
Volume 26, issue 6, 1 December 2007 ISSN 0167-6296
ELSEVIER
JOURNAL OF
HEALTH
ECONOMICS
25th Anniversary Issue

Available online at www.sciencedirect.com

ScienceDirect

ELSEVIER

Journal of Health Economics 26 (2007) 1128–1150

JOURNAL OF HEALTH ECONOMICS

www.elsevier.com/locate/econbase

Optimal health insurance for prevention and treatment☆

Randall P. Ellis [a,*], Willard G. Manning [b]

[a] *Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215, USA*
[b] *Harris School of Public Policy Studies, The University of Chicago, 1155 East 60th Street, Room 176, Chicago, IL 60637, USA*

Received 23 May 2007; received in revised form 5 September 2007; accepted 7 September 2007
Available online 2 October 2007

Abstract

This paper reexamines the efficiency-based arguments for optimal health insurance, extending the classic analysis to consider optimal coverage for prevention and treatment separately. Our paper considers the tradeoff between individuals' risk reduction on the one hand, and both *ex ante* and *ex post* moral hazard on the other. We demonstrate that it is always desirable to offer at least some insurance coverage for preventive care if individual consumers ignore the impact of their preventive care on the health premium. Using a utility-based framework, we reconfirm the conventional tradeoff between risk avoidance (by risk sharing) and moral hazard for insuring treatment goods. Uncompensated losses that reduce effective income provide a new efficiency-based argument for more generous insurance coverage for prevention and treatment of health conditions. The optimal coinsurance rates for prevention and for treatment are not identical.

JEL classification: I11; D80; L80

Keywords: Insurance; Prevention; Moral hazard; Risk aversion

1. Introduction

One of the major themes in health economics since the field began has been the behavior of patients and providers in the presence of health insurance that covers part or all of the cost of healthcare. Much of the economic literature on optimal health insurance focuses on "the fundamental tradeoff of risk spreading and appropriate incentives"; see Cutler and Zeckhauser

☆ An earlier version of this paper was presented at the Harvard Symposium and Dinner Celebrating the 25th Anniversary of the *Journal of Health Economics*, April 20, 2007.

* Corresponding author. Tel.: +1 617 353 2741; fax: +1 617 353 4449.
E-mail addresses: ellisrp@bu.edu (R.P. Ellis), w-manning@uchicago.edu (W.G. Manning).

0167-6296/$ – see front matter
doi:10.1016/j.jhealeco.2007.09.002

(2000, p. 576) for a review of this literature. Specifically, it examines either the dead-weight losses from moral hazard, the tradeoff between moral hazard and the gains from insuring against financial risk, or the differential coverage of multiple goods with varying degrees of risk.

Our primary interest in this paper has to do with extending the literature to directly address issues of optimal insurance in markets with two healthcare goods, namely preventive activities/care and treatment of poor health when it arises. For this paper we define preventive care to mean a service with no direct utility or benefit other than it reduces the probability of being in relatively sick states of the world. Although our main focus is on primary prevention, which prevents bad health states from happening altogther (as with immunization), a simple extension of our model partially captures certain dimensions of secondary prevention which include the detection and prevention of a deteriation of a chronic disease.

In the process of deriving an expression for the optimal coverage of preventive care, we contrast how privately and socially optimal choices of preventive care differ, and how these choices are affected by the presence of insurance coverage for treatment services. We also examine how the optimal coverage of preventive care should be influenced by cost sharing on healthcare treatment in a second best optimal insurance policy that covers both prevention and treatment. One of the questions that we answer is: should preventive services be covered at all? Should they be covered to the same extent as other healthcare treatments, such as for accidents, curative care, or paliative care to relieve pain? There is a view held by some in the field that prevention is less uncertain than illness itself and may thus merit less generous coverage. Others have argued for coverage for prevention based on criteria other than economic efficiency (e.g., concerns about cost sharing hampering compliance among those with serious chronic health conditions). However, few have addressed the efficiency arguments for covering prevention based on an expected utility framework.

We present a series of models where consumers' choices about prevention affect expected health status as well as affect expenditures on healthcare treatment. Risk averse consumers value their health, their consumption of non-health goods and services, and protection from financial risks. We provide theoretical support for coverage of prevention to reduce insurance premiums and the cost of bearing risk, especially when the individual's premiums do not fully reflect savings from his or her own individual preventive activity.

This paper makes three basic contributions to the existing literature. First, we differentiate coverage for prevention from that for treatment in determining optimal insurance. Second, we examine how health insurance coverage for both prevention and healthcare treatment are influenced by the presence of uncompensated losses. For example, in a world where consumers are imperfectly insured for loss of income from ill health, or there are uncompensated healthcare treament costs (such as for over-the-counter drugs and supplies) how do these uncompensated losses affect the optimal cost sharing on healthcare treatment? In a parallel manner, how do uncovered costs of prevention (the time and discomfort costs of screening tests, for instance) affect optimal cost sharing on preventive services and healthcare treatment? The third contribution is that we explicitly examine the implications of non-monetary losses, such as blindness or pain, that healthcare treatment may ameliorate but not eliminate. How do such direct utility losses of poor health affect optimal insurance design? As we show below, uncompensated treatment and prevention losses provide an additional rationale for reducing cost sharing both for both preventive care and healthcare treatment goods, while direct utility losses impact optimal preventive care cost sharing only.

Since optimal insurance is a topic of considerable interest to many researchers and policymakers – both economists and non-economists – much has been written on this topic. We do not address these other rationales for insurance coverage that can be found in the health economics and public health literatures. These other important rationales for insurance coverage include correcting for externalities, such as those that can occur with communicable diseases; altruism or public good arguments; corrections of informational problems (i.e., uninsured consumers may make the wrong *ex ante* decisions about health insurance or preventive activity); distributional concerns that may underlie some forms of social insurance (such as goals of elimination of poverty, or achieving social solidarity); or fostering more complete coordination among healthcare providers and patients. Without denying the relevance of these other arguments for more generous health insurance coverage, we reexamine the efficiency-based arguments for insurance, and derive new results which refine our understanding of the value of generous insurance coverage from the consumer's point of view in the absence of other market or social imperfections.

2. Literature and theory review

There is a relatively small literature on "economically optimal" insurance for prevention services *per se*. Some of what is there suggests little or no coverage for prevention. Zweifel and Breyer (1997) suggest that because preventive care is a choice itself by the consumer, it is not itself subject to uncertainty[1] and therefore is uninsurable in the usual insurance sense of an insurable risk. In his review of the literature on prevention, Kenkel (2000) does not highlight any literature on the optimal coverage for preventive services, although he does review the diverse factors that should influence the coverage of prevention.

There is a substantial literature on the overall tradeoff between the welfare losses from moral hazard and the welfare gains from insuring against the riskiness of healthcare expenditures, starting with the seminal work by Arrow (1963, 1971, 1976) and Zeckhauser (1970); see Cutler and Zeckhauser (2000) for a detailed review of this literature. Much of this work has been based on either a one-period model or a two-period model where the consumer selects the coinsurance before knowing his or her realized state of health. Healthcare expenditures are chosen conditional on the state of the world that occurs. The common conclusion of this literature is that one should select the optimal coverage in a plan with a constant copayment or coinsurance rate such that the marginal gains from risk reduction from a change in the coinsurance (copayment) rate just equal the marginal costs of increasing moral hazard. Blomqvist (1997) extends the theory to nonlinear insurance schedules.

The consensus of much of this literature is that insurance does not simply create dead-weight losses because of moral hazard. When the value of avoiding or reducing an individual's risk is included, optimal levels of cost sharing involve neither fully insured (zero out-of-pocket cost), nor being uninsured. Depending on the formal model approach and the data employed, optimal coinsurance rates range from the 50–60% range (Feldstein and Friedman, 1977; Manning and Marquis, 1996) down to values that are in the mid 20% range or lower, possibly with a deductible and/or stop-loss (Blomqvist, 1997; Buchanan et al., 1991; Feldman and Manning, 1997; Newhouse et al., 1993).

Besley (1988) provides a multi-good extension to this literature. His model implies that goods and services that are more uncertain (variable) or less price elastic (with respect to out-of-pocket costs) should have more generous coverage–lower coinsurance rates or copayments. Our results for treatment are not inconsistent with his findings, but the results for prevention are much more complex.

[1] Although its consequences may be.

Recently, Goldman and Philipson (2007) model the case with multiple goods (or technologies) in which the healthcare treatment goods are substitutes or complements. In the case where the goods are substitutes, the insured good should have a lower level of cost sharing than otherwise under traditional rules, because lowering the cost sharing increases the use of those other goods. For complements, the cost sharing should be higher. This clearly has implications for the design of coverage of healthcare services where drug compliance (a form of secondary prevention) is concerned; see their discussion of several recent studies in this area. We return to the contrast with our findings below.

