



International Association for the  
Study of Insurance Economics

# Études et Dossiers

---

Extract from

## Études et Dossiers No. 302

**World Risk and Insurance  
Economics Congress**

**Inaugural Conference**

7 – 11 August 2005  
Salt Lake City, Utah, USA

November 2005

**Working Paper Series of  
The Geneva Association**

© Association Internationale pour l'Etude de l'Economie de l'Assurance

The Geneva Association Working Paper Series “Études et Dossiers” appear at irregular intervals about 10 - 12 times per year. Distribution is limited.

The “Études et Dossiers” are the working paper series of The Geneva Association. These documents present intermediary or final results of conference proceedings, special reports and research done by The Geneva Association. As they contain work in progress or summaries of conference presentations, the material must not be cited without the express consent of the author in question.

Layout & Distribution: Valéria Kozakova

# **EGRIE Salt Lake City 2005 The transfer problem in copayment insurance**

John M. Marshall\*  
Department of Economics  
University of California  
Santa Barbara, CA 93106  
marshall@econ.ucsb.edu

## **Abstract**

The transfer problem in copayment insurance is the limited ability of price changes to transfer wealth into those health states in which the consumer has high marginal utility of wealth. The paper examines the transfer problem in the absence of, and in the presence of, the better-known substitution problem. Three contexts are considered: zero elasticity of substitution, unitary elasticity of substitution, and indivisible treatments. Adequacy of wealth transfers and health expenditures is measured by reference to the first-best optimum using complete markets in contingent wealth, and the copayment rate is always that of the second-best optimum. In every context the paper identifies inadequate expenditure on health care in case of grave illness and excessive spending in relatively healthy states.

---

\*Rod Garratt has amicably withdrawn as coauthor of this paper. I gratefully acknowledge his contributions.

## 1 Introduction

The transfer problem in copayment insurance is the limited ability of price changes to transfer wealth from states of health having low marginal utility of wealth to those having high marginal utility. The paper begins by studying wealth transfers in a context of zero substitution. Subsequently, it compares the substitution problem and the transfer problem. Then it examines the same issues in the context of indivisible treatments.

There is a belief that stems from Zeckhauser (1970) and is maintained in Cutler and Zeckhauser (2000) and many current textbooks, that in a copayment system, excessive expenditure on health care will occur in very state of health. The doctrine is important because it is the intellectual basis for a pervasive emphasis on cost containment in public policy towards health care. Despite the distinguished authority maintaining the excessive-expenditure position, today some authors have a more nuanced view. Pauly, whose 1968 note and diagram initiated study of the substitution problem, writes in 2000 that the situation is complex. Marshall (1976) and de Meza (1983) tangentially suggest the same thing, and Nyman in several recent papers finds that under spending is possible and even likely. The present paper shows again that the problem of spending too little is worthy of attention. Excessive expenditure in all states is a theoretical possibility, depending on specific preferences and endowments, but overspending is not universal in the models examined here. Instead, the typical pattern is inadequate spending in case of grave illness and excessive spending in relatively healthy states. Thus the analysis directs attention to grave illnesses, which in fact absorb the greater part of medical spending.

In examining these features, the first-best optimum is defined using complete markets in contingent claims, and the comparison is made to the optimum copayment model. The exercise is carried out in three models of consumer demand, each of which has its own section of the paper. The first model involves ex post preferences represented by Leontief utility, and the second uses Cobb-Douglas preferences. In each case, the ex ante demand for insurance is governed by logarithmic preferences. The distinction between ex post and ex ante preferences demonstrates the role of "intensity" of state demand. The third model of consumer preferences involves indivisible treatments. Again the ex ante and ex post demands are coordinated in a natural way. In that section it is shown almost incidentally that some medical treatments for which the cost exceeds the reservation price are efficient, contradicting the Arrow(1968, p. 538) remark that health insurance should allow only "...cost items that are regarded as 'normal'...where normality means roughly what would

have been bought in the absence of insurance."

The conclusions after completing this series of exercises are mainly negative. Copayment insurance has been misunderstood and misrepresented in too many instances. Its inefficiencies are not what they are often thought to be, but they are of such magnitude that it is unlikely that copayment systems were ever viable for health insurance. Present-day health plans operate far differently, and there is no evidence for the allegation that past systems were dominated by response of consumers to copayment incentives. In that sense, the paper encourages the research now being undertaken into the actions of physicians in recommending and ordering medical treatments.

## 2 Zero elasticity of substitution

Much literature concerns itself with substitution effects of the price subsidies and ignores the transfer problem. The clearest way to see the transfer problem is to turn the tables and ignore substitution. Thus it is useful to work in the context of elasticities of substitution equal to zero. The attendant L-shaped indifference curves remove the substitution effects. Many authors mistakenly write that zero demand elasticity is needed, but the crucial condition is zero elasticity of substitution, which coexists in the case examined here and in general with non-zero elasticity of demand.

### 2.1 Preliminaries

Health care and the numeraire consumption good are the two commodities in the model. They are produced under conditions of constant costs and scaled so one unit of either commodity has a production cost and market price of unity. Under copayment insurance the consumer can buy health care at a lower price,  $\theta$ . There are health states  $s = 0, 1, \dots, S$  with associated probabilities  $\pi_s$  and associated endowments of wealth  $\bar{w}_s$ . Consumption of the health care commodity is denoted  $h_s$  and consumption of ordinary goods is  $c_s$ . Utility  $u^s(c_s, h_s)$  carries a great deal of information because it represents ex ante preference for insurance and ex post preferences for consumption. A full analysis requires consideration of both aspects.

Ex post preferences are of the Leontief type<sup>1</sup>. In a state of health,  $s$ , a fraction  $\gamma_s$  of wealth is spent on health care, regardless of prices. The  $\gamma_s$  parameter will be referred to as the gravity of the health state. Thus

---

<sup>1</sup>Generalized Leontief would also satisfy the requirements but at the expense of two extra parameters to be studied.

ex post preferences are represented by the utility function

$$u^s(c_s, h_s) = \min\left(\frac{c_s}{1 - \gamma_s}, \frac{h_s}{\gamma_s}\right) \quad (1)$$

Without loss of generality the states  $s$  are numbered so that gravity  $\gamma_s$  is increasing in  $s$ .

The price of the consumption good is always unity, and the price of the health care good is  $\theta$ , a nonnegative number. Then the ex post demand functions are

$$c(w_s, 1, \theta; \gamma_s) = \frac{(1 - \gamma_s)w_s}{1 - \gamma_s(1 - \theta)} \quad (2)$$

$$h(w_s, 1, \theta; \gamma_s) = \frac{\gamma_s w_s}{1 - \gamma_s(1 - \theta)} \quad (3)$$

Also associated with the utility representation is the indirect utility function

$$\tilde{V}(w_s, 1, \theta; \gamma_s) = \frac{w_s}{1 - \gamma_s(1 - \theta)} \quad (4)$$

## 2.2 Intensity of state demand

The indirect utility function represents ex post behavior but not ex ante motives to insure. Strictly Leontief consumers have constant marginal utility of wealth and consequently do not demand anything recognizable as insurance. More information is needed in the form of a concave function  $g_s(\cdot)$  applied to the indirect utility, with the resulting Bernoullian utility function  $V$ , where

$$V(w_s, 1, \theta; \gamma_s) = g_s(\tilde{V}(w_s, 1, \theta; \gamma_s)) = g_s\left(\frac{w_s}{1 - \gamma_s(1 - \theta)}\right) \quad (5)$$

The embedding of ex post preferences in the ex ante preferences maintains a type of consistency. That is, in demanding insurance the individual correctly foresees what his ex post demands and satisfactions will be.

The examples considered here use a specific form for  $g_s(\cdot)$  in which there is an intensity parameter  $b_s \geq 0$  that indicates whether the disease depresses marginal utility of wealth or intensifies it. Specifically,

$$g_s(\tilde{V}) = b_s \ln(\tilde{V}) \quad (6)$$

$$V(w_s, 1, \theta; \gamma_s) = b_s \ln\left(\frac{w_s}{1 - \gamma_s(1 - \theta)}\right) \quad (7)$$

Intensity is important. In the first-best optimum the consumer chooses an optimum bundle of contingent wealths – contingent upon the health

state – and the prices of the contingent wealths are actuarially fair. The consumer’s optimized ex ante utility is, recalling that  $\theta$  is unity,

$$M(\bar{w}_0, \bar{w}_1, \dots, \bar{w}_S) = \text{maximum } \sum_0^S \pi_s b_s \ln w_s \quad (8)$$

$$\text{subject to } \sum_{s=0}^{s=S} \pi_s w_s = \sum_{s=0}^{s=S} \pi_s \bar{w}_s \quad (9)$$

The second order conditions are satisfied by concavity of  $\ln(\cdot)$ . Denote the solution as  $\{w_0^*, w_1^*, \dots, w_S^*\}$  and, for the multiplier,  $\lambda^*$ .

