out in 2011 and summarizes the amazingly rich set of insights and research findings of Kahneman and others who have helped developed this field since the 1970s and 80s. The fundamental thesis of the Kahneman book is that there are two different mental processes that drive the way we think. System 1 performs fast, intuitive and emotional judgments, where it is important to quickly process and make decisions in the presence of the overflow of information that constantly surrounds us. System 2, is slower, more deliberative, and logical, the type of analytical mind which we economists excel at using for interpreting the world, and which contributes to our view of the world as largely rational. Kahneman's book is full of insights

about decisions and behavior that are not well-captured by only thinking about the analytical System 2. I clearly don't have time to go over all of his insights, but I have included a list of key concepts which is presented in Table 1. I encourage you to go through the book and pull out the different insights. Each of us may have been aware of a few of these concepts, but it was really quite stunning to me to realize the breadth of ideas that Daniel Kahneman has been working on throughout his lifetime. The ideas have been out there for a long time; they just have not been well-integrated into all of our health economics literature.

A second book that is of great interest is "Nudge…" by Richard Thaler and Kass Sunstein (2008). This book has been out for a while, and even includes the word health in the extended title, so perhaps you have already had a chance to read it. It provides great insights on how choice architecture (which may combine priming, framing, and cognitive ease) can improve health and many other lifestyle choices. It will be particularly important for thinking about how to set up health insurance exchanges.

A final book with a behavioral economics emphasis is Dan Ariely's (2008) book, "Predictably Irrational." I want to end by illustrating how the insights from Ariely's book may relate back to a classic in the literature by Will Manning, Joe Newhouse and others from AER in 1987. We've all seen pictures of the nonlinear demand curve for medical care from the Rand Health Insurance Experiment shown in Figure 6, which highlights that, yes, the demand for medical care is price elastic, and for low levels of cost-sharing it's more responsive than for higher levels of cost sharing. I agree completely with Manning et al that the demand curve is non-linear, but let's think: What does the behavioral economics literature suggest? One of the key concepts in Dan Ariely's book is that "free" is fundamentally different from just offering services at a low price. Marketing people understood this concept decades ago. We see signs of it every day: buy one get one free, get 30% more shampoo for free, buy three tires and get one free. Marketers try to avoid saying "one-third off." Instead they use the word "free" over and over again on their packaging and advertisements, because the beauty of "free" is you can disengage your mindset and think: "Oh, it's free, I don't have to worry about what it costs." That's what "free" does; it liberates your mind to not even think about cost.

So if we take the classic demand curve, we might ask: What if all the action is happening at the first dollar of cost sharing. Perhaps Figure 7 is the correct one, there is a discontinuity, and what really matters to patients (or to the doctors who are recommending more treatment) is whether the treatment costs you something at all. If the price is positive, then you have to at least pause long enough to ask, "Well, is it worth the dollar, or the ten dollar visit fee?" But if the service is free, then you don't have to ask that question. Perhaps this helps explain why most HMOs and PPOs have moved away from a percentage copayment toward a fixed dollar

amount for your visit, typically $10 or $20. It makes you think about whether you want to make that first visit. And that gets you to pause before that first visit. But by moving away from the 20% co-payment world that we used to live in, they've removed the effort of questioning everything else that is done to you. You no longer need to ask about the value of one more lab test, or one more procedure. After the initial visit fee, all of the additional services are free. Perhaps the medical system has taken advantage and knows that by not having to raise the question of the cost of a service they can avoid having people become price-responsive.

One final thought in the same vein. What if something similar is going on for physician behavior in response to a mixed payment system? Perhaps the physician's supply curve also contains a discontinuity at the marginal cost, so that the quantity supplied increases discontinuously when the marginal reimbursement to doctors exceeds the marginal cost. If the marginal loss to the provider, or the bonus penalty to the doctor were even a few dollars for referring to a particular lab test, doctors might be less eager to recommend as many lab tests or as many specialist visits.