There is also a literature that examines the tradeoff between preventive care and conventional insurance. The classic article in this literature is Ehrlich and Becker (1972), which explores the moral hazard problems caused by insurance in an expected utility model where consumers have uncertainty about two possible states of the world. Consumers can protect themselves from risk by purchasing market insurance, self-insuring, or engaging in self-protection. They use the term "self-protection" to describe what we call preventive care in this paper. They model self-protection as influencing probabilities of different outcomes, but do not consider optimal, market-based insurance for this "self-protection". Self-insurance includes those activities that reduced the size of the loss once the state occurs. This is a close analog to secondary prevention in the public health literature. Ehrlich and Becker find that the effect of market insurance on preventive effort depends on whether the premiums faced by the consumer are adjusted for the individual's level of preventive effort. More recently Barigozzi (2004) has also modeled prevention and optimal insurance issues, without worrying about the interaction between treatment and prevention care coverage. Both papers show that when preventive effort is not observable to the insurer, the presence of market insurance reduces self-protection. We reaffirm the findings from these studies while exploring only models in which preventive effort is unobservable to the insurer.

In the subsequent health economics literature, the effects of insurance on preventive care have come to be referred to as *ex ante* moral hazard. This differentiates it from the type of moral hazard that occurs *ex post* the realization of the health state. Much of the empirical literature on demand for and expenditures on healthcare focuses on either *ex post* moral hazard or some aggregate responses that reflect a combination of the two; see Zweifel and Manning (2000), for a review of this literature. For a survey of the theoretical literature in this area, see Zweifel and Breyer (1997, esp. Chapter 6) for a series of models which explore both *ex ante* and *ex post* moral hazard.

Recently there has been great interest in secondary prevention (Barigozzi, 2004; Kenkel, 2000; Newhouse, 2006) which in our framework differs from primary prevention in two dimensions. First, the consumers who benefit from secondary prevention have already realized an unfavorable state of the world, and preventive care may prevent even worse states of the world in the future. Hence this preventive care does not return a consumer to what their health would have been absent ever having the disease or condition addressed by that specific type of prevention, but only to a better outcome than not having had this type of preventive care.[2] At a cost of complicating our notation, this dimension of secondary prevention is readily added to our model. The other feature of secondary prevention is that often it combines elements of treatment and prevention. Secondary prevention activities such as heart surgery typically provide immediate benefits (increased vigor, reductions in angina) as well changes in probabilities of future states of the world (reduced risk of heart failure). Services which combine treatment and prevention dimensions cannot be readily

[2] This allows for both side effects of prevention, and the detection at an earlier stage of the disease.

incorporated in our theoretical model as we have formulated it. Examining optimal insurance for this type of secondary prevention is an important topic for future research.

Our paper considers the tradeoff between risk protection on the one hand, and both *ex ante* and *ex post* moral hazard on the other. Thus it is related to both the optimal insurance literature for an overall health good, as well as the ones that consider the tradeoff in the context of self-protective effort. Beyond the papers by Besley, by Goldman and Philipson, and the ones that follow the Ehrlich–Becker framework, we have found little work on optimal insurance coverage for prevention that also incorporates coverage for healthcare treatment.

3. Model assumptions

We examine a series of models that involve two health goods, preventive care (Z) and healthcare treatment (X). Appendix Table A.1 lists the notation that we use throughout the paper. The individual's utility function (V) is defined over health state (or health status) and the consumption of other goods and services (Y) other than prevention and healthcare treatment. In the underlying behavioral model, there is a health production function that transforms healthcare X into health status. Preventive activity Z increases the probability of better health $\alpha(Z)$.[3] At the outset, before the state of one's health is determined, the consumer chooses the level of prevention that will affect the probability of various health states (sick vs. well) but not the level or variability of expenditure on healthcare. To simplify notation, we assume initially that prevention has no adverse side effects or expensive false positives: it simply increases the probability of being healthy while reducing the probability of all sick states of the world proportionally. Also to simplify the discussion, we assume initially that there are no healthcare expenditures when the consumer is healthy, but expected healthcare expenditures are positive when sick. Later we return to consider more general cases where healthcare treatment expenditures are positive in all states, and where prevention may increase the probabilities of multiple states of the world. For simplicity, we ignore the possibility of death. Death could be included in an expanded set of health states. We assume that the moments of spending to treat healthcare shocks do not depend on the level of cost sharing, prevention, or income, considering only briefly the case where these variables might also affect the distribution of health shocks, not just consumer choices.[4]

Following much of the literature (for example, Cutler and Zeckhauser, 2000), we examine only health insurance plans with constant coinsurance rates $0 \leq c_X \leq 1$ for healthcare treatment (X) and c_Z for the preventive care good (Z), where $c_Z \leq 1$.[5] We assume premiums, π, are competitively determined and depend on the copayment rates and the demand structure. The premium is (one plus a loading factor δ) times the expected covered healthcare expenditure. The premium does not vary across individuals based on their own individual levels of preventive activity Z_i. The overall premium π does depend on the population level or average of preventive activity. The units of medical care have been normalized so that the market price is 1, which we also assume to be the marginal cost and hence the efficient price.

[3] All that is required to generate the results shown below is that prevention increases the expected health of the consumer, which can happen even if in some states of the world prevention may make the consumer worse off.

[4] We assume that the error term θ in the demand for healthcare X is homoscedastic in cost sharing and prevention. The literature on healthcare expenditures and demand indicates that the variance is an increasing function of the mean response. This implies that the variance should fall as the coinsurance rises. As an alternative, see Feldman and Manning (1997) for such an extension to the basic model based on a constant coefficient of variation property for healthcare expenditures.

[5] The insurance policy is a pure coinsurance plan. It has no deductible, stop-loss, or limit on the maximum expenditure or level(s) of covered services.

In addition to the costs of healthcare treatment and prevention that are relevant from the insurers' point of view, we incorporate the idea that illness may involve other uninsured, uncompensated losses. These losses can be of two types: uncompensated out-of-pocket costs which we assume are proportional to covered costs (denoted as prices L_X and L_Z), and uncompensated health shock losses $L_\theta(\theta)$ that affect consumer utility but do not affect the marginal utility of income.[6] An example of L_X could be the time and travel costs of receiving care, or uncovered over-the-counter medications. An example of L_Z could be the discomfort and time cost of preventive activities such as a colonoscopy or vaccination. An example of the $L_\theta(\theta)$ might be the possible loss of vision even if a detached retina occurs and is optimally treated.

As we noted in the introduction, we do not consider externalities from the consumption of preventive services or healthcare treatment. We also ignore tax subsidies for health insurance and healthcare spending. These assumptions eliminate various arguments for using reduced cost sharing as a Pigouvian remedy for those market failures.

We begin by developing a basic model with one healthcare treatment good that generates the usual results found in the literature. Our basic model uses consumer surplus, Harberger welfare triangles, proportional insurance loading factors, and the Arrow–Pratt approximation for the cost of risk (Pratt, 1962; Arrow, 1963). Results from this basic model serve as a benchmark against which to contrast subsequent results, and are shown in an appendix to be equivalent to a utility-based formulation with special assumptions. We then develop a utility-based model with of prevention and contrast the privately optimal and socially optimal levels of preventive care. We also derive an expression for the optimal coinsurance rate for preventive care.

4. Basic model with one healthcare treatment good and no prevention

As a starting point, consider a basic one-period model with one healthcare treatment good, but no preventive activity. For the welfare loss from moral hazard, we employ the Harberger triangle, following on the work by Buchanan et al. (1991) and Newhouse et al. (1993, Chapter 4). For the cost of risk bearing, we use the Arrow–Pratt measure for constant absolute risk aversion, R^{A}.

Using these approximations for the case of a linear demand curve ($\mathrm{X} = \mu_X - B(c_X + L_X) + \theta$), we can write the various components of the welfare loss to be minimized as follows:

Ex post dead-weight loss from moral hazard $= (1/2)(1 - c_X)^2 B$.
Cost of insurance loading factors $= \delta[(1 - c_X)(\mu_X - B(c_X + L_X))]$.
Ex ante cost of risk bearing (Arrow–Pratt) $= [R^{\mathrm{A}}(c_X + L_X)^2 \sigma_\theta^2/2]$.

Uncompensated losses due to ill health shock $= L_\theta(\theta)$, with expectation $\bar{L}_\theta$.