In this problem the intensity parameter  $b_s$  regulates the size of the marginal utility of wealth, which is

$$\frac{\partial}{\partial w_s} b_s \ln(w_s) = b_s \frac{1}{w_s} \quad (10)$$

The solution to the consumer maximization involves the conditions

$$\forall s \quad w_s^* = \frac{b_s}{\lambda^*} \quad (11)$$

It follows that

$$c_s^* = (1 - \gamma_s) w_s^* = (1 - \gamma_s) \frac{b_s}{\lambda^*} \quad (12)$$

Consider the whole vector of intensity parameters  $(b_0, b_1, \dots, b_S)$ . It is clear that the optimum consumptions of both goods are homogeneous of degree zero in them. Thus without loss of generality, take  $b_0 = \frac{1}{1-\gamma_0}$ . Now there is a recognizable neutral case in which  $b_s = \frac{1}{1-\gamma_s}$  for all states. It is appropriately called neutral because consumption of the ordinary good is constant across all states (from equation (12)).

More generally, then, some illnesses might be "depressive" in the sense that  $b_s < \frac{1}{1-\gamma_s}$ , leading to a reduction in optimum consumption of the ordinary good, relative to the neutral case. Other states could be "intensive" if  $b_s > \frac{1}{1-\gamma_s}$  meaning that consumption is larger, relative to the neutral case, in those ill states. A positive intensity parameter has the effect of raising ex ante demand for transfers of wealth into sick states.

Intensity is a reasonable consideration. Certainly there can be depressive states of health, hopeless afflictions that make survival a miserable exercise. Patients may refuse treatment ex post in such dire situations, and ex ante they might demand less insurance in them. However, depressive states are exceptional, even for grave illnesses. In thinking of serious health threats, the natural unit of analysis is not the individual but the family. The family increases consumption of the ordinary good as it reallocates resources to support the patient emotionally, to replace the patient’s contributions to household production, and to maintain

some ordinary activities that have become problematic because of the illness. The final year of a patient's life can be quite expensive for the family, even apart from the cost of medical care, and these expenses correspond to increased consumption of ordinary goods. Intense ex ante preferences represent that reality.

### 2.3 The transfer problem under copayment

The solution in the first best maximization is  $\{w_s^*\}_{s=0}^{s=S}$ . The attendant consumptions of both goods are points in the commodity space  $(c_s^*, h_s^*)$ . These first-best allocations become the targets towards which copayment insurance strives. The instruments for hitting the targets are an insurance premium  $P$  and a copayment rate  $\theta$ , which satisfy

$$P = (1 - \theta) \sum_{s=0}^{s=S} \pi_s h_s^* \quad (13)$$

Thus in each state the consumer pays a fraction  $\theta$  of the health care costs and the insurer pays the fraction  $1 - \theta$ . In the copayment regime, demand for  $h$  is found from equation (3) which here becomes

$$h(w_s, 1, \theta; \gamma_s) = \frac{\gamma_s w_s}{1 - \gamma_s(1 - \theta)} \quad (14)$$

In the first-best the demand  $h_s^*$  is  $\gamma_s w_s^*$ . All targets would be hit if for  $s = 0, 1, \dots, S$ ,

$$\frac{\gamma_s(\bar{w}_s - P)}{1 - \gamma_s(1 - \theta)} = \gamma_s w_s^* \quad (15)$$

The condition reduces to a linear expression in  $P$  and  $\theta$ , that is, for  $s = 0, 1, \dots, S$ ,

$$\gamma_s(\bar{w}_s - P) = (1 - \gamma_s(1 - \theta))\gamma_s w_s^* \quad (16)$$

a total of  $S + 1$  linear requirements plus the constraint connecting  $P$  and  $\theta$ .

One of the requirements is redundant. To see that, divide by  $\gamma_s$  on both sides in equation (16) and recognize that  $h_s^* = \gamma_s w_s^*$ . The result is

$$\bar{w}_s - P = w_s^* - (1 - \theta)h_s^* \quad (17)$$

Take expectations on both sides, noting that  $E[\bar{w}_s] = E[w_s^*]$ , and the result is the same as equation (13). Therefore a solution to the  $S + 1$  conditions of equation (16) automatically satisfies the restriction on  $P$  and  $\theta$ . Given two instruments and  $S + 1$  targets, it appears unlikely that all the targets can be hit. Success is possible, but it is improbable and non generic.

## 2.4 An exceptional success

In one special case the targets are all hit and the needed transfers are achieved. Suppose that illness has no consequences for wealth, that is for all  $s = 0, 1, \dots, S$ ,  $\bar{w}_s = \bar{w}$ , and that the intensity parameters are neutral,  $b_s = \frac{1}{1-\theta_s}$ . The constancy of wealth across states ignores the time cost of illness, especially serious illness, and the partial offsets of sick pay and disability insurance. Neutral intensity means that, as described above, optimum consumption is the same  $c^*$  in every state of health.

In this circumstance, the first-best optimum is attained by the copayment mechanism that takes  $\theta = 0$  and  $P = \bar{w} - c^*$ . Staying in the special context of state-independent wealth and neutral intensity, and setting aside the problem of substitutability, copayment insurance transfers wealth optimally into sick states. The result supplies a fresh rationale for the Arrow (1963) conclusions and clarifies their very special requirements. In that paper full insurance implied constancy of consumption across health states because illness did not affect utility directly but was instead a purely financial problem. In the present context that case is clearly special. It is sustainable only when intensity is neutral and wealth is independent of the health state.

## 2.5 Generic failure

There are numerous dimensions in which the results for the neutral-intensity, wealth-independent case are not generic. Consider first intensity of preference. Look at a two-state model and assume, quite reasonably, that the first best consumption in the grave state is above that in the routine illness state. Now the sick target is northeast of the well target, as shown in Figure 1.

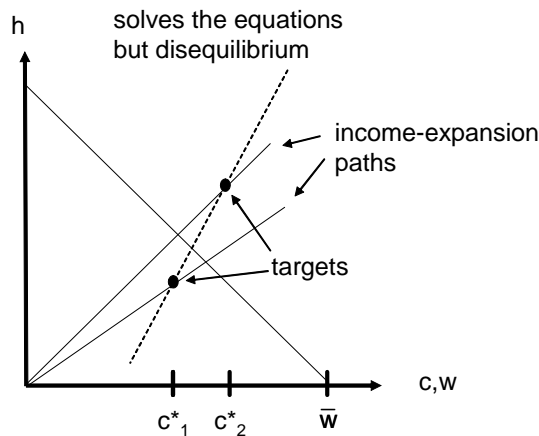


Figure 1: The transfer problem with intensity

A unique combination of  $P$  and  $\theta$  solves the equations (16), but the

solution is not a negatively-sloped budget constraint. It is a positively-sloped line – the dotted one – through the two targets. It has  $\theta < 0$  which would induce infinite demand for both goods. Thus the solution does not support the targets as an equilibrium, and because the conditions are linear, there is no second solution. The feasible second-best solution is a vertical budget constraint splitting the gap between the two points, as illustrated in Figure 2.

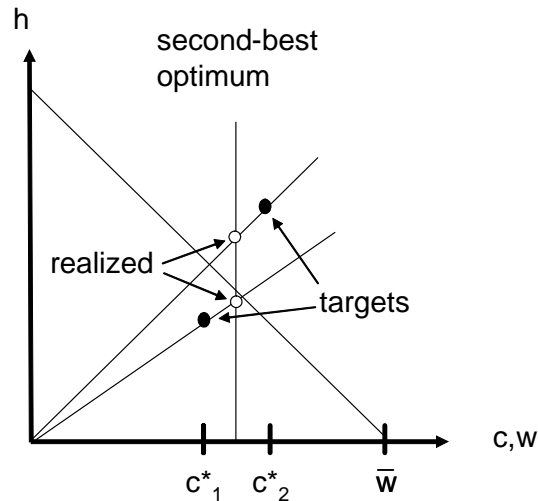


Figure 2: Second-best transfers

The second-best solution misses both targets by an optimum amount and supplies too much consumption and health care in the well state and too little of both in the ill state.

Thus, even when the targets and instruments are equally numerous, the second-best does not imitate the first-best. The nature of the failure is typical. Not enough wealth is transferred into the sick state. In consequence, over spending on health care occurs in the well state, but under spending prevails in the sick state. One expects more generally that when higher intensity is associated with graver illnesses, wealth transfers into the graver states will be inadequate, and wealth remaining in the less grave states will be excessive. The pattern is then under spending on health care in sick states and over sending in well states.

So much for lack of genericity with respect to preference intensity. Similar problems arise from variations in endowed wealth. For exposition, assume again that  $S = 1$  and intensity is neutral in all states, so that the first-best consumption is the same in both states. The targets lie on a vertical line. Consider then that wealth is less in the sick state. For any fixed premium  $P$  the budget line differs from one state to the next. The budget constraint in the sick state lies to the left of that for the well state. Again, the constraints are vertical, indicating  $\theta = 0$ , and

again there is more than optimum health care in the well state and less than the optimum in the sick state.

In summary, recall that the purpose here was to study copayment insurance in the absence of inefficiency from substitution effects. The findings are that inefficiency still prevails because of the inability of the two-parameter copayment system to match the multidimensional requirements of the first-best optimum. Over-spending prevails in some states, typically the well state and those with relatively minor illness – and possibly also a few depressive states – while under-spending is characteristic of states in which illness is more serious and, in consequence, intensity of preference or loss of wealth is a concern.

## 2.6 Implications of the transfer problem

The discussion above is somewhat informal. It is useful at this point to prove two propositions that tend to justify the diagrammatics. First, consider whether total spending on health care under copayment insurance can exceed total spending in the optimum.