I will end by summarizing some of the key points I have made. First, health economists could pay more attention to variation in states of the world than taste heterogeneity, worrying more about non-financial health shocks and bankruptcy costs. Supply-side cost sharing remains attractive as an alternative to demand-side cost sharing because of the reduced financial risk on consumers. American economists have spent a great deal of time studying whether competition promotes high quality or lowers costs (often not) relative to examining alternatives to competition given that competition is generally not working well. Many of the most interesting payment and insurance innovations will require careful risk adjustment, especially in the new area of bundled primary care payment. Behavioral economics continues to provide new insights into how health care choices are made, and we would all benefit from examining and building upon its rich set of concepts. As one specific example, health economists have done remarkably little work in worrying about discontinuities in health care markets, both at the zero prices for consumers, and for supply prices at or below marginal cost on the supply side.

Figure 1: A health shock increases the demand for health care

Figure 2: Allowing ex post plan choice worsens the moral hazard problem.

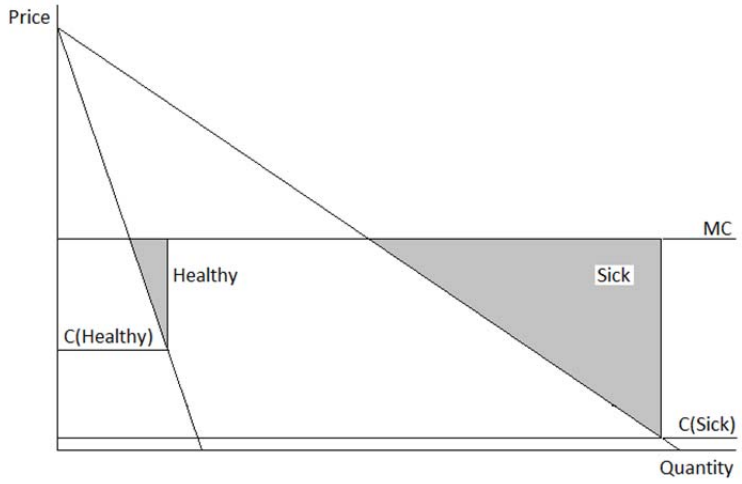
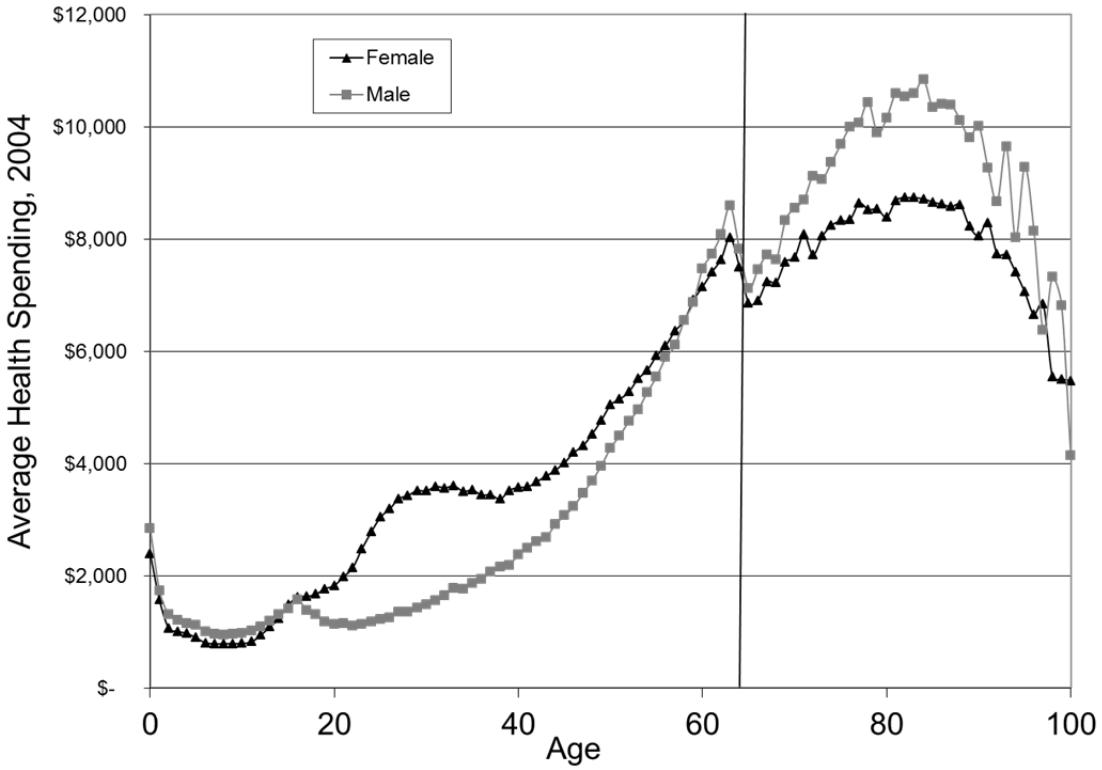
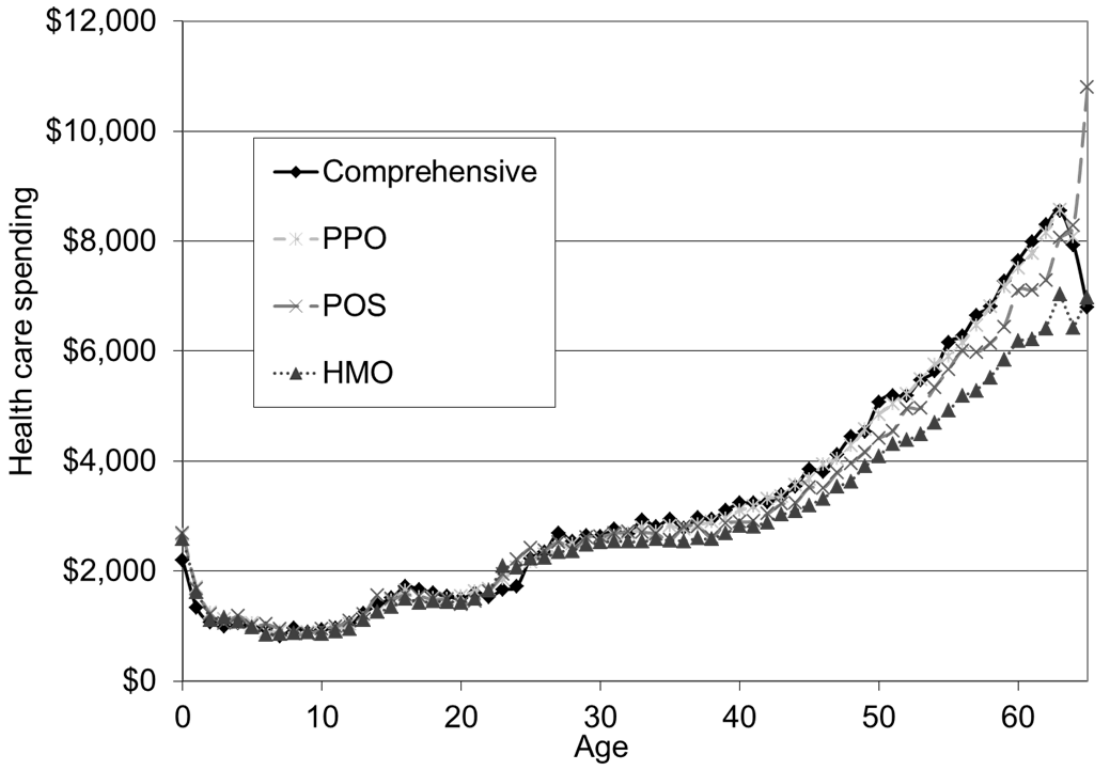