With this notation, the overall welfare loss WL^1 to be minimized through the choice of c_X can be written:

$$\min \mathrm{WL}^1 = \left(\frac{1}{2}\right)(1 - c_X)^2 B + \delta(1 - c_X)(\mu_X - BL_X - Bc_X) + R^A(c_X + L_X)^2 \frac{\sigma_\theta^2}{2} + \bar{L}_\theta \tag{1}$$

[6] Here the term θ refers to uncertainty that remains in the demand for healthcare and for health-care related other losses, given that the individual is in the sick state *ex post*. This could occur because of residual uncertainty about the effectiveness of medical care for a specific patient's condition. This follows on the suggestion of Dardanoni and Wagstaff (1990).

The first-order condition for the minimization of WL^1 is

$$\text{FOC}^1 = -(1 - c_X)B - \delta(\mu_x - BL_X - 2Bc_X) - \delta B + R^A(c_X + L_X)\sigma_\theta^2 = 0 \tag{2}$$

Eq. (2) can be rearranged to yield the following expression for the optimal coinsurance rate c_X^* for healthcare treatment X:

$$c_X^* = \frac{B + (\mu_X + B)\delta - (B\delta + R^{\text{A}}\sigma_\theta^2)L_X}{B + 2B\delta + R^{\text{A}}\sigma_\theta^2} \tag{3}$$

There are four broad types of terms influencing the optimal coinsurance rate for healthcare treatment c_X^*; demand parameters B and μ_X, the cost of risk $R^{\text{A}}\sigma_\theta^2$, insurance loading factor δ, and uncompensated costs of treatment L_X. We interpret (3) in a series of special cases below.

4.1. Pure case with no insurance loading or uncompensated losses

The simplest case is where there is no insurance loading factor ($\delta = 0$), and no uncompensated health losses that affect income ($L_X = 0$). The optimal choice of cost sharing c_X^* becomes

$$c_X^* = \frac{B}{B + R^{\text{A}}\sigma_\theta^2} \tag{4}$$

As long as the consumer is risk averse ($R^{\text{A}} > 0$), there is any variation in spending ($\sigma_\theta^2 > 0$), and the demand for healthcare treatment is not perfectly inelastic ($B > 0$), then it will be optimal to offer partial insurance with $0 < c_X^* < 1$. Optimal c_X^* is decreasing as R^{A} or σ_θ^2 increase or as demand becomes more elastic ($|B|$ is increasing).

4.2. One healthcare good, with insurance loading

Taking the pure case, we next relax the assumption of no insurance loading factor so that $\delta > 0$. The new expression for the optimal coinsurance rate becomes

$$c_X^* = \frac{B + (\mu_X + B)\delta}{B + 2B\delta + R^{\text{A}}\sigma_\theta^2} \tag{5}$$

New terms involving δ have appeared in the numerator and denominator. Since $\mu_X - B > 0$ is a necessary condition for demand being positive when $c_X = 1$, it follows that $\mu_X + B > 2B$ and hence the optimal cost sharing c_X^* is increasing in δ. As long as the insurance loading factor δ is not prohibitively large, then $0 < c_X^* < 1$. Also, the new term μ_X enters the expression. Insurance loading factors, because they increase the inefficiency of insurance, make it desirable to less fully insure treatment costs as they become more expensive on average.

4.3. Adding uncompensated costs of treatment

In many optimal health insurance models, a fully insured consumer bears no financial risk. This result ignores the real world feature that there are often uncompensated costs of treatment, as we have discussed above. In our derivation of expressions characterizing optimal prevention and optimal cost sharing for prevention, uncompensated costs of prevention and uncompensated costs of health shocks both played an important role of ensuring that even a consumer with zero copayments will wish to purchase some preventive care Z. This happened in our model because the expected utility when sick was less than the expected utility when healthy. Kenkel (2000) also has pointed out that this is probably the key reason why economists' concerns about inefficiencies due to *ex ante* moral hazard may be overstated. In this section, we demonstrate that incorporating uncompensated costs of treatment has a parallel effect, increasing the optimal coverage for treatment services.

Using our general model, we consider the special case where $\delta = 0$ so that there is no insurance loading factor. But we now allow $L_X > 0$. This loss might reflect lost worker productivity (e.g., sick days without pay), the value of time spent receiving treatment, or uncompensated healthcare spending (e.g., over-the-counter medicine or home care). It is readily shown that

$$c_X^* = \frac{B - R^{\mathrm{A}}\sigma_\theta^2 L_X}{B + R^{\mathrm{A}}\sigma_\theta^2} \tag{6}$$

Uncompensated treatment costs reduce the optimal coinsurance rate in a straightforward way, by increasing the cost of risk facing consumers, and making it desirable to transfer more income into less healthy states. Because insurance against these losses is missing, reducing cost sharing on covered treatment services is a second best solution, and cost sharing rates should be kept lower than would occur otherwise. Note that unlike preventive care, health shock losses $L_\theta(\theta)$ do not affect optimal treatment cost sharing.

It is possible for c_X^* to be negative or reach a corner solution where $c_X^* = 0$ for either large L_X or small B. Eq. (6) provides an efficiency-based rationale for why full insurance can be second best optimal in some situations: the absence of complete insurance markets to transfer income into particular ill health states of the world means that coinsurance rates are set at or closer to zero than they would have been if the alternative insurance markets were complete and consumers were able to insure against all healthcare losses. Incomplete insurance markets provide a rationale for more generous insurance coverage of healthcare treatment, even when welfare losses due to moral hazard and insurance loading may be important.

The comparative statics for the optimal cost sharing are summarized in Table 1:

This basic model generates the usual cost sharing results underlying the optimal proportional insurance literature. Optimal cost sharing should be smaller the less price responsive the consumer (because the welfare loss from moral hazard is smaller), the more risk averse the consumer (because the cost of risk bearing to the consumer is higher), the greater the variance σ_θ^2 (because the cost of risk bearing is higher), and the lower the insurance loading factor (the proportional departure from actuarial fair premiums). If there are insurance loading costs ($\delta > 0$), then optimal insurance should be reduced when mean spending is increased. In this simple model if insurance

Table 1
Comparative statics on optimal c_X^* using the basic model

Effect of		on c_X^*
Healthcare treatment demand slope	$\lvert B\rvert$	+
Risk aversion parameter	R^{A}	−
Variance of spending on healthcare treatment	σ_θ^2	−
Healthc are treatment demand intercept	μ_X	+
Insurance loading factor	δ	+
Uncovered costs of healthcare treatment	L_X	−
Uncompensated health losses directly affecting utility	$\bar{L}_\theta$	0

is not actuarially fair and variance of spending is unrelated to mean spending, then increasing the mean cost of the good being insured implies higher cost sharing.

With the exception of the effects of uncompensated costs of treatment L_X, the optimal cost share c_X is always bounded away from zero. Uninsured costs of treatment $L_X > 0$ introduce the possibility that the optimal cost share on treatment may be zero or even negative. If treatment costs covered by insurance are only a part of the total burden facing the consumer, then it is optimal to offer relatively more generous coverage on the insured costs. Hence this model would argue for relatively more generous coverage for treatment services such as oxygen, or prosthetic devices, which are often associated with reduced earning ability or uncompensated health costs.

While we do not formally model the impact of tax subsidies on the insurance choices, their impact follows naturally from this specification. The US tax system affects both health insurance premiums and copayments through the extent of deductibility of premiums and cost sharing payments. From the point of view of private employers and the self-insured consumers, a proportional premium tax subsidy is analogous to a negative insurance loading factor (in our model, $\delta < 0$).[7] Our model confirms the Pauly (1974) result that tax subsidies for premiums that reduce the cost of insurance make it privately optimal to overinsure by choosing cost sharing rates that are inefficiently low relative to the social optimum.

5. A utility-based model of prevention

The basic model provides some important insights, but is conditioned on a given specific probability distribution of health shocks that may themselves be affected by prevention decisions. In this section, we develop a utility-based model in order to make some statements about the role of preventive effort, uncompensated health losses, and the interplay between cost sharing on prevention and treatment goods. Mathematical derivations of the results are presented in an appendix.

The sequence of steps underlying our model is the following:

(1) The social planner chooses the coinsurance rates c_X and c_Z for healthcare treatment and prevention.
(2) The consumer chooses the preventive care level Z to maximize *ex ante* utility. Commitments to pay for π and Z are made before health shocks are known.
(3) Nature decides on the consumer's state of illness as random health shocks θ affecting the demand for a vector of healthcare goods X.
(4) The consumer chooses quantities X and Y to maximize *ex post* utility conditional on health shock θ.