**Proposition 1** *Health expenditure under the optimum copayment cannot exceed the first-best targets in all states. If the optimum is exceeded in any state, there is another that shows a short-fall.*

**Proof.** Suppose that the optimum copay implies higher expenditure on health care for every state. Because proportions are fixed, it also implies higher expenditure on the ordinary consumption good in every state. Thus the total of health care and ordinary consumption is higher than the optimum levels in every state. In consequence, the program is infeasible. The conclusion is that health spending is higher than optimum in some states and lower in others. ■

The next proposition deals with the tendency for the optimum copayment rate to be zero. For brevity of notation, let  $a_s = \frac{1}{1-\gamma_s}$ . Then  $b_s = a_s$  for all  $s$  is the baseline of neutral intensity. It is already known that optimum copayment rate is zero in the neutral case. The following proposition extends that to cases of increasing intensity.

**Proposition 2** *Suppose utility has the logarithmic form. Suppose that the intensity parameters satisfy  $b_0 = a_0$  and for  $s = 1, \dots, S$ ,*

$$\frac{b_s}{\pi_0 a_0 + \pi_1 b_1 + \dots + \pi_s b_s} \geq \frac{a_s}{\pi_0 a_0 + \pi_1 a_1 + \dots + \pi_s a_s} \quad (18)$$

*then the optimum copayment rate is zero.*

**Proof.** The idea of proof is to compare any second best optimum of the copay variety having  $\theta > 0$  with the  $\theta = 0$  case. Refer to the  $\theta > 0$  case as the challenger and the  $\theta = 0$  case as the incumbent. In the incumbent state the premium is  $P$  and in the challenger it is  $P^0$ . In the challenger state also define  $a'_s = \frac{1}{1-\gamma_s(1-\theta)}$ . Let any summation denoted by  $\Sigma$  without explicit limits be recognized as running over  $s = 0, 1, \dots, S$ . As before the  $\pi_s$ 's are probabilities of states.

Utility in the challenger is

$$U^c = \Sigma \pi_s b_s \ln(\bar{w} - P^0) a'_s \quad (19)$$

which becomes

$$U^c = \Sigma \pi_s b_s \ln(\bar{w} - P^0) + \Sigma \pi_s b_s \ln a'_s \quad (20)$$

The premium is found from

$$P^0 = (1 - \theta) \Sigma \pi_s \gamma_s a'_s (\bar{w} - P^0) \quad (21)$$

To simplify, observe that

$$(1 - \theta) \gamma_s a'_s = a'_s - 1 \quad (22)$$

from which

$$P^0 = [\Sigma \pi_s (a'_s - 1)] (\bar{w} - P^0) \quad (23)$$

Now write

$$\bar{w} - P^0 = \bar{w} - [\Sigma \pi_s (a'_s - 1)] (\bar{w} - P^0) \quad (24)$$

Solve for  $\bar{w} - P^0$  and take the logarithm solve to find

$$\ln(\bar{w} - P^0) = \ln \bar{w} - \ln(\Sigma \pi_s a'_s) \quad (25)$$

Now substitute in equation (20) with the result

$$U^c = \Sigma \pi_s b_s \ln(\bar{w}) - (\Sigma \pi_s b_s) \ln(\Sigma \pi_s a'_s) + \Sigma \pi_s b_s \ln a'_s \quad (26)$$

That completes characterization of the  $\theta > 0$  case:

The incumbent,  $\theta = 0$ , case is similar except that  $a_s$  replaces  $a'_s$  throughout. Thus

$$U^i = \Sigma \pi_s b_s \ln(\bar{w}) - (\Sigma \pi_s b_s) \ln(\Sigma \pi_s a_s) + \Sigma \pi_s b_s \ln a_s \quad (27)$$

Saving steps, form a measure of the difference between expected utilities in the two states:

$$\begin{aligned} \frac{U^i - U^c}{\Sigma \pi_s b_s} &= \ln(\Sigma \pi_s a'_s) - \ln(\Sigma \pi_s a_s) - \frac{\Sigma \pi_s b_s \ln \frac{a'_s}{a_s}}{\Sigma \pi_s b_s} \\ &= \ln \left( \frac{\Sigma \pi_s a_s \frac{a'_s}{a_s}}{\Sigma \pi_s a_s} \right) - \frac{\Sigma \pi_s b_s \ln \frac{a'_s}{a_s}}{\Sigma \pi_s b_s} \end{aligned} \quad (28)$$

Add and subtract cleverly the same thing and finally end with

$$\begin{aligned} \frac{U^i - U^c}{\Sigma \pi_s b_s} &= \ln(\Sigma \pi_s a'_s) - \ln(\Sigma \pi_s a_s) - \frac{\Sigma \pi_s b_s \ln \frac{a'_s}{a_s}}{\Sigma \pi_s b_s} \\ &= \ln \left( \frac{\Sigma \pi_s a_s \frac{a'_s}{a_s}}{\Sigma \pi_s a_s} \right) - \frac{\Sigma \pi_s a_s \ln \frac{a'_s}{a_s}}{\Sigma \pi_s a_s} + \frac{\Sigma \pi_s a_s \ln \frac{a'_s}{a_s}}{\Sigma \pi_s a_s} - \frac{\Sigma \pi_s b_s \ln \frac{a'_s}{a_s}}{\Sigma \pi_s b_s} \end{aligned} \quad (29)$$

In this equation, the first two terms on the right combine to yield a positive number because of the concavity of  $\ln(\cdot)$ .

Now examine the last two terms on the right. The method of proof is to show that their combined value is positive for every value of  $\theta$  under the assumption on intensity. To see that result, notice that each of the two right-most terms is a weighted sum of the values

$$\ln \frac{a'_0}{a_0} \geq \ln \frac{a'_1}{a_1} \geq \dots \geq \ln \frac{a'_S}{a_S} \quad (30)$$

where the inequality follows from the convention, assumed without loss of generality, that

$$\gamma_0 \leq \gamma_1 \leq \dots \leq \gamma_S \quad (31)$$

and from the negative association of  $\frac{a'_s}{a_s}$  and  $\gamma_s$ .

$$\begin{aligned} \frac{\partial}{\partial \gamma_s} \frac{a'_s}{a_s} &= \frac{\partial}{\partial \gamma_s} \frac{1 - \gamma_s}{1 - \gamma_s(1 - \theta)} \\ &= \frac{-\theta}{(1 - \gamma_s(1 - \theta))^2} \end{aligned} \quad (32)$$

The condition on intensity assures that the weighting that comes from the intensity parameters  $b_s$  puts more weight on the higher values of  $s$  than does the weighting from  $a_s$ . The proof of the condition is by induction. To begin, for  $s = 1$ , the condition reduces to  $b_1 \geq a_1$  from which it follows that

$$\begin{aligned} \frac{\pi_0 a_0}{\pi_0 a_0 + \pi_1 a_1} \ln \frac{a'_0}{a_0} + \frac{\pi_1 a_1}{\pi_0 a_0 + \pi_1 a_1} \ln \frac{a'_1}{a_1} \\ \geq \\ \frac{\pi_0 a_0}{\pi_0 a_0 + \pi_1 b_1} \ln \frac{a'_0}{a_0} + \frac{\pi_1 b_1}{\pi_0 a_0 + \pi_1 b_1} \ln \frac{a'_1}{a_1} \end{aligned} \quad (33)$$

Assume now the induction hypothesis for  $s = 0, 1, \dots, n$  and demonstrate it for  $n + 1$ . The induction hypothesis is that if, for all  $a_0, b_1, \dots, b_n$  satisfying, for  $s = 1, 2, \dots, n$ ,

$$\frac{\Sigma_{s=0}^{s=n} \pi_s a_s \ln \frac{a'_s}{a_s}}{\Sigma_{s=0}^{s=n} \pi_s a_s} \geq \frac{\Sigma_{s=0}^{s=n} \pi_s b_s \ln \frac{a'_s}{a_s}}{\Sigma_{s=0}^{s=n} \pi_s b_s} \quad (34)$$

Let  $b_{n+1}$  satisfy

$$\frac{b_{n+1}}{\sum_{s=0}^{s=n} \pi_s b_s + \pi_{n+1} b_{n+1}} \geq \frac{a_{n+1}}{\sum_{s=0}^{s=n} \pi_s a_s + \pi_{n+1} a_{n+1}} \quad (35)$$

It must be shown that

$$\begin{aligned} & \frac{\sum_{s=0}^{s=n} \pi_s a_s \ln \frac{a'_s}{a_s}}{\sum_{s=0}^{s=n} \pi_s a_s + \pi_{n+1} a_{n+1}} + \frac{\pi_{n+1} a_{n+1} \ln \left( \frac{a'_{n+1}}{a_{n+1}} \right)}{\sum_{s=0}^{s=n} \pi_s a_s + \pi_{n+1} a_{n+1}} \\ & \geq \\ & \frac{\sum_{s=0}^{s=n} \pi_s b_s \ln \frac{a'_s}{a_s}}{\sum_{s=0}^{s=n} \pi_s b_s + \pi_{n+1} b_{n+1}} + \frac{\pi_{n+1} b_{n+1} \ln \left( \frac{a'_{n+1}}{a_{n+1}} \right)}{\sum_{s=0}^{s=n} \pi_s b_s + \pi_{n+1} b_{n+1}} \end{aligned} \quad (36)$$