Figure 3: US Privately-Insured Health Care Spending, by Age and Gender



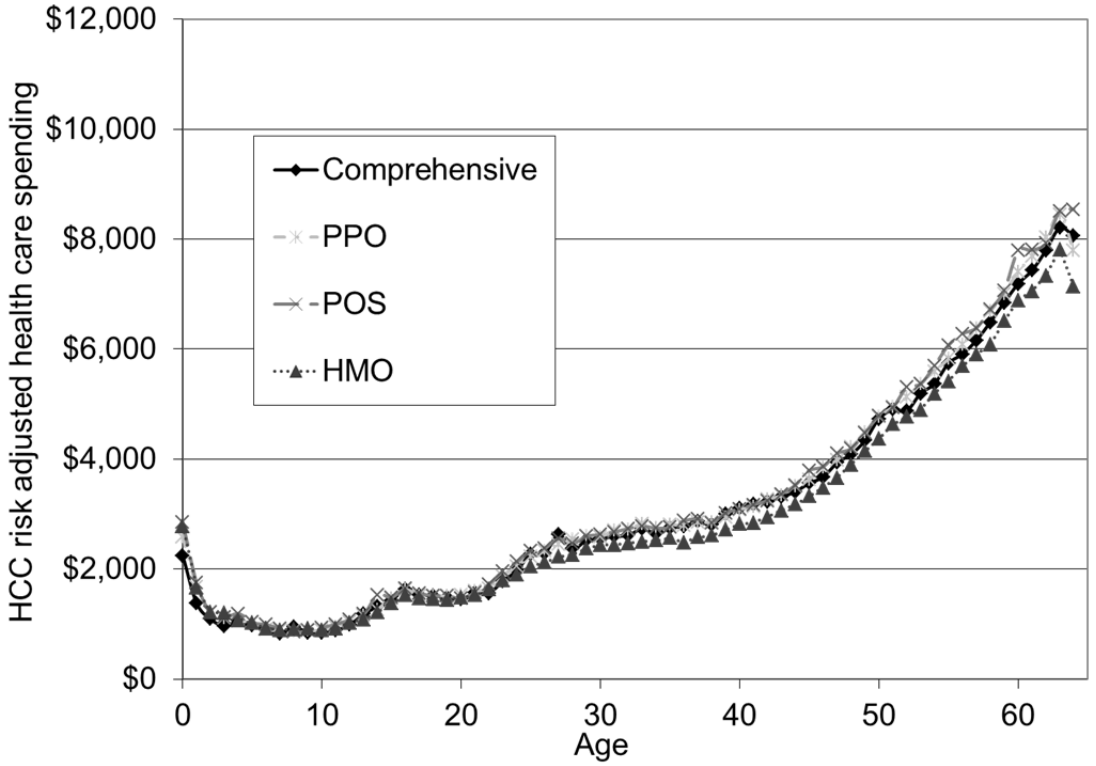
Note: Each point plots the average annualized health care spending (i.e., allowed charges) per year for men and women by one-year age levels in using 2004 MEDSTAT (Thomson-Reuters) MarketScan commercial claims and encounter data, N=14.6 million.

Figure 4: Unadjusted US Privately-Insured Health Care Spending by Age, by Health Plan Type



Note: Each point plots the average annualized health care spending (i.e., allowed charges) per year by health plan type by one-year age levels in using 2004 MEDSTAT (Thomson-Reuters) MarketScan commercial claims and encounter data, N=13.0 million, after dropping low frequency health plan types.

Figure 5: Risk-adjusted US Privately-Insured Health Care Spending by Age, by Health Plan Type



Notes: Each point plots the risk adjusted average annualized health care spending (i.e., allowed charges) per year by health plan type by one-year age levels in the 2004 MEDSTAT (Thomson-Reuters) MarketScan commercial claims and encounter data, N=13.0 million. Risk adjustment was done by dividing the average spending for each plan type by that plan type's average relative risk score for that age, and then multiplying this ratio by the overall (all plan type) average age-specific relative risk score. Risk scores calculated using the concurrent HCC model of DxCG Release 6.1 software. Source: Ellis (2008).