For simplicity, we assume that the demand for health care treatment X is perfectly income inelastic. The evidence from the literature is that medical care demand has some income response, but that is quite inelastic (Newhouse et al., 1993). This assumption about the income elasticity, along with the assumption of a linear response, enables us to derive closed form solutions for the optimal coinsurance rates. We avoid corner solutions by assuming that income is always

[7] We do not model subsidies to out-of-pocket health expenses by the tax system. These could also affect optimal insurance coverage. If we were to include this subsidy, the results would be qualitatively similar to what we present here.

sufficient to pay for at least some of all other goods Y after paying for X and Z at their insured prices.[8]

There are three classes of goods, one medical treatment good X, one composite other good Y, and preventive care Z. After normalizations, $P_Y = 1$, and the consumer's full prices for goods X and Z including uncompensated costs become $P_X = c_X + L_X$ and $P_Z = c_Z + L_Z$, respectively. There are two possible broad states of the world facing the consumer: healthy and sick. Disposable income J available for spending on Y and X is $J = I - \pi - (c_Z + L_Z)Z$, where I is the consumer's income, and π is the insurance premium. Expressions for the premium π are derived below. The indirect utility when healthy is

$$V(I, P, \theta) = V^{\mathrm{H}}(I - \pi - (c_Z + L_Z)Z) \equiv V^{\mathrm{H}}(J) \tag{7}$$

When the consumer is sick, which happens with probability $\alpha(Z)$, the demand curve for each medical service is linear in price the consumer's healthcare price $(c_X + L_X)$, and hence

$$X = A - B(c_X + L_X) \tag{8}$$

where A and B are positive. The indirect utility function consistent with this demand function is shown in the appendix. Stochastic healthcare treatment demand is introduced by letting $A = \mu_X + \theta$ where $\theta \sim F(\theta)$ with $E(\theta) = 0$. We assume that the variance of θ is a constant, and specifically does not depend on the out-of-pocket price or the level of prevention.[9] This corresponds to the horizontal intercept of the demand curves having a mean of μ_X when the out-of-pocket price is zero.

In addition to the uncompensated costs of prevention and treatment, which are proportional to prevention and treatment activities Z and X, we also allow health shocks to directly cause losses in consumer utility independent of the level of medical care. Medical conditions (such as mental illness or injuries) may reduce a person's productivity on the job or ability to obtain work. While some injuries or disabilities may be eligible for imperfect compensation through insurance programs such as disability insurance, many are not. We denote these utility losses as $L_\theta(\theta)$ and assume that they directly affect utility independent of a person's income or levels of treatment. The expected loss is denoted $\bar{L}_\theta$.

In order to introduce risk aversion, we apply a monotonically increasing concave function $V^{\mathrm{S}}(\ldots)$ to the indirect utility function consistent with (8). In the sick state, disposable income available to spend after paying the health insurance premium and preventive care is $J = I - \pi - (c_Z + L_Z)Z$. Using this notation, we write the indirect utility function with one treatment good as

$$\begin{aligned} V^{\mathrm{S}}(J, c_X + L_X, \theta) = V^{\mathrm{S}}\left(J - (c_X + L_X)\mu_X + \frac{B(c_X + L_X)^2}{2} - (c_X + L_X)\theta\right) \\ - L_\theta(\theta) \text{ for } (c_X + L_X)[\mu_X - B(c_X + L_X) + \theta] \le J \end{aligned} \tag{9}$$

[8] Nyman (1999) highlights the importance of insurance when it enables the consumer to purchase healthcare treatment goods that would not be affordable in the absence of insurance, and emphasizes that some of what is often called moral hazard is actually an income effect. Since there are no income effects in our model, we have a pure moral hazard problem.

[9] We return to this issue below.

Using the linear demand equation for X, the insurer's break-even condition for the insurance premium is

$$\pi = (1+\delta)[(1-\alpha(Z))(1-c_X)(\mu_X - B(c_X + L_X)) + (1-c_Z)Z] \tag{10}$$

where δ is the administrative loading factor such that insurance costs proportion δ more than actuarially fair insurance.

Taking expectations and combining (7), (9), and (10), the expected utility when healthy and when sick can be written as

$$\begin{aligned} &EV = \alpha(Z)V^{\mathrm{H}}(J) + (1-\alpha(Z))\{E_\theta[V^{\mathrm{S}}(J-K-(c_X+L_X)\theta) - L_\theta(\theta)]\} \\ &\text{where} \\ &J = I - \pi - (c_Z + L_Z)Z \\ &K = (c_X+L_X)\mu_X - \frac{B(c_X+L_X)^2}{2}, \text{ and} \\ &\pi = (1+\delta)[(1-\alpha(Z))(1-c_X)(\mu_X - B(c_X+L_X)) + (1-c_Z)Z] \end{aligned} \tag{11}$$

While we have used the somewhat restrictive assumptions of linear demand, additive errors, and zero income effects, this specification allows us to derive closed form solutions in an intuitive way. We next characterize private and socially optimal choices of prevention.

5.1. Optimal prevention

Two types of prevention choices are of interest, the private optimum and the social optimum. The privately optimal choice of preventive effort is found by maximizing (11) over Z while taking the premium as given from the individual's perspective. The social optimum is found by maximizing (11) while taking into account the effect of Z on the premiums. It is shown in the appendix that the privately and socially optimal choices of Z condition should satisfy:

$$\alpha'(Z^{\mathrm{Priv}}) = \frac{c_Z + L_Z}{[V^{\mathrm{H}} - E[V^{\mathrm{S}}]]/E[V_I]} \tag{12}$$

$$\alpha'(Z^{\mathrm{Soc}}) = \frac{[\partial\pi/\partial Z + c_Z + L_Z]}{[V^{\mathrm{H}} - E[V^{\mathrm{S}}]]/E[V_I]} \tag{13}$$

where $(\partial\pi/\partial Z) = -\alpha'(Z)(1+\delta)(1-c_X)(\mu_X - B(c_X+L_x)) + (1+\delta)(1-c_Z)$.

Both the private and the social optimums involve the denominator $[V^{\mathrm{H}} - E[V^{\mathrm{S}}]]/E[V_I] \equiv Q$ which is the utility gain from being healthy rather than sick, converted into income units using the expected marginal utility of income. If either $L_X > 0$ or $L_\theta(\theta) > 0$, then Q will be positive even when healthcare treatment is fully insured ($c_X = 0$). The social and private optima differ precisely because the social optimum takes into account the change in the premium due to preventive effort. In the absence of any insurance coverage for preventive care, so that $c_Z = 1$, then $\partial\pi/\partial Z$ will always be negative. For any concave increasing function $\alpha(Z)$, the social optimum will always require more prevention effort than is privately optimal. Hence it is always optimal to offer at last some insurance coverage for preventive care in this model.

Eq. (13) can be rearranged to yield the following equation characterizing the socially optimal Z:

$$[Q + (1+\delta)(1-c_X)(\mu_X - B(c_X+L_X))]\alpha'(Z) = 1 + \delta(1-c_Z) + L_Z \tag{14}$$

Here the left hand side is the marginal benefit of an increase in Z, while the right hand side is the marginal cost. Each extra unit of Z improves the probability of being healthy by $\alpha'(Z)$, which is multiplied by the two terms in square brackets. These consist of the full social value of prevention, which is Q, the monetary value of utility gain from prevention, plus the insurance savings from avoiding being sick. The marginal cost of Z is $1+L_Z$ if there are no loading costs or no insurance for prevention, but if both are present then there is an additional cost of $\delta(1-c_Z)$.

Even with full insurance for healthcare treatment, so that $c_X = 0$, the incentive for preventive care does not drop to zero as long as healthcare alone does not fully restore the individual to the healthy state, i.e., as long as $V^{\mathrm{H}} > V^{\mathrm{S}}$. This will occur if there are ever any uncompensated losses of types 1 or 2 that create a difference between expected utility when healthy and sick.

5.2. *Optimal cost sharing on prevention*

The optimal insurance rate for preventive care, c_Z^* can be found by equating (12) and (13), and then solving for c_Z^*. In the appendix, for our linear demand specification, this is shown to result in the following:

$$c_Z^* = \frac{Q - L_Z(1-c_X)(\mu_X - B(c_X + L_X))}{Q + (1-c_X)(\mu_X - B(c_X + L_X))} \quad \text{where } Q = \frac{[V^{\mathrm{H}} - E_\theta[V^{\mathrm{S}}]]}{E[V_I]} \tag{15}$$

Technically, this is not a closed-form solution for the optimal coinsurance rate for prevention c_Z^*, because it depends on c_X; it applies for any value of c_X, not just its optimal value. If we wished, we could substitute into the above equation the solution for c_X^* from (3) and we would have a closed form expression characterizing c_Z^*. We use this conditional expression here because it facilitates the intuition below when comparing the optimal c_Z^* and c_X^*.