Rewrite the required relations as

$$\begin{aligned} & \frac{\sum_{s=0}^{s=n} \pi_s a_s \ln \frac{a'_s}{a_s}}{\sum_{s=0}^{s=n} \pi_s a_s} \frac{\sum_{s=0}^{s=n} \pi_s a_s}{\sum_{s=0}^{s=n} \pi_s a_s + \pi_{n+1} a_{n+1}} + \frac{\ln \left( \frac{a'_{n+1}}{a_{n+1}} \right)}{\pi_{n+1} a_{n+1}} \frac{\pi_{n+1} a_{n+1}}{\sum_{s=0}^{s=n} \pi_s a_s + \pi_{n+1} a_{n+1}} \\ & \geq \\ & \frac{\sum_{s=0}^{s=n} \pi_s b_s \ln \frac{a'_s}{a_s}}{\sum_{s=0}^{s=n} \pi_s b_s} \frac{\sum_{s=0}^{s=n} \pi_s b_s}{\sum_{s=0}^{s=n} \pi_s b_s + \pi_{n+1} b_{n+1}} + \frac{\ln \left( \frac{a'_{n+1}}{a_{n+1}} \right)}{\pi_{n+1} b_{n+1}} \frac{\pi_{n+1} b_{n+1}}{\sum_{s=0}^{s=n} \pi_s b_s + \pi_{n+1} b_{n+1}} \end{aligned} \quad (37)$$

Here there is a weighted sum on each side. On the right, the weight on the new term is greater than in the left, tending to the required result. Moreover, the new term is slightly less on the right. The first term on each side already has the required relation. Thus, on the left, each of the weighted terms is smaller than on the right, and moreover the weights favor the smaller term. Thus the induction step is complete, and so is the proof. ■

As is apparent from the proof, the sufficient condition on intensity is far from tight. The monotone increase in intensity is not necessary. Various intensities, even including some depressive states, would still be consistent with an optimum  $\theta$  of zero. For a check on the result, note that if all the inequality requirements on intensity are satisfied with equality the result is the same as that obviously true in the neutral-intensity state. It is unclear from the proof how to proceed if endowed wealth depends upon the state of health, at it reasonably should. That extension would be interesting, especially if it showed that  $\theta = 0$  is optimum in a wider class of cases.

## 2.7 Numerical examples with varying wealth and intensity

Consider the problems raised by differences among states in the amount of endowed wealth. It is reasonable that in case of serious illness, the patient and his family suffer a loss of wealth. Time lost from work, permanent disability and reduction of working life span, obviously diminish wealth. The family of the ill patient loses earnings and household services, perhaps partially offset by sick leave or disability insurance. A recent study has estimated that approximately half of all U.S. personal bankruptcies involve health care debts (Himmelstein et al. 2005) and projected that about two million persons per year would be affected in such bankruptcies. Clearly a theory of health insurance must take account of wealth variations. Diagrammatic arguments like those given above are intriguing, but perhaps the best procedure is to compute some cases that further illustrate the nature of the transfer problem.

The numerical example is based on the functional forms described above and on data about health expenditures collected by Berk and Monheit (1992) and reproduced in Cutler and Zeckhauser (2000).

Table 1. Distribution of Medical Spending in 1987

share of distribution	share of total spending
1%	30%
5%	58%
10%	72%
50%	98%
remainder	100%

The breakdown leads to the five-state model in Table 2 below. The probabilities match those in the above data (column one of Table 2) using endowed wealth (column two) and preference parameters (column three) that lead to expenditures (column four) that match the Berk and Monheit shares (column five). In all examples the expectation of endowed wealth is unity, and the share of medical spending is 10% of wealth. Then the share by state of health care in total wealth should be as shown in the last column. To match that column, I computed the shares under optimum copayment and adjusted the health care demand parameters for a close fit. The resulting parameters are in the third

column and the spending is shown in column four.

Table 2: Calibration of the neutral model

Probability of state	Endowed wealth	Health care parameter $\gamma_s$	Spending out of $E[\bar{w}] = 1$	Berk & Monheit share at 10%
.01	.8	0.81	0.029823296	0.03
.04	.9	0.465	0.02779854	0.028
.05	1.006315	0.24	0.014303696	0.014
.4	1.006315	0.068	0.026438369	0.026
.5	1.006315	0.0045	0.002047497	0.002
expectation	1		0.100411397	0.1

Searching for exact matches was pointless because the data are already rounded. Numerous insignificant decimals from my computations are included in the tables to facilitate the work of those who want to repeat and improve the exercise.

The significant thing is to compare the optimum copayment with the first-best optimum, looking particularly at the wealth transfers, upon which everything else depends.

Table 3: Adequacy and Inadequacy in the second-best optimum

gravity	First best state wealth	copayment state wealth	First best health care	copayment health care
0.81	4.679888382	3.681888334	3.79070959	2.982329551
0.465	1.662016435	1.494545154	0.772837642	0.694963496
0.24	1.169972096	1.191974666	0.280793303	0.28607392
0.068	0.954054499	0.971998869	0.064875706	0.066095923
0.0045	0.893198184	0.909998559	0.004019392	0.004094994
Expectation	0.999999038	0.99999825	.110820245	.100411

Even in the absence of substitution problems, the wealth transfers of the copayment system are wrong. Grave states get too little wealth, routine states get too much.

So far there are no intensity differences. They make the transfer problem worse without changing its qualitative features. Intensities are chosen arbitrarily in the following table. Given those intensities, the other preference parameters are readjusted to match the Berk and Mon-

heit data, with results given in Table 4.

Table 4: Calibration of the intensity model

Prob.	Endowed wealth	Gravity $\gamma_s$	Intensity $b_s$	Health spending out of $\bar{w}$	Berk-Monheit share at 10%
.01	.8	0.81	1.8	0.029823296	0.03
.04	.9	0.47	1.5	0.028362514	0.028
.05	1.006315	0.24	1.3	0.014303696	0.014
.4	1.006315	0.067	1	0.02602165	0.026
.5	1.006315	0.0045	1	0.002047497	0.002
exp.	1			0.100558652	0.1

The gravity parameter in the sickest states must be very high to make spending match the data. That happens for a by-now-familiar reason: wealth in the sickest states is below the optimum. As a standard for calibrating health spending, the choice is between the first-best optimum and the optimum copayment. Neither is totally compelling. I used the optimum copayment.

Results follow the same qualitative pattern as before: not enough wealth, consumption, or health care in the gravest states, and too much of everything in the relatively well states.

Table 5: Adequacy in the intensity model

state	1st best	copayment	1st best	copayment
gravity	state wealth	state wealth	health care	health care
0.81	7.737051113	3.681888334	6.267011401	2.982329551
0.47	2.311383194	1.50864437	1.086350101	0.709062854
0.24	1.396967562	1.191974666	0.335272215	0.28607392
0.067	0.875336258	0.970957081	0.058647529	0.065054124
0.0045	0.820380441	0.909998559	0.003691712	0.004094994
Expectation	0.99999894	1.0001455	0.148192597	0.100558652

It may help to see this data in a graph. For clarity of visualization, each value, for instance,  $h$ , is graphed as  $\ln(1 + h)$ .

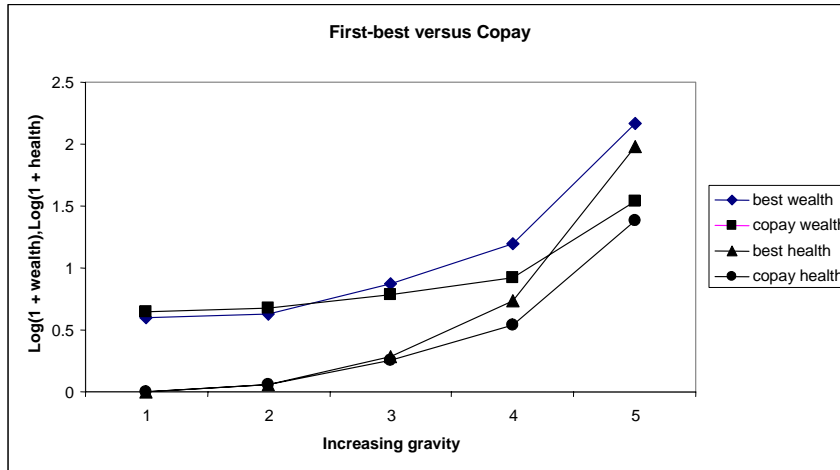


Figure 3: Optimum versus copay

The crossings in the graph between states two and three represent the observation that copayment wealth and health care are excessive in the routine states and inadequate in the gravest ones. The other crossing, between states four and five, shows that optimum health spending in the gravest state is greater than the whole wealth available in that state under the copayment system.

A remarkable feature of this example is the relation of total health spending across the systems

Table 6: Expected health care expenditure

	Expected health care
Endowed	0.064709231
First best	0.148192597
copayment best	0.100558652

The fully efficient transfer of wealth leads to more than twice as much health care in the first best, of which a third is sacrificed by using a copayment system instead.

These explorations of the transfer problem are a thorough indictment of copayment systems as vehicles for insurance. Each failure asks for an obvious modification. A good system would, for instance, separate the gravest states and transfer wealth into them by some other mechanism, as actual health insurance systems do and have always done. These transfers would account for a large fraction of total expenditures, somewhere between thirty and sixty percent, based on the data of Berk and Monheit. Whether the remaining copayment insurance would be worth having is doubtful. To explore that idea, I assumed that the two gravest states – the worst 5% – were handled by an optimum lump-sum, and I

optimized the copayment rate for the three least-ill states. The result was an optimum copayment rate of 97%, which is almost no insurance at all. These considerations leave the question: what real phenomenon is modeled by the copayment model? Quite possibly none at all.