Table 1. Remarkable array of “irrational” behavior discussed in Kahneman (2011)

i)	**Priming**
ii)	**Framing**
iii)	**Base rate neglect**
iv)	**Cognitive ease**
v)	**Anchors**
vi)	**Availability bias**
vii)	**Intuitive prediction**
viii)	**Optimism bias**
ix)	**Prospect theory**
x)	**The endowment effect**
xi)	**The possibility effect**
xii)	**The certainty effect**
xiii)	**Overestimation/over-weighting rare events**
xiv)	**Avoiding regret**
xv)	**WYSIATI (What you see is all there is)**

Figure 6. Demand for outpatient care from the Rand Health Insurance Experiment from Manning et al (1987)

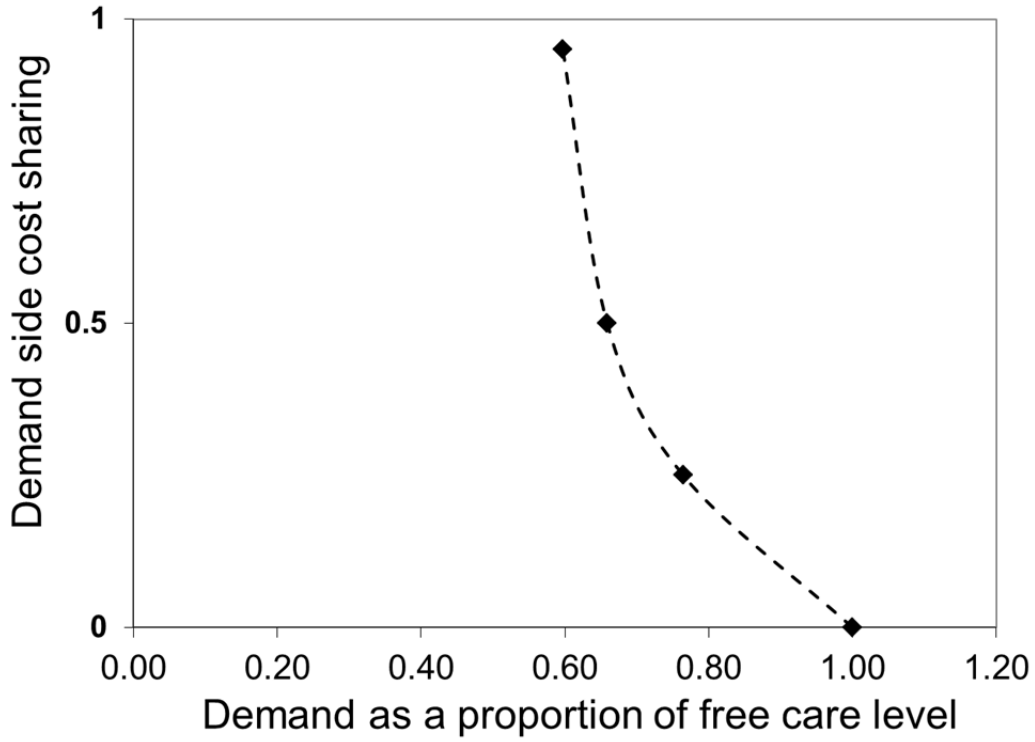


Figure 7. Possible demand for outpatient care consistent with Manning et al (1987)