The optimal cost share on preventive care will be strictly less than one as long as $c_X < 1$. This seems plausible: if there is no health insurance, then the consumer bears the full cost of any illness, and makes the correct tradeoff between preventive care and other goods. The term $(1-c_X)(\mu_X - B(c_X+L_X))$ is the insured cost of X, which will be decreasing in c_X as long as demand for treatment is inelastic. Both the numerator and denominator terms involving this term imply that c_Z^* should be increased as c_X is increased. Because Q is increasing in health losses directly affecting utility, then preventive care should be subsidized less as uncompensated losses increase (the consumer has more incentive to do prevention anyway). Greater risk aversion and greater variability in spending when sick affect c_Z^* through Q. Increasing either parameter will tend to increase the difference between healthy and sick states for an incompletely insured consumer. Therefore, c_Z^* will tend to be decreasing in the risk aversion parameter and variance of spending.

An important new insight is the role of uncompensated costs of treatment and prevention, L_X and L_Z. Both uncovered costs affect the optimal insurance coverage for preventive care. As L_Z increases, it is desirable to lower c_Z^*. The reason for this is not to compensate the consumer for these out-of-pocket costs (since doing so is inefficient–both premiums and preventive costs are paid in all states of the world). Rather the reason to lower the prevention cost share is because these costs make it more likely that the consumers will underinvest in prevention, which has the externality of lowering premiums for all. The extent to which L_Z affects c_Z^* depends on the size of the externality caused by insurance on treatment.

Also important are the uncompensated health losses directly affecting utility, which have expectation $\bar{L}_\theta$. These losses ensure that Q will always be strictly positive, even with full insurance and with L_X and L_Z both equal to zero. Uncompensated health losses motivate consumers to do prevention and reduce the need for coverage of prevention.

It may strike some readers as odd that the expression for the optimal coinsurance rate for preventive care, c_Z^*, does not depend on how responsive preventive care Z is to this price, which in our model is $\alpha'(Z)$, as long as preventive care is productive ($\alpha'(Z) > 0$). This happens for the same reason that the efficient competitive price in a supply and demand framework depends only on the marginal cost, not on anything related to demand. Eq. (14) reveals that the optimal choice of Z does depend on how effective preventive care is ($\alpha'(Z)$), but the optimal cost share c_Z^* is not influenced by this probability responsiveness.

Before continuing, we would like to briefly highlight that although we have used a number of very specific assumptions in deriving Eq. (15), not all of these assumptions are necessary for the final result. For this model, we assumed that prevention increases the probability of a single state of the world labeled "healthy" and reduced the probability of all remaining states of the world by an equal proportion. We show in the appendix that a similar result can be derived under more general assumptions. In the more general framework, we relax the assumption of linear demand, and consider more general preventive care activities. Specifically we consider the case where prevention increases the probability of a proper subset of states of the world (healthy set of states), while reducing the probability of a different proper subset of the states of (a relatively sick set of states), which ensures that all probabilities sum to one. The comparative static results in Table 2 continue to hold. In this case the new optimal c_Z^* becomes

$$c_Z^* = \frac{Q - L_Z(1 - c_X)(E(X^{\mathrm{S}}) - E(X^{\mathrm{H}}))}{Q + (1 - c_X)(E(X^{\mathrm{S}}) - E(X^{\mathrm{H}}))} \tag{16}$$

where $Q = [EV^{\mathrm{H}} - EV^{\mathrm{S}}]/EV_I$.

The new insight from this more general formulation is that, in general, it is the difference in expected costs covered by insurance that should drive the choice of optimal preventive coverage. This is consistent with the results in Goldman and Philipson (2007), who argue that healthcare treatment goods that substitute for other treatment goods deserve lower cost sharing. In our case preventive care reduces expected insured costs because of the savings in moving from relatively sick to healthy states, and hence greater coverage of prevention is optimal.

In general, the solutions to Eqs. (3) and (15) will generate different coinsurance rates for health care prevention and treatment (c_Z^* and c_X^*, respectively). The optimal coinsurance rate for health care treatment in Eq. (3) reflects a tradeoff between losses from moral hazard and the gains from reducing financial risk, including those from the uncompensated losses associated with health care

Table 2
Comparative statics on optimal cost sharing for prevention and healthcare treatment (c_Z^* and c_X^*)

Effect of		on c_Z^*	on c_X^*
Healthcare treatment demand slope	$\lvert B \rvert$	+	+
Risk aversion parameter	R^{A}	(−)	−
Variance of spending on healthcare treatment	σ_θ^2	(−)	−
Healthcare treatment demand intercept	μ_X	−	+
Insurance loading factor	δ	+	+
Uncovered costs of healthcare treatment	L_X	(+)	−
Uncovered costs of prevention	L_Z	−	?
Uncompensated health losses directly affecting utility	$\bar{L}_\theta$	+	0

treatment when sick. In contrast the optimal coinsurance rate for prevention in Eq. (15) depends on uncovered costs of prevention L_Z, the amount to which insurance coverage for treatment costs create a wedge between private and socially optimal coverage for treatment, and the money value of the difference in utility of being in the well vs. being in the sick state. Further, the larger the insurer's expected payment for treatment (conditional on being sick) $[(1-c_X)(\mu_X - B(c_X + L_X))]$, the lower the optimal coinsurance rate c_Z for prevention, a variant of the argument that the higher the cost of treatment, the lower should be the out-of-pocket cost paid by the consumer to prevent being sick.

Table 2 summarizes the comparative statics findings in this paper for the various parameters considered for both prevention and healthcare treatment goods. The parameters in the first three rows of Table 2 reaffirm conventional results found in the existing literature, while the terms at the bottom reflect our results that extend the previous literature. Results that can only be shown in special cases of our model are shown in parentheses.[10]

It is well established that optimal cost sharing on healthcare treatment should be higher as demand becomes more elastic consumers become less risk averse, or variance of spending decreases (Besley, 1988; Cutler and Zeckhauser, 2000). We confirm these findings with our model, and also show that these effects hold for special cases for preventive care as well as for healthcare treatment. Our model shows that an increase in the demand for healthcare treatment (by a shift in the intercept μ_X) has different effects on optimal cost sharing for preventive care and treatment. The implied increase in spending on treatment (absent any behavioral response) makes it more attractive to cover preventive care more generously. Because of insurance loading factors, the higher expected costs from a rise in μ_X make it less attractive to cover healthcare treatment as generously. We believe that this finding could differ for alternative demand functions and utility functions. Nevertheless, this result is consistent with intuition for our basic and utility-based models.

Our model also generates different coinsurance rates for preventive care and healthcare treatment when sick. This echoes a similar finding by Eeckhoudt et al. (2004). In our model before uncompensated costs of preventive care are incorporated, there is only one reason why preventive care should be covered: to offset the fact that consumers will ignore the premium savings from preventive effort when they remain healthy. A richer model could add perhaps a second rationale: consumer overdiscounting of future costs and health risks. Such consumer misperceptions provide a rationale for increased insurance for healthcare prevention, as they do for all kinds of time-dependent investments (retirement, education, housing, appliances, etc.) that are not unique to healthcare.

6. Discussion

In this article, we have extended the literature on efficiency-based models of optimal insurance for both health care treatment and prevention, while focusing purely on demand-side incentives. Using a linear demand specification to derive the major findings[11] but which exhibits zero income effects on the demand for healthcare treatment, we have been able to derive closed form solutions for optimal cost sharing on both preventive care and healthcare treatment in a variety of specifications.

Our work indicates that insuring preventive care can be second-best optimal when either of two major conditions hold. First, if preventive activity is not contractable between the insurer

[10] For certain specific functional forms for $V^{\mathrm{H}}(\ldots)$ and $V^{\mathrm{S}}(\ldots)$ we have been able to sign these derivatives. However in general, they require additional assumptions.

[11] A less restrictive version of the prevention model is developed in Appendix A.

and the consumer, then the privately optimal coinsurance rate does not equal the socially optimal rate, because the consumer will ignore his or her impact on premium costs. Second, the second-best optimal coinsurance rate for both treatment and prevention depend on health-related losses typically not covered by health insurance. In fact, if the losses are large and strongly correlated with health care treatment, then it may be optimal to have a negative coinsurance rate for prevention.

Although we have derived these results under a simple set of assumptions, we fully expect that they can be generalized qualitatively to a broader set of circumstances. These would include situations with externalities (contagious disease), inadequate information and other decision-making problems, decisions made by other parties (such as an employer) rather than by a social planner, as well as a wider range of assumptions about the specification of the demand for medical care. Our model examines a simple form of third-party insurance, but could apply in part to managed indemnity, managed care, and various supply-side arrangements. Although the details of each alternative are important, the addition of moral hazard on the part of the consumer (with respect to his or impact on insurance premiums) and non-health losses that are associated with health care use will both require some type of correction in the incentive structures to patients under these more elaborate extensions or assumptions about behavior. For example, if managed care or pay-for-performance rewards providers for keeping up immunization or screening rates, then there is still a concern that the patient may still not fully internalize the costs of his or her actions or to compensate for health-related but not otherwise insurable actions. To do that may require lowering the copayment or coinsurance rate as well. How far would depend on the specifics of the supply-side incentives, as well as the demand parameters that we have modeled.