### 3 Unitary elasticity of substitution:

The Leontief case has elasticity of substitution equal to zero. This section compares the Leontief case to the Cobb-Douglas case of unitary elasticity of substitution. The development is parallel to that of the previous section. As before the health states are  $s = 0, 1, \dots, S$  with associated probabilities  $\pi_s$  and associated endowments of wealth  $\bar{w}_s$ . The health care commodity is  $h_s$  and consumption of ordinary goods is  $c_s$ . In a state of health,  $s$ , a fraction  $\gamma_s$  of wealth is spent on health care at equal prices. Again  $\gamma_s$  grows with  $s$  without loss of generality. Ex post preferences are represented by

$$u^s(c_s, h_s) = c_s^{1-\gamma_s} h_s^{\gamma_s} \quad (38)$$

The price of the health care good is  $\theta$ , a nonnegative number and the ex post demand functions are

$$c(w_s, 1, \theta; \gamma_s) = (1 - \gamma_s)w_s \quad (39)$$

$$h(w_s, 1, \theta; \gamma_s) = \frac{\gamma_s w_s}{\theta} \quad (40)$$

and the resulting indirect utility function is

$$\tilde{V}(w_s, 1, \theta; \gamma_s) = \frac{w_s}{\theta^{\gamma_s}} (1 - \gamma_s)^{1-\gamma_s} \gamma_s^{\gamma_s} \quad (41)$$

and for convenience denote the constant term  $(1 - \gamma_s)^{1-\gamma_s} \gamma_s^{\gamma_s}$  by  $K_s$ .

#### 3.1 Intensity of state demand

The Bernoullian utility function  $V$  comes from the indirect utility function and a concave function  $g_s(\cdot)$  applied to the indirect utility, with the result

$$V(w_s, 1, \theta; \gamma_s) = g_s(\tilde{V}(w_s, 1, \theta; \gamma_s)) = g_s\left(\frac{w_s}{\theta} K_s\right) \quad (42)$$

The examples use a specific form for  $g_s(\cdot)$ , namely

$$g_s(\tilde{V}) = b_s \ln(\tilde{V}) = b_s \ln\left(\frac{w_s}{\theta} K_s\right) \quad (43)$$

As before, the intensity parameter  $b_s$  indicates whether the disease depresses consumption or magnifies it. When  $\theta = 1$  the optimizing condition is

$$\forall s \quad w_s^* = \frac{b_s}{\lambda^*} \quad (44)$$

from which it follows that

$$c_s^* = (1 - \gamma_s)w_s^* = (1 - \gamma_s)\frac{b_s}{\lambda^*} \quad (45)$$

Again there is a recognizable neutral case in which  $b_s = \frac{1}{1-\gamma_s}$  for all states, and that is the one examined next and compared to the previous, zero elasticity of substitution case. Thus the consumer's state utility function is

$$V(w_s, 1, \theta; \gamma_s) = \frac{1}{1 - \gamma_s} \ln\left(\frac{w_s}{\theta^{\gamma_s}} K_s\right) \quad (46)$$

Obviously this is similar to the Leontief preference model. In fact, ex post demands of Leontief and Cobb-Douglas preferences coincide exactly along the income expansion paths, which are rays from the origin with slope  $\frac{\gamma_s}{1-\gamma_s}$ . Utilities along the rays differ by a factor  $K_s$ , but because of the logarithm, there is no difference in marginal utilities.

The consumer's optimization is

$$\begin{aligned} M(\bar{w}_0, \bar{w}_1, \dots, \bar{w}_S) = & \text{maximum } \sum_0^S \pi_s b_s \ln\left(\frac{w_s}{\theta} K_s\right) \\ \text{subject to } & \sum_{s=0}^S \pi_s w_s = \sum_{s=0}^S \pi_s \bar{w}_s \end{aligned} \quad (47)$$

Again, the numerical example is based on data about health expenditures cited above. The state probabilities are familiar. Health care parameters are shown in the second column of Table 7. Using those parameters and computing demands at the optimum copayment rate (which turned out to be 0.27130), the third column shows spending in the states as a fraction of total wealth  $\bar{w}$ . The last column gives the contribution by state implied by the Berk and Monheit data.

Table 7. Calibration under Unitary Elasticity of Substitution

Probability of state	Health care parameter $\gamma_s$	Spending out of $\bar{w}$	Berk & Monheit share at 10%
.01	.9	0.030700107	0.03
.04	.23	0.028653434	0.028
.05	.08	0.013644492	0.014
.4	.02	0.027288984	0.026
.5	.0012	0.002046674	0.002
Expectation		0.103078976	0.1

The very large expenditures in the sickest state are difficult to achieve in the Cobb-Douglas model, and all the calibration is sensitive to the gravity  $\gamma_s$  in that state. Undue sensitivity is another suggestion that a copayment system is inappropriate for any environment in which such

large expenditures are likely, that is, it is inappropriate for health insurance in general.

Returning to the example, the same parameters are used as in the Leontief version. By construction, both versions have the same first-best optimum spending on health care, shown in column two of Table 8. They also have the same first-best optimum state wealth, as shown in column four. Moreover, consumers under the Leontief model receive the first-best optimum quantities in the optimum copayment system, which has a copayment rate of zero. The situation is different in the Cobb-Douglas model. Health care is in the third column and the corresponding state wealth is in the last column.

Table 8: Adequacy under Unitary Elasticity

gravity $\gamma_s$	first-best health care	copayment health care	first best state wealth	copayment state wealth
0.9	8.080847422	3.070010734	8.978719358	3.162553726
0.21	0.238674818	0.716335838	1.136546754	1.447425468
0.08	0.078075821	0.272889843	0.975947756	1.124285362
0.02	0.018323917	0.068222461	0.916195853	0.975143774
0.0012	0.001078741	0.004093348	0.898950677	0.928412744
Expected	0.102128195	0.102333691		copayment =.27130

From Table 8 it is clear that state wealth and health care are excessive under the optimum copayment system in all but the sickest state. Consumption of ordinary goods is also excessive in some of those states, a consequence of inappropriately high wealth. Health care, state wealth, and consumption are inadequate in the sickest state. Total expected expenditure on health care is nearly the same in both systems but slightly more under the optimum copayment. Undoubtedly, minor modifications could reverse that comparison.

The frequent misconception is that copayment insurance increases health care at the expense of ordinary consumption. That is not true when the copayment outcomes are examined state by state. States showing excessive health care can show excessive consumption because they are states with excessive wealth, or consumption can be lower and health-care higher than efficient, as happens twice in the present example. There is no consistent rule.

From experimentation with these models I can report that unequal endowed wealth in states often leads to optimum copayment rates greater than unity. That is, the consumers would like to anti-insure their health risk by paying extra for health care in exchange for a premium they receive with certainty. That is beneficial in the theory because it raises

their wealth in states with the low endowed wealth. It seems to have no real-world counterpart. The result suggests, again, the inappropriateness of copayment systems to deal with major health risks.

Pauly (1968) and Zeckhauser (1970) showed that copayment systems are problematical. It was tempting to read that as showing that the private sector cannot insure health care efficiently and that government health insurance in the form of Medicare and Medicaid would also be inefficient. What it really showed was that the private insurers could not survive using a copayment system and that the government, too, would have to seek some other model for transferring wealth into sick states. Private sector health insurance at the time employed a variety of coverages, including a major medical coverage that dealt with serious illness. It is wrong to think that health care under copayment insurance is universally excessive, but it is right to say that copayment systems are inefficient. The price subsidies are weak tools for wealth transfer.

### **3.2 Constrained optimum**

Because the results undermine a consensus, some further discussion is needed. One might object to the concept of optimum used. In particular, the optimum allocates wealth to states for the purpose of optimizing consumption as well as health expenditures. In that way it insures not only health care but also lost wages, reduced productivity, and disability. The optimum assumes the use of an instrument, disability insurance, that is not in the possession of health insurers. This is an argument that should be considered seriously.

After due consideration, however, there are three grounds for rejecting the argument. The first is that a social optimum should, by its nature, comprehend all possibilities including coordination of disability insurance with health insurance. Otherwise what is the meaning of optimum? The second ground for rejection is that the concept of contingent claims optimum was originated by Arrow (1954) and first applied to health insurance by Zeckhauser (1970), and in the intervening years no other concept of optimum has been offered in the health insurance literature.

The third ground is more complicated and more compelling. In the general economics literature there is a concept of constrained optimum that might be applied here to study a health insurance that has been stripped of the disability element. The optimum is then constrained by the same thing that limits the health insurer, namely the lack of a franchise to coordinate with disability insurance. Let us see where this argument takes us. It will not take us far.

The second-best optimum is like the first best (equation 47) except

for the addition of constraints of the form

$$c_s \leq \bar{w}_s \text{ for } s = 0, 1, 2, \dots, S \quad (48)$$

The new constraints say that consumption of the ordinary good cannot exceed the endowed wealth in the state. For analysis, think of the constraints as being introduced one at a time. Thus when only the constraint in state zero is added the solution is  $c_s^{[0]}$ ,  $h_s^{[0]}$ , and  $\lambda_s^{[0]}$ . When constraints in states zero and one are present, the optimum is  $c_s^{[1]}$ ,  $h_s^{[1]}$ , and  $\lambda_s^{[1]}$ . Continue in this fashion until all constraints are present, in which case the optimum is  $c_s^{[S]}$ ,  $h_s^{[S]}$ , and  $\lambda_s^{[S]}$ . The latter is the second-best, constrained optimum.