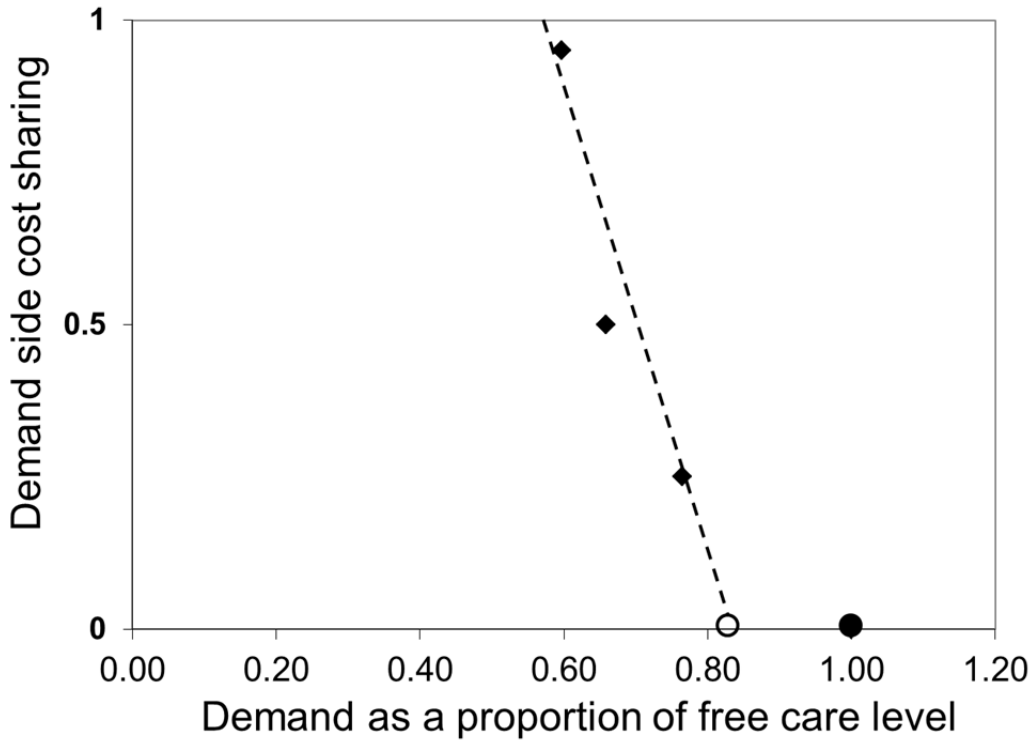
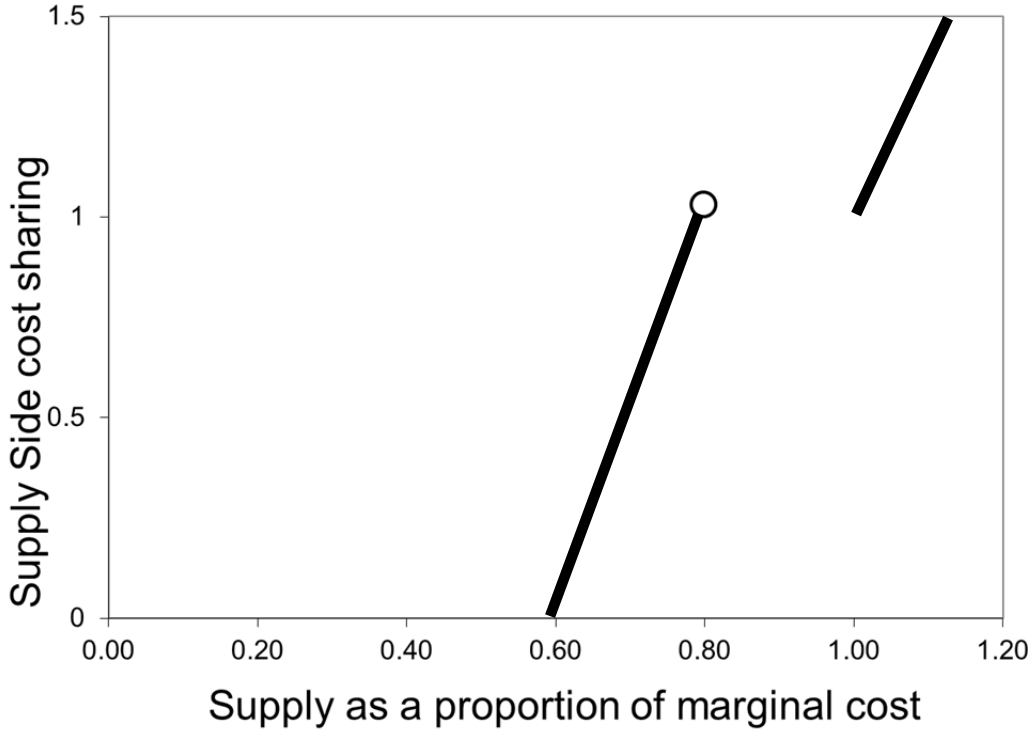


Figure 8. Possible supply curves in response to supply-side cost sharing



References

Abaluck, J., & Gruber, J. (2011) Choice Inconsistencies among the Elderly: Evidence from Plan Choice in the Medicare Part D Program. *American Economic Review* 101 (June): 1180–1210.