The new results that we find most interesting are those for uncompensated costs of treatment and prevention, and uncompensated health shock losses that have not received significant attention in the previous literature on optimal insurance. Uncovered costs of treatment and prevention include the time costs of these services, as well as perhaps the uninsured losses of income, or need to purchase uninsured medical care goods (over the counter drugs, for instance). The additional financial risk imposed by these types of losses provides a rationale for reducing cost sharing on both prevention and treatment services. On the other hand, uncompensated health shock losses – direct health losses that do not affect the marginal utility of income – make coverage of preventive care less important and do not affect optimal treatment cost sharing. The intuition is clear. On the one hand, if consumers face uncovered health-related losses, then they already have an incentive to expend effort on prevention, and do not need as much financial inducement to do so. On the other hand, if consumers face uncertain income losses which are correlated with healthcare spending shocks on certain treatment goods, then "over-insuring" (more generously covering them than would occur otherwise) those treatment goods is a second best response to reduce the individual's associated risk.

It is worth highlighting the limitations of our study. While we were able to generalize our optimal preventive care cost sharing results to an arbitrary demand and utility functions, our optimal treatment coverage results were all derived using a very specific demand structure. Both in our basic model and our utility-based framework, we consider only one period, one good demand functions that are linear in prices and have additive errors with constant variance.[12] We

[12] Such an extension would be useful for several reasons. First, it would allow us to generalize the findings of Goldman and Philipson (2007) to deal with complementarity and substitutability over time. Second, it would allow us to deal with issues of the insurer's shorter time horizon than the social planner's in a world where insurance contracts, especially through employment, may have a shorter duration and the insurer may be myopic or at least ignore some terms in the stream of future costs and benefits of healthcare treatment and prevention.

have ignored supply side incentives throughout, and hence our results assume that the level of healthcare treatment spending when uninsured is *ex post* efficient in the sense that the marginal cost is equated to the *ex post* marginal benefit.

We have also assumed that the demand for treatment is income inelastic, which is troubling to both us and no doubt many of the journal readers.[13] Our zero income assumption may be problematic. By making this choice, we assume away income distributional issues and corner solutions, which are particularly relevant in any equity discussions of optimal health insurance. In our model, subsidizing healthcare does not affect relative incomes, although it does affect those with poor health. We recognize that these are relatively restrictive assumptions, although our models remain more general than many others that have used only consumer surplus or assumed only two healthcare states or one healthcare good.

Our zero income elasticity assumption sidesteps the concerns of Nyman (1999) that there are income effects on behavior generated by insurance coverage. We are not very worried about this possibility for three reasons. The first is that the Health Insurance Experiment's findings on income effects suggest that they are quite modest once one adequately controls for the endogeneity of insurance and have reliable measures of health status. The second reason is that neither of us are convinced that the positive income effects of a few individual's with large healthcare expenditures are not offset by the cumulative smaller negative effects of the resulting premiums on a larger population. We are aware of no direct evidence on this compositional effect, given that almost none of the literature explicitly measures and estimates the two components. The third and more pragmatic reason for not being concerned about large income effects is we are not considering the effects of going from no insurance to optimal insurance, but rather the characteristics on the margin of an optimal insurance policy, where redistributive effects are smaller.

We have also repeatedly used a linear approximation of the marginal utility of income which is consistent with approximating the utility function with a constant absolute risk aversion function. We are not especially troubled by this assumption because our results should hold as an approximation for any arbitrary function, as long as the absolute risk aversion parameter is not varying too much across states of the world. Our uncompensated loss function and optimal savings function were also approximated using linear functions. Again, we believe that our results should hold as an approximation for more general nonlinear functions.

The other restrictive assumption that we have made for tractability sake is that the variance in healthcare expenditure is a constant, conditional on the health state. Specifically we have assumed that the variance (as well as the higher order moments) in healthcare treatment does not depend on either the level of preventive activity Z, nor the level of cost sharing, that is $\partial\sigma_\theta^2/\partial Z = 0$ and $\partial\sigma_\theta^2/\partial c_X = 0$. We do not need these restrictive assumptions to achieve our conclusions. Conditional on being sick, we might expect that more preventive activity Z also reduces the variance in expenditures conditional on health state. For example, earlier detection leads to earlier, less expensive and less variable expenditures. This would lead to a lower coinsurance rate for preventive activity, in part because there would be an additional element in the $\partial\pi/\partial Z$ term that acts as a wedge between the socially and privately optimal levels of prevention.

[13] The empirical literature finds that demand is income inelastic overall, especially in the absence of adverse selection on insurance coverage (Newhouse et al., 1993). But demand for specific healthcare treatment services may be more highly income elastic and yield different results.

Perhaps the area most in need of empirical work is to document the magnitude of uncompensated treatment and prevention costs to consumers. Significant uncompensated costs provide a rationale for zero or even negative cost shares on treatment goods and prevention goods, reflecting a second best correction in the absence of perfect insurance markets. It would be interesting to know how large are the adjustments needed to the conventional model results.

Acknowledgements

We would like to thank Shenyi Jiang for her assistance in the literature review, model development and checking of the derivations. We have also benefited from comments on earlier versions presented at Boston University, Harvard, and the iHEA World Congress in Copenhagen. We would also like to thank Kate Bundorf, Don Kenkel, Albert Ma, Tom McGuire, Joe Newhouse, Frank Sloan, and three anonymous reviewers for this journal for their useful comments on earlier versions of this paper. The opinions expressed are those of the authors, and not those of either Boston University or the University of Chicago.

Appendix A

This appendix derives selected analytical results in the main paper. For convenience, Table A.1 presents our notation. Equation numbers shown without an A prefix correspond to numbering in the main text.

A.1. Optimal cost sharing rates for healthcare treatment

We condition on a given level of preventive care *Z*, so that probabilities of different states of health are given. Treatment good *X* is assumed to have has linear demand curves of the form:

Table A.1
Notation

Z = quantity of preventive care
X = quantity of health care treatment
Y = quantity of all other consumption goods
I = consumers total income
J = disposable income after premiums and prevention spending
π = premium paid by consumer
P_X, P_Y, P_Z = demand prices of X, Y, and Z
c_X, c_Z = cost share rates on X, Z (share of insurable costs paid by consumer)
$\alpha(Z)$ = probability of being perfectly healthy
θ = random health shocks affecting the demand for X
μ_X = mean of health care spending on X with zero cost sharing X
B = slope of demand curves when written in the form $X=-Bc+\theta$
δ = insurance loading factor
σ_θ^2 = variance of health care shocks affecting spending X_i
V = the consumer's utility function
R^{A} = absolute risk aversion constant = $-V_{II}/V_I$
L_X = per unit uncompensated costs of treatment
L_Z = per unit uncompensated costs of prevention
$L_\theta(\theta)$ = uncompensated health losses that reduce utility but not income

$$X = A - B\frac{P_X}{P_Y} \tag{A.1}$$

This demand function is consistent with a risk neutral indirect utility function of

$$\tilde{V} = \frac{I}{P_Y} + \frac{BP_X^2}{2P_Y^2} - A\frac{P_X}{P_Y} \tag{A.2}$$

Using the normalizations $P_Y = 1$, $P_X = c_X + L_X$, $A = \mu_X + \theta$, and noting that $J = I - \pi$ will be available to spend on goods Y or X after paying the premium, applying the concave transformation $V(\ldots)$, and introducing L_X and L_θ as two forms of uncompensated treatment and health losses, this yields an indirect utility function that can be written as

$$\begin{aligned} &EV = \{E_\theta[V(J - K - (c_X + L_X)\theta) - L_\theta(\theta)]\} \\ &\text{where} \\ &J = I - \pi, \\ &K = (c_X + L_X)\mu_X - \frac{B(c_X + L_X)^2}{2}, \text{ and} \\ &\pi = (1 + \delta)[(1 - c_X)(\mu_X - B(c_X + L_X))] \end{aligned} \tag{A.3}$$

Taking partial derivatives with respect to c_X yields:

$$\frac{\partial EV}{\partial c_X} = [\{E_\theta[V_I(J - K - (c_X + L_X)\theta)]\}]\left[-\frac{\partial \pi}{\partial c_X} - \theta - \mu_X + B(c_X + L_X)\right] \tag{A.4}$$