The major result is that, for  $s = 0, 1, 2, \dots, S$ ,

$$h_s^* \leq h_s^{[0]} \leq h_s^{[1]} \leq \dots \leq h_s^{[S-1]} \leq h_s^{[S]} \quad (49)$$

In words, optimum health care in each state in the constrained optimum is, if anything, greater than in the unconstrained optimum. Contrary to first impressions that lack of disability insurance reduces optimum health expenditures, it can only increase them.

To establish the result, consider in detail the addition of the first constraint,  $c_0 \leq w_0$ . Let the multiplier associated with the constraint be  $\psi_0$ . First order conditions for equation (47) are modified slightly to become

$$\begin{aligned} u_1^0(c_0^{[0]}, h_0^{[0]}) - \lambda^{[0]} - \psi_0 &= 0 & (50) \\ u_2^0(c_0^{[0]}, h_0^{[0]}) - \lambda^{[0]} &= 0 \\ \text{and for } s = 1, 2, \dots, S & \\ u_1^s(c_s^{[0]}, h_s^{[0]}) - \lambda^{[0]} &= 0 \\ u_2^s(c_s^{[0]}, h_s^{[0]}) - \lambda^{[0]} &= 0 \\ \sum_{s=0}^S \pi_s \bar{w}_s - \sum_{s=0}^S \pi_s (c_s^{[0]} + h_s^{[0]}) &= 0 \end{aligned}$$

In case the constraint is not binding,  $\psi_0 = 0$ ,  $\lambda^{[0]} = \lambda^*$ ,  $c_s^{[0]} = c_s^*$ , and  $h_s^{[0]} = h_s^*$ . In the interesting case, the constraint is binding,  $\psi_0 > 0$ , and  $c_0^{[0]} = w_0 < c_0^*$ . The resources freed by the binding constraint amount to  $\pi_0(c_s^* - c_s^{[0]})$  and are available to support health expenditure in all states and consumption in states  $s = 1, 2, \dots, S$ . Distributing the increment optimally requires equating marginal utilities in all states, and because the utility functions are concave,  $\lambda^{[0]} < \lambda^*$ . Marginal utility in all the other states is less than before. Recalling that both goods are normal at unitary prices, it follows that in every state

$$h_s^{[0]} > h_s^* \quad (51)$$

and in all states except the first

$$c_s^{[0]} > c_s^* \tag{52}$$

However, the important thing is that, in either case,

$$h_s^{[0]} \geq h_s^* \tag{53}$$

The lack of disability insurance has, if anything, increased the optimum health expenditures in every state. The result is intuitive. A binding constraint releases resources to be used in other states and therefore, in every state, increases consumption of the health-care good.

When the second, third, and so on constraints are added to the problem, the same analysis applies. Thus is proved the result in equation (49).

The analysis pursued in every section of this paper directs attention to the gravest states and shows that copayment systems tend to supply too little wealth and too little health care in them. Using the constrained optimum does not ameliorate the problem but makes it worse. The main findings of the paper survive limitations on disability insurance or other direct wealth transfers into the sickest states.

### 3.3 Summary so far

Copayment insurance is a weak tool for insurance even in the fixed-coefficients, Leontief case. Adding a non zero elasticity of substitution makes it totally ineffective. The way it fails, however, is quite different from what is imagined in much of the literature. There is no tendency for universal overspending on health care. Instead, typical health care spending is excessive in states of routine illness and inadequate in the gravest states. States in the middle have various attributes, some even showing excessive consumption of both goods. In short, copayment systems don't necessarily result in too much health care, but they do cause health care and ordinary consumption to be inappropriately distributed over health states.

## 4 Indivisible treatments

In Nyman's writings about inadequate spending on health care under copayment insurance, he frequently cites grave illnesses and expensive treatments. In many of these examples an element of indivisibility is explicitly recognized. The effect of indivisibility is to make substitution inelastic in possibly relevant ranges. The thought that indivisibility partly remedies the faults of copayment insurance is an attractive one and needs more attention.

Given that treatments are indivisible, there might be several possible treatments in each sick state, but addressing the problem of treatment selection would require analysis that does not advance current goals. Assume therefore that the unique available treatment in any sick state is known. The question of over or under spending then boils down to whether the right treatments are accepted or not. Over spending occurs when the treatment is not in the optimum but is provided to the consumer. Under spending means an efficient treatment – a treatment in the optimum – is omitted. Given this background, the preferences of the consumer in the sick state  $s$  are shown in Figure 4.

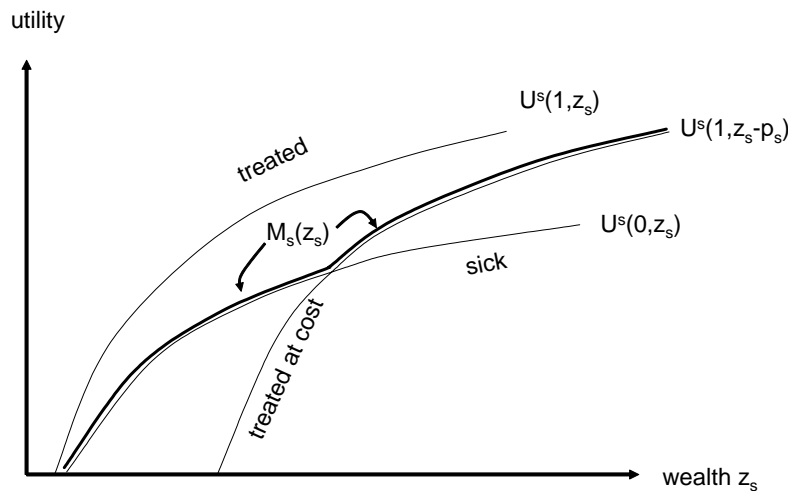


Figure 4. Utility of a sick person choosing to undergo or avoid an indivisible treatment

The horizontal axis measures consumption of an ordinary, divisible, numeraire good that represents all non health commodities. The “sick” utility function is the one that holds in the absence of a restorative medical treatment. The “treated” utility function applies when the patient receives the only available treatment, which is indivisible. For the moment let the horizontal axis measure *consumption* of the numeraire in state  $s$ . The utilities are denoted  $U^s(0, z_s)$  and  $U^s(1, z_s)$ , corresponding to sick with no treatment, and sick with treatment. For given health status and treatment choice, the consumer is always risk averse in consumption of the ordinary good – that is  $U_{22}^s(0, z_s) < 0$  and  $U_{22}^s(1, z_s) < 0$ .

Now change viewpoints and suppose that  $z_s$  on the horizontal axis represents the wealth available in state  $s$ . The utility of a person who chooses not to be treated is still  $U^s(0, z_s)$  because all wealth is spent on the consumption good. In case he chooses the treatment and pays the cost  $p_s$ , consumption of the ordinary good is  $z_s - p_s$ . Utility of wealth  $z_s$  is  $U^s(1, z_s - p_s)$ , which is obtained by shifting the curve  $U^s(1, z_s)$

to the right by a distance  $p_s$ , as is done in Figure 4, resulting in the treated-at-cost curve.

Thus in each state the indirect utility function is the upper envelope of  $U^s(0, z_s)$  and  $U^s(1, z_s - p_s)$ . It is illustrated in Figure 4 by the heavier, non-concave curve

$$M_s(z_s) = \max[U^s(0, z_s), U^s(1, z_s - p_s)] \quad (54)$$

Demand for treatment is read from the diagram. At low wealth the utility of being treated at cost lies below the untreated curve  $U^s(0, z_s)$ , and the consumer prefers not to be treated. At a critical wealth the consumer becomes indifferent, and for wealth above that level he prefers to have the treatment in spite of its cost. The indirect utility function in Figure 4 is of the risk-loving type identified by Friedman and Savage (1948). Consumers might be expected to seek gambles in wealth prior to deciding on treatment.<sup>2</sup> In the present application, however, gambling is of negligible importance, as further analysis will show. There are of course other sick states and possibly some healthy ones. Each is associated with its own state-dependent utility functions and  $M_s(z_s)$  curve.

Values of treatments in different sick states are commensurate through a measure of efficacy, which is derived as follows: Define  $\widehat{M}_s(z_s)$  to be the concave hull of  $M_s(z_s)$ , that is, the least concave function that is nowhere less than  $M_s(z_s)$ . It is illustrated in Figure 5.