Ariely, D. (2008) *Predictably Irrational: The hidden forces that shape our decisions*. New York, New York: Harper.

Ash, A. S. & Ellis, R.P. (2012) Risk-adjusted payment and performance assessment for primary care. *Medical Care*, vol. 50(8):643-653.

Baicker, K., Mullainathan, S.,Schwartzstein, J. (2012) "Choice Hazard in Health Insurance" Dartmouth working paper, August 13.

Dafny, L. S., Duggan, M. G. & Ramanarayanan, S. (2012) Paying a premium on your premium? Consolidation in the US health insurance industry. *American Economic Review*, vol. 102(2):1161-85.

Ellis, R P. (1989) Employee Choice of Health Insurance. *Review of Economics and Statistics*, vol. 71(2):215-23.

Ellis, R. P., & Ash, A. S. (2012) Payments in Support of Effective Primary Care for Chronic Conditions. *Nordic Economic Policy Review*.

Ellis, R. P., & Manning, W. G. (2007). Optimal health insurance for prevention and treatment. *Journal of Health Economics*, vol. 26(6):1128-1150.

Ellis, R. P. & McGuire, T. G., (1986) Provider behavior under prospective payment: Cost sharing and supply. *Journal of Health Economics*, vol. 5(2):129-151.

Enthoven, A. C. & Kronick, R. G. (1989) A consumer-choice health plan for the 1990s: Universal health insurance in a system designed to promote quality and economy. *New England Journal of Medicine*, vol. 320(2):94-101.

Glazer, J, & McGuire, T. G. (2000) “Optimal Risk Adjustment in Markets with Adverse Selection: an Application to Managed Health Care.” *American Economic Review*, 90(4): 1055-1071.

Goroll, A. H., Berenson, R. A. Schoenbaum, S. C, & Gardner, L. B. (2007) Fundamental reform of payment for adult primary care: comprehensive payment for comprehensive care, *Journal of General Internal Medicine*, 22, 410-415.

Goroll, A. H. (2008) Reforming physician payment, *New England Journal of Medicine*, 359, 2087-2090.

Himmelstein, D. U., Thorne, D., Warren, E., & Woolhandler, S. (2009) Medical bankruptcy in the United States, 2007: Results of a national study. *American Journal of Medicine*, vol. 122(8):741-6.

Kahneman, D. (2011) *Thinking, Fast and Slow*. New York, New York: Farrar, Straus and Giroux.

Kocher, R., & Sahni, N. R. (2011) Hospitals' race to employ physicians - The logic behind a money-losing proposition. *New England Journal of Medicine*, vol. 364(19):1790-1793.

Manning, W. G., Newhouse, J. P., Duan, N., Keeler, E. B., Leibowitz, A., & Marquis, M. S. (1987) Health insurance and the demand for medical care: Evidence from a randomized experiment. *American Economic Review*, vol. 77(3):251-277.

Miller, M. M. (2012) Who files for bankruptcy? State laws and characteristics of bankrupt households. Rutgers Working Paper (Available at: http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1983503)

Nyman, J. (2003) *The Theory of Demand for Health Insurance*, Stanford U Press.

Sinaiko, A. D., & Hirth, R. A. (2011) Consumers, health insurance and dominated choices. *Journal of Health Economics*, 30(2): 450–457.

Schlesinger, M. J. (2010) Choice cuts: parsing policymakers' pursuit of patient empowerment from an individual perspective. *Health Economics, Policy and Law*, 5, 365-387 doi:10.1017/S174413311000006X

Thaler, R. H., & Sunstein, C. R. (2008) *Nudge: Improving Decisions About Health, Wealth, and Happiness.* Yale University Press.