Now using a Taylor series approximation of V_I around $J - K$ and rearranging we can write:

$$E_\theta\left\{\begin{array}{l} V_I(J - K)\left[\dfrac{\partial \pi}{\partial c_X} - \mu_X + B(c_X + L_X)\right] \\ -V_{II}(J - K)(c_X + L_X)\theta^2 - V_I(J - K)\theta \\ -V_{II}(J - K)(c_X + L_X)\theta\left[\dfrac{\partial \pi}{\partial c_X} - \mu_X + B(c_X + L_X)\right] \end{array}\right\} \approx 0 \tag{A.5}$$

Taking expectations and noting that the last two terms will have an expectation of zero:

$$V_I(J - K)\left(\frac{\partial \pi}{\partial c_X} - \mu_X + B(c_X + L_X)\right) - V_{II}(J - K)(c_X + L_X)\sigma_\theta^2 = 0 \tag{A.6}$$

Dividing through by $-V_I(J - K)$ and substituting $\partial\pi/\partial c_X = (1 + \delta)[-\mu_X - B + B(2c_X + L_X)]$ and $R^{\mathrm{A}} = -V_{II}/2V_I$:

$$(1 + \delta)(\mu_X + B - B(2c_X + L_X)) - \mu_X + B(c_X + L_X) - R^A(c_X + L_X)\sigma_\theta^2 = 0 \tag{A.7}$$

For the general case the optimal cost share c_X^* is (3)

$$c_X^* = \frac{B + (\mu_X + B)\delta - (B\delta + R^A\sigma_\theta^2)L_X}{B + 2B\delta + R^A\sigma_\theta^2}$$

This result is identical to the result for the basic model (3) in the main text, and hence the result can be seen as correct for a second approximation of any arbitrary utility function with zero income effects.

A.2. Choice of prevention and its optimal cost share

There are three classes of goods: healthcare treatment (X), prevention (Z), and all other goods and services (Y). There are two broad states of the world, healthy and sick; X is only demanded when sick. Using the normalization that the marginal insurable cost of Z is one, the cost to the consumer of preventive care is $P_Z = c_Z + L_Z$, and the indirect utility when healthy is

$$V^{\mathrm{H}} = V^{\mathrm{H}}\left(\frac{I - \pi - (c_Z + L_Z)Z}{P_Y}\right) \tag{A.8}$$

When sick, the linear demand curve for healthcare treatment X is

$$X = A - B\frac{P_X}{P_Y}, \tag{A.9}$$

with $A > 0$, and $B > 0$.

Using Roy's identity, and ignoring the possibility of corner solutions, the risk-neutral indirect utility function $\tilde{V}$ consistent with this demand function is

$$\tilde{V}^{\mathrm{S}}(I, P_X, P_Y) = \frac{I}{P_Y} + \frac{BP_X^2}{2P_Y^2} - A\frac{P_X}{P_Y} \tag{A.10}$$

Let $A = \mu_X + \theta$ where $\theta \sim F(\theta)$ with $E(\theta) = 0$. We apply a monotonically increasing concave function $V^{\mathrm{S}}(\ldots)$ to introduce risk aversion. Uncompensated costs $L_\theta(\theta)$ reduce utility without affecting the marginal utility of income and are assumed to be linear in θ. Using this notation, and normalizing $P_Y = 1$ and $\mathrm{MC_X} = 1$ so that $P_X =$ and $c_X + L_X$, the indirect utility function in the sick states will be (9)

$$V^S = V^S\left(I - \pi - (c_Z + L_Z)Z + -(c_X + L_X)\mu_X + \frac{B(c_X + L_X)^2}{2} - (c_X + L_X)\theta\right)$$
$$- L_\theta(\theta) \text{ for } (c_X + L_X)[\mu_X - B(c_X + L_X) + \theta] \le I - \pi - (c_Z + L_Z)Z.$$

The competitively determined insurance premium is (10)

$$\pi = (1 + \delta)[(1 - \alpha(Z))(1 - c_X)(\mu_X - B(c_X + L_X)) + (1 - c_Z)Z]$$

where δ is the administrative loading factor. The expected utility is then (11)

$$EV = \alpha(Z)V^{\mathrm{H}}(J) + (1 - \alpha(Z))\{E_\theta[V^{\mathrm{S}}(J - K - (c_X + L_X)\theta) - L_\theta(\theta)]\}$$

where

$$J = I - \pi - (c_Z + L_Z)Z,$$

$$K = (c_X + L_X)\mu_X - \frac{B(c_X + L_X)^2}{2}, \text{ and}$$

$$\pi = (1 + \delta)[(1 - \alpha(Z))(1 - c_X)(\mu_X - B(c_X + L_X)) + (1 - c_Z)Z]$$

The consumer's problem is to choose Z so as to maximize (11) while taking the premium π as given. Differentiating (11) with respect to Z while holding π constant characterizes the consumer's choice of Z:

$$\frac{\partial EV}{\partial Z} = \alpha'(Z)[V^{\mathrm{H}} - E_\theta[V^{\mathrm{S}}]] - (c_Z + L_Z)\{E_\theta[V_I]\} = 0 \tag{A.11}$$

This privately optimal choice of Z condition can be rearranged as (12)

$$\alpha'(Z) = \frac{c_Z + L_Z}{[V^{\mathrm{H}} - E_\theta[V^{\mathrm{S}}]]/E[V_I]}$$

The socially optimal choice of Z takes into account the effect of Z on the premium:

$$\frac{\partial EV}{\partial Z} = \alpha'(Z)[V^{\mathrm{H}} - E_\theta[V^{\mathrm{S}}]] - \left[\frac{\partial \pi}{\partial Z} + c_Z + L_Z\right]\{E_\theta[V_I]\} = 0 \tag{A.12}$$

Rearranging this result yields an expression characterizing the socially optimal choice of Z (13):

$$\alpha'(Z) = \frac{[\partial\pi/\partial Z + c_Z + L_Z]}{[V^H - E_\theta[V^{\mathrm{S}}]]/E_\theta[V_I]}$$

where $\partial\pi/\partial Z = -\alpha'(Z)(1+\delta)(1-c_X)(\mu_X - B(c_X + L_X)) + (1+\delta)(1-c_Z)$.

This result can be further rearranged to yield (14):

$$[Q + (1+\delta)(1-c_X)(\mu_X - B(c_X + L_X))]\alpha'(Z) = \ 1 + \delta(1-c_Z) + L_Z$$

If $c_Z = 1$ and $\alpha''(\cdot) < 0$, then $Z^{\mathrm{Soc}} > Z^{\mathrm{Priv}}$.

The optimal insurance rate for preventive care, c_Z^* can be found by solving (12) and (13) for c_Z^* when $Z^{\mathrm{priv}} = Z^{\mathrm{Soc}}$. This yields (15)

$$c_Z^* = \frac{Q - L_Z(1-c_X)(\mu_X - B(c_X + L_X))}{Q + (1-c_X)(\mu_X - B(c_X + L_X))} \quad \text{where } Q = \frac{[V^{\mathrm{H}} - E_\theta[V^{\mathrm{S}}]]}{E[V_I]}$$

This result can also be derived under more general assumptions about how prevention affects the probabilities of different states of the world. Assume there are N discrete states of the world, indexed by i, and let p_i be the probability that state i occurs when there is zero preventive care. Let θ_i be the health shocks occurring in state i. Let there be some proper subset of N called N^{H} (i.e., relatively healthy states) which is made more likely by preventive activity and N^{S} be the set of health states that are made less likely by prevention (relatively sick states). Define

$$p^{\mathrm{H}} = \sum_{i \in N^{\mathrm{H}}} p_i,\ p^{\mathrm{S}} = \sum_{i \in N^{\mathrm{S}}} p_i,\ P^{\sim\mathrm{H}\sim\mathrm{S}} \equiv 1 - p^{\mathrm{H}} - p^{\mathrm{S}}$$

The set J^{H} is preferred to the set J^{S} in the following sense:

$$EV^{\mathrm{H}} \equiv \left[\frac{\sum_{i \in N^{\mathrm{H}}} p_i V(I, c_X + L_X, \theta_i)}{p^{\mathrm{H}}}\right] > \mathrm{EV}^{\mathrm{S}} \equiv \left[\frac{\sum_{i \in N^{\mathrm{S}}} p_i V(I, c_X + L_X, \theta_i)}{p^{\mathrm{S}}}\right]$$

for all I and c_X. (A.13)

We will also need

$$EX^{\mathrm{H}} \equiv \left[\frac{\sum_{i \in N^{\mathrm{H}}} p_i X(I, c_X + L_X, \theta_i)}{p^{\mathrm{H}}} \right] \quad \text{and} \quad EX^{\mathrm{S}} \equiv \left[\frac{\sum_{i \in N^{\mathrm{S}}} p_i X(I, c_X + L_X, \theta_i)}{p^{\mathrm{S}}} \right] \tag{A.14}$$