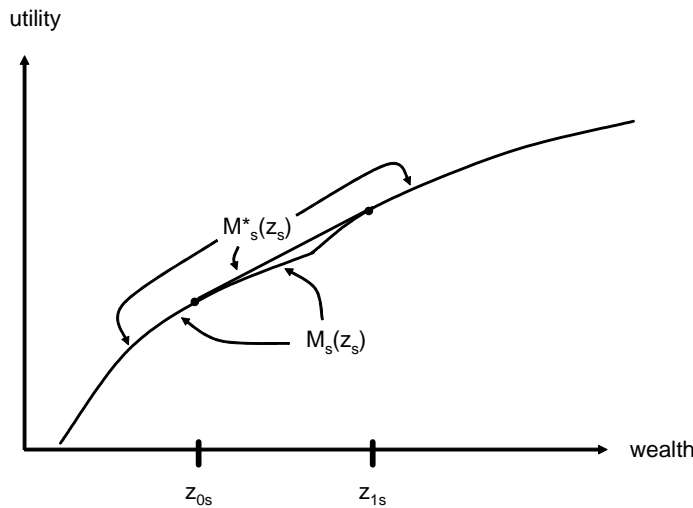


Figure 5. Surrogate utility

Now define the efficacy of the treatment as the slope of the linear portion

<sup>2</sup>The role for randomization arises in a similar fashion in Marshall (1984), Bergstrom (1986) and Garratt and Marshall (1994,1995).

of  $M_s^*(z_s)$ . Clearly efficacy has the units of marginal utility. Its role in optimum health insurance is described below.

States belong to a finite set  $S$ . The probability of a state is  $\pi_s$ . Since contingent wealths are fairly priced,  $\pi_s$  is also the price of a unit of wealth contingent upon its state.

The consumer's endowed wealth in state  $s$ , before entering the contract, is a given quantity  $\bar{w}_s$ . As previously argued, a theory of health insurance should take account of wealth variations.

### 4.1 Optimum

Insurance consists of trading the contingent endowed wealths  $\bar{w}_s$  for contingent insured wealths  $z_s$ . The goal is to optimize the expectation of the indirect utility functions  $M_s(z_s)$ , but these functions are inconvenient because they are not concave. Therefore substitute the concave hulls  $\widehat{M}_s(z_s)$  at this point and notice later on that the solutions using them are actually solutions for the non concave utilities as well. Therefore

$$\text{Maximize}_{\{z_s\}} \sum_{s \in S} \pi_s \widehat{M}_s(z_s) \tag{55}$$

subject to the budget constraint

$$\sum_{s \in S} \pi_s z_s = \sum_{s \in S} \pi_s \bar{w}_s \tag{56}$$

The problem in equations (2) and (3) is a concave programming problem. The solution  $\{z_s^*\}_{s=1}^S$  involves a Lagrangian multiplier  $m^*$  satisfying, for all states  $s$

$$\widehat{M}_s'(z_s) = m^* \tag{57}$$

The condition means that marginal utility is equalized across states. It is a familiar condition in the demand for fairly-priced insurance and is the result of the consumer trading wealth out of states that have low marginal utility and into those having high marginal utility until no further trades of that kind can be made.

Possibly in some states  $z_s^*$  falls in the linear portion of  $\widehat{M}_s(z_s)$ . In any such state the solution requires that the consumer is given a fair gamble between the wealths defining the end points of the linear portions of  $\widehat{M}_s(z_s)$ , the points  $z_{0s}$  and  $z_{1s}$  in Figure 5. However, the optimum expected utility can always be reached with an allocation requiring a gamble in no more than one state<sup>3</sup>. A single state is negligible in a state space that consists of all possible diagnoses, but notice that in the

---

<sup>3</sup>Assume there is more than one  $z_s^*$  in the interior of a flat spot. Move wealth downward in one such state and upward by a compensating (at fair price) amount

numerical examples below which use small numbers of states, gambling solutions are needed to find the optima.

The solution just described is also the solution to the non-concave problem in which the maximization at equation (55) has the "real" utility functions  $M_s(z_s)$  in place of the surrogates  $\widehat{M}_s(z_s)$ . This is so because the maximand in the real problem already equals the optimum of the surrogate problem and cannot in any case exceed it. That completes the definition of optimality.

In order to compare copayment plans with the full optimum, it is convenient to restrict the form of the utility function. From now on suppose that in each sick state there is a quantity  $B_s$  which is the utility benefit from treatment. Utility in the state satisfies

$$U^s(1, z_s) = U^s(0, z_s) + B_s \quad (58)$$

The form of utility assures that consumption of the ordinary good is the same regardless of whether the treatment is chosen or not. The difference between the points  $z_{1s}$  and  $z_{0s}$ , which denote the endpoints of the linear portions of  $\widehat{M}_s(z_s)$ , as shown in Figure 5, is under the restrictions on utility just the cost of the treatment,  $p_s$ . The efficacy of treatment is the slope of the linear portion, that is, the utility increment  $B_s$  divided by  $z_{1s} - z_{0s} = p_s$ . Thus the efficacy of the treatment is  $B_s/p_s$ .

In the optimum there is a critical efficacy  $m^*$  and the treatment of disease  $s$  is included in the optimum plan if and only if  $B_s/p_s \geq m^*$ , *i.e.*,

$$p_s \leq B_s/m^* \quad (59)$$

The condition in equation (59) characterizes the optimum.<sup>4</sup>

## 4.2 Copayment

In the absence of insurance, treatments are selected differently. Supposing endowed wealth of  $\bar{w}_s$  in state  $s$ , a fair premium of  $P$  in all states, then the reservation price  $R_s(B_s, \bar{w}_s - P)$  satisfies

$$U^s(0, \bar{w}_s - R_s(B_s, \bar{w}_s - P)) + B_s = U^s(0, \bar{w}_s - P) \quad (60)$$

The function  $R(B_s, \bar{w}_s - P)$  is concave, as can be seen in a short derivation from equations (58) and (60), because the utility functions are concave.

---

in the other. The transfer continues until, in one state or the other, one end of the linear portion is reached. At that point the number of states having  $z_s^*$  in the linear portion has been reduced by one. If there are still two such states, repeat the operation. Eventually, only a single "gambling" state remains.

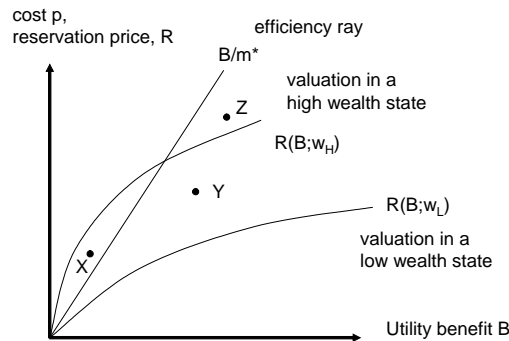
<sup>4</sup>If  $p_s = B_s/m^*$  then the treatment is included in the sense that it is part of an optimizing gamble. As mentioned previously, there needs to be, at most, one such treatment.

In the absence of insurance, a person chooses the treatment in state  $s$  if and only if  $R_s(B_s; \bar{w}_s) \geq p_s$ . Under a copayment plan with rate  $\theta$ , a treatment is chosen if and only if  $\theta p_s \leq R_s(B_s; \bar{w}_s - P)$ , i.e.,

$$p_s \leq \frac{R_s(B_s; \bar{w}_s - P)}{\theta} \tag{61}$$

The curve  $R(B_s; \bar{w}_s - P)/\theta$  represents the critical relationship of cost and benefit. From equation (61) a treatment is purchased in the copayment plan if and only if it lies below the curve.

Although the derivations are not complex, it is useful to fix the ideas by seeing them in a diagram. Figure 6 illustrates the relation of no-insurance to the optimum. Figure 7 concerns itself with copayment insurance. In both figures the region of efficiency is the portion lying below the "efficiency ray," the slanting line  $p_s = B_s/m^*$ , as dictated by equation (59).



Choice of treatment depends upon wealth in the state. High wealth leads to high reservation prices. A person without insurance may purchase inefficient treatments like X when wealth is high or omit efficient treatments like Z in high-wealth and low-wealth states.

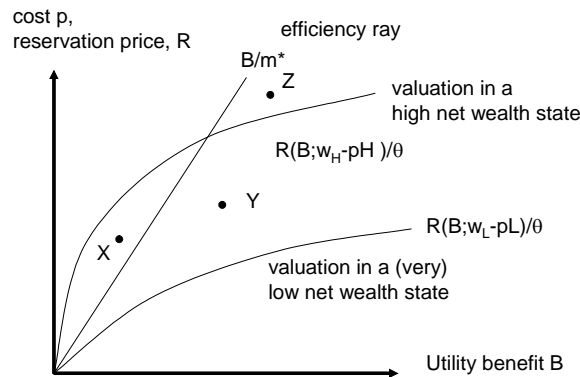
Figure 6. Inefficiency of the no-insurance standard

In Figure 6 several treatments  $X$ ,  $Y$ , and  $Z$  are represented by their benefits on the horizontal axis and their prices on the vertical axis. The reservation curve in the absence of insurance,  $R(B_s; \bar{w}_s)$  is drawn for two situations, a high-wealth state and a low-wealth state. It is easy to show that, when state wealth  $\bar{w}_s$  equals the optimum state wealth  $z_s^*$ , the reservation curve  $R(B_s; \bar{w}_s)$  is tangent to the efficiency ray at the origin. If wealth in the state is lower, a curve like the lower one in Figure 6 represents the reservation price, and if wealth in the state is higher, the higher curve applies.

Arrow, cited above, suggested that efficient treatments should be approximately the same as those chosen in the absence of insurance. The

idea is a hidden assumption in much public discussion of moral hazard. Figure 6 illuminates the two ways in which the idea is unsound. Efficient treatments like  $Z$  lie above the reservation price even in the high-wealth state, while inefficient ones like  $X$  are below it. The efficient  $Z$  would be omitted in the absence of insurance, but inefficient treatments like  $X$  would be chosen. Additionally, the inclusion of treatments like  $Y$  will depend upon the level of wealth in the state. Thus for efficiency, allowable treatments cannot be judged by standards that exist in the absence of insurance.

Figure 7 illustrates demands for treatments under a copayment insurance. The efficiency ray is exactly the same as in the previous figure, but the reservation curve is inflated by the copayment and diminished by the cost of the premium to become  $R(B_s; \bar{w}_s - P)/\theta$ . In typical instances the curves lie higher than ones in Figure 6.



Choice of treatment depends upon wealth in the state. High wealth leads to high reservation prices. A person without insurance may purchase inefficient treatments like  $X$  when wealth is high or omit efficient treatments like  $Z$  in high-wealth and low-wealth states.

Figure 7: Inefficiencies of copayment insurance

Nevertheless, as before, copayment plans lead to sub optimum treatment choices in two ways: Consumers purchase some low-cost treatments like point  $X$  that are not part of the optimum, and they omit treatments like  $Z$  that should be included in the optimum. The intuition behind the omissions is obvious. The client is very poor in this ill state and the treatment is very expensive. Either way, copayment insurance neglects efficient treatments. Overspending is also a possibility, as illustrated by point  $X$ .

## 5 Numerical examples: indivisible treatments

In the example the healthy state is state zero and the sick states are  $s = 1, 2, \dots$ . The utility functions in states are

$$\begin{aligned} u^s(z, 0) &= \ln(z) \\ u^s(z, 1) &= \ln(z) + B_s \\ u^s(z - p_s, 0) &= \ln(z - p_s) + B_s \end{aligned} \tag{62}$$

Healthy utility is singled out by having a very small benefit value  $B_0$  and an even smaller price  $p_0$ . It seems odd that utility in a sick state can be higher than in the healthy one, as occurs when  $B_s$  is large and  $p_s$  is small. However, the remedy for that would be to subtract appropriate constants from the utility function in the sick states. Such constants play no essential role and therefore they are omitted.

Data are generated mostly randomly and shown in Table 9. Probabilities are chosen arbitrarily, although the probability of state zero, the no illness state, is fifty percent, and that is not too far from the Berk and Monheit number. For computational reasons it is convenient to have a treatment in state zero, but the treatment is almost costless and the interpretation is still no illness. Wealth in state zero is unity. Data for the other states is produced by generating random numbers distributed uniformly between zero and unity, with the exception of state four. In that state the benefit and price are inserted specifically for the purpose of demonstrating the possibility of underspending on health care in the copayment regime. Thus, the data are

Table 9: The mostly random data

state	probability	$B_s$	$p_s$	$\bar{w}_s$
0	.5	0.1	0.0001	1
1	.05	0.976852911	0.30686951	0.760244605
2	.05	0.705818029	0.050185522	0.788647915
3	.05	0.083337802	0.671881087	0.099843509
4	.05	2	1.28	0.351779741
5	.05	0.840681488	0.544217268	0.617511774
6	.05	0.969113434	0.874360534	0.371805647
7	.05	0.30613548	0.767879619	0.291932152
8	.05	0.557065686	0.153426888	0.854338878
9	.05	0.016796868	0.173367424	0.250795889
10	.05	0.852796766	0.528367329	0.85510351
	1			0.7621

The endowed-wealth  $\bar{w}_s$  in column five of Table 9 can equally be interpreted as a utility shift. A small endowed wealth is equivalent to a very adverse shift in utility. The effect of such a shift is like the effect of a decrease in wealth: it raises the marginal utility of wealth. In what follows, the shift will be interpreted as a wealth effect.

In Table 9 the number in the lower right is the expected value of contingent wealths, and in the lower left is a check sum for the probabilities.

Optimum insurance in this context is characterized by a choice of critical efficacy  $m^*$  which assures that the expected costs of consumption and medical treatments are just equal to the expected value of wealth. The following table illustrates the optimum.

Table 10: First-best optimum

$m = 1.54773$	treatment included?	optimum state wealth
0	1	0.646207525
1	1	0.952977042
2	1	0.696293047
3	0	0.646107525
4	1	1.926107525
5	0	0.646107525
6	0	0.646107525
7	0	0.646107525
8	1	0.799534414
9	0	0.646107525
10	1	1.174474855

Expectation = .7621

The program is optimum because the condition of equal marginal utility across states is satisfied by construction and the budget constraint, which says that the expected value of consumption – here .7621 in column "required wealth in state" – is the same as the expected value of wealth – here .7621 in the  $\bar{w}_s$  column – is also satisfied. "Solver" in Microsoft Excel was used to find the marginal utility  $m$  at which the budget constraint is satisfied.

Computation of values in states is illustrated by considering state one. The efficacy there is  $B_1/p_1 = .976853/.30687$  and is greater than the critical efficacy level of  $m = 1.54773$ . Therefore the treatment is included. Because the treatment is included the client is consuming at a point on the "treated" utility function  $u^1(z_1 - p_1, 1)$ . Thus to find  $z_1$

solve

$$m = \frac{\partial u^1(1, z_1 - p_1)}{\partial z_1} \tag{63}$$

which is

$$1.54773 = \frac{1}{z_1 - .30687} \tag{64}$$

The solution of the latter is the number in the right-most column,  $z_1 = .952977$ . And so on for the other states, using a zero in the treatment-included column when the efficacy of treatment is too low and in that case solving

$$m = \frac{\partial u^1(0, z_1)}{\partial z_1} \tag{65}$$

Incidentally, all of the states in which the treatment is avoided have the same state wealth of 0.646107525.

Now consider the same problem under copayment-style insurance. To demonstrate the main result it is adequate to take a reasonable level for the copayment rate, twenty percent. This is not the optimum copayment rate, and the implications of that statement will be discussed below. Twenty percent is a reasonable rate.

Table 11: Copayment at 20 percent

$\theta = .2$	$Ez_s = .7621$	$P = .08085$		
state	$p_s$	$R/\theta$	treatment included?	state wealth
0	0.0001	0.437343436	1	0.91923
1	0.306869517	2.118032733	1	0.924890218
2	0.050185522	1.791774338	1	0.747946332
3	0.671881087	0.007593576	0	0.018993509
4	1.28	1.17131694	0	0.270929741
5	0.544217268	1.525685368	1	0.972035588
6	0.874360534	0.90280734	0.5	0.640699861
7	0.767879619	0.278325685	0	0.211082152
8	0.153426888	1.651833331	1	0.896230389
9	0.173367424	0.014153593	0	0.169945889
10	0.528367329	2.221251163	1	1.196947374
				0.762100053

The distributions of benefit, price, and wealth are the same as before, and expected wealth of the client is the same, .7621. The copayment rate is a

given constant theta, and the premium is denoted by  $P$ . Knowing it, the client has net wealth equal to the state wealth less the premium. That net wealth permits him to compute a reservation price  $R(\bar{w}_s - P, B_s)$ . Using the fixed copayment rate, the client will pay up to  $R/\theta$  for the health care and when the price is less than that number, as it is in state one, the treatment is included. Knowing these calculations, the insurer's cost in, for instance, state one is  $(1 - \theta) * p_1 = .8 * .30687 = .2454956$ . Using the costs, the insurer needs to be paid the expected value, which is the  $P$ . The optimum for the consumer is found when the premium paid is equal to the expected cost. As mentioned earlier, indivisibility of the treatments implies discontinuities in demand. When a discontinuity is encountered, as it is here, the solution requires giving the treatment with a probability strictly between zero and unity. In the solution here the treatment in state six is given with probability .5.

Computations completed, Table 12 compares the two allocations.

Table 12: Compare the first-best with a 20 percent copayment allocation

state	first-best state wealth	copayment state wealth	first best treatment included?	copayment treatment included?
0	0.646207525	0.91923	1	1
1	0.952977042	0.924890218	1	1
2	0.696293047	0.747946332	1	1
3	0.646107525	0.018993509	0	0
4	1.926107525	0.270929741	1	0
5	0.646107525	0.972035588	0	1
6	0.646107525	0.640699861	0	0.5
7	0.646107525	0.211082152	0	0
8	0.799534414	0.896230389	1	1
9	0.646107525	0.169945889	0	0
10	1.174474855	1.196947374	1	1
Expectation	.7621	0.762100053		

States five and six illustrate the well-known over-expenditure problem. Treatments are included that are not part of the optimum. In state four, however, the copayment plan omits a treatment that is part of the first-best. The reason it is excluded can be seen in the endowed wealth (or utility shifter) ( $w_4$ ). Wealth in the state is low (or the utility shift is large). The patient has net wealth, after paying the premium, of  $.35178 - .08085 = .27263$ , and the cost of the treatment would be  $.2 * 1.28 = .256$ . At this level of expense, the patient prefers to consume

instead of getting medical treatment.

Twenty percent is not the optimizing level of copayment. The optimum level of copayment is 100%, that is, no insurance. Clearly the treatments purchased under no insurance are fewer than those purchased in the first best optimum. Thus the numerical example dramatically reinforces the general argument of this paper, which is that treatments under optimum copayments are likely to be inadequate, rather than excessive. The random data illustrate again that copayment systems are not viable, and they again call into question whether health insurance ever existed under such a system.

## 6 Concluding remarks

Given the implausibility of a pure copayment system in health insurance, what aspects of the present analysis can be salvaged for examination of other wealth transfer systems? The distinction among neutral, depressive, and intensive states is essential to any model of transferring wealth among health states. Similarly, grave illnesses reduce wealth