Note that this notation does not require either that all states in N^{H} are preferred to all states in N^{S} or that all states in N^{S} are less preferred to those in N^{H}, only that on average N^{H} is preferred. Preventive care is assumed to increase the probability that the subset N^{H} (healthy states) occurs and decrease the probability that subset N^{S} (sick states) occurs in the following way. Let $\alpha(Z)$ be a probability multiplier for the increased probability that N^{H} occurs, with the properties $\alpha(0) = 1$, and $\alpha'(Z) > 0$ and $\alpha''(Z) < 0$ for all Z. Using this notation, the consumers expected utility can be written as follows:

$$\begin{aligned} EV = & \sum_{i \in N^{\mathrm{H}}} \alpha(Z) p_i [V(J, c_X + L_X, \theta_i) + L_\theta(\theta_i)] \\ & + \sum_{i \in N^{\mathrm{S}}} \left(\frac{1 - p^{\sim\mathrm{H}\sim\mathrm{S}} - \alpha(Z) p^{\mathrm{H}}}{p^{\mathrm{S}}} \right) p_i [V(J, c_X + L_X, \theta_i) + L_\theta(\theta_i)] \\ & + \sum_{i \in N^{\sim\mathrm{H}\sim\mathrm{S}}} p_i [V(J, c_X + L_X, \theta_i) + L_\theta(\theta_i)] \end{aligned} \tag{A.15}$$

where

$$J = I - \pi - (c_Z + L_Z)Z,$$

$$\pi = (1 + \delta) \left[\begin{array}{l} (1 - c_X) \sum_{i \in N^{\mathrm{H}}} \alpha(Z) p_i X(I, c_X + L_X, \theta_i) \\ (1 - c_X) \left(\frac{1 - p^{\sim\mathrm{H}\sim\mathrm{S}} - \alpha(Z) p^{\mathrm{H}}}{p^{\mathrm{S}}} \right) \sum_{i \in N^{\mathrm{S}}} p_i X(I, c_X + L_X, \theta_i) \\ (1 - c_X) \sum_{i \in N^{\sim\mathrm{H}\sim\mathrm{S}}} p_i X(I, c_X + L_X, \theta_i) + (1 - c_Z) Z \end{array} \right] \tag{A.16}$$

To find the maximum, we take partial derivatives with respect to Z:

$$\begin{aligned} \frac{\partial EV}{\partial Z} = & \, \alpha'(Z) \sum_{i \in N^{\mathrm{H}}} p_i [V(J, c_X + L_X, \theta_i) + L_\theta(\theta)] - \frac{\alpha'(Z) p^{\mathrm{H}}}{p^{\mathrm{S}}} \sum_{i \in N^{\mathrm{S}}} p_i [V(J, c_X + L_X, \theta_i) \\ & + L_\theta(\theta)] + \sum_{i \in N} p_i [V_I(J, c_X + L_X, \theta_i)] \left[-\frac{\partial \pi}{\partial Z} - (c_Z + L_Z) \right] = 0 \end{aligned} \tag{A.17}$$

which can be rewritten to characterize the social and privately optimal choices.

$$\alpha'(Z^{\mathrm{SOC}}) p^{\mathrm{H}} = \frac{\partial \pi / \partial Z + c_Z + L_Z}{[EV^{\mathrm{H}} - EV^{\mathrm{S}}]/EV_I} \quad \text{and} \quad \alpha'(Z^{\mathrm{PRIV}}) p^{\mathrm{H}} = \frac{c_Z + L_Z}{[EV^{\mathrm{H}} - EV^{\mathrm{S}}]/EV_I} \tag{A.18}$$

The private choice Z^{PRIV} ignores the effect of Z on the premium. These are identical to the equations in the main text for the more specific case where prevention only affects one state of the world.

Using

$$\frac{\partial \pi}{\partial Z} = (1+\delta)[(1-c_X)\alpha'(Z)p^{\mathrm{H}}[EX^{\mathrm{H}} - EX^{\mathrm{S}}] + (1-c_Z)]$$

the optimal insurance rate for preventive care, c_Z^* can be found by solving for c_Z^* when $Z^{\mathrm{priv}} = Z^{\mathrm{Soc}}$. This yields in the general case shown in the main text (16):

$$c_Z^* = \frac{Q - L_Z(1-c_X)(E(X^{\mathrm{S}}) - E(X^{\mathrm{H}}))}{Q + (1-c_X)(E(X^{S}) - E(X^{\mathrm{H}}))}$$

where $Q = [EV^{\mathrm{H}} - EV^{\mathrm{S}}]/EV_I$.

References

Arrow, K.J., 1963. Uncertainty and the welfare economics of medical care. American Economic Review 53, 941–973.

Arrow, K.J., 1971. Essays in the Theory of Risk-bearing. Markham Publishing Company, Chicago, IL.

Arrow, K.J., 1976. Welfare analysis of changes in health coinsurance rates. In: Rossett, R. (Ed.), The note of Health Insurance in the Health Services Sector. National Bureau of Economic Research, New York.

Barigozzi, F., December, 2004. Reimbursing preventive care. In: Geneva Papers on Risk and Insurance-theory, vol. 29, pp. 165–186.

Blomqvist, A., 1997. Optimal non-linear health insurance. Journal of Health Economics 16, 303–321.

Besley, T.J., 1988. Optimal reimbursement health insurance and the theory of Ramsey taxation. Journal of Health Economics 7, 321–336.

Buchanan, J.L., Keeler, E.B., Rolph, J.E., Holmer, M.B., 1991. Simulating health expenditures under alternative insurance plans. Management Science 37, 1067–1090.

Cutler, D.M., Zeckhauser, R.J., 2000. The anatomy of health insurance. In: Culyer, A., Newhouse, P. (Eds.), Handbook of Health Economics. North Holland.

Dardanoni, V., Wagstaff, A., 1990. Uncertainty and the demand for medical care. Journal of Health Economics 9, 23–38.

Eeckhoudt, Louis, Marchand, Maurice, Pestieau, Pierre, Piaser, Gwenael. Are differential co-payment rates appropriate in the health sector? October, 2004. CORE Discussion Paper No. 2004/70.

Ehrlich, I., Becker, G.S., 1972. Market insurance, self-insurance, and self-protection. Journal of Political Economics 80, 623–648.

Feldstein, M., Friedman, B., 1977. Tax subsidies, the rational demand for health insurance, and the health care crisis. Journal of Public Economics 7, 155–178.

Feldman, R., Manning, W.G., 1997. Une Formule Simple du Taux Optimale pour Une Police D-Assurance Maladie. In: S. Jacobzone (Ed.), Economie de la Sante: Trajectoires du Futur, Insee Methodes, number 64–65, pp. 133–144, French National Ministry of Health (English Translation: A Simple Formula for the Optimal Coinsurance Rate in a Health Insurance Policy).

Goldman, D., Philipson, T.J., 2007. Integrated insurance design in the presence of multiple medical technologies. American Economic Review 97 (2), 427–432.

Kenkel, D.S., 2000. Prevention. In: Culyer, A.J., Newhouse, J.P. (Eds.), Handbook of Health Economics 1A. North Holland, Amsterdam.

Manning, W.G., Marquis, M.S., 1996. Health insurance: the trade-off between risk pooling and moral hazard. Journal of Health Economics 15 (5), 609–639.

Newhouse, J.P., The Health Insurance Group, 1993. Free-For-All: Health Insurance, Medical Costs and Health Outcomes: The Results of the Health Insurance Experiment. Harvard University Press, Cambridge.

Newhouse, J.P., 2006. Reconsidering the moral hazard-risk avoidance tradeoff. Journal of Health Economics 25, 1005–1014.

Nyman, J.A., 1999. The economics of moral hazard revisited. Journal of Health Economics 18, 811–824.

Pauly, M.V., 1974. Overinsurance and public provision of insurance: the roles of moral hazard and adverse selection. Quarterly Journal of Economics 88, 44–62.

Pratt, J.W., 1962. Risk aversion in the small and the large. Econometrica 32, 122–136.
Zeckhauser, R., 1970. Medical insurance: a case study of the tradeoff between risk spreading and appropriate incentives. Journal of Economic Theory 2, 10–26.
Zweifel, P., Breyer, F., 1997. Health Economics. Oxford University Press, New York.
Zweifel, P.J., Manning, W.G., 2000. Consumer incentives in health care. In: Newhouse, J.P., Culyer, A.J. (Eds.), Handbook of Health Economics, 1A. Elsevier, Amsterdam, Chapter 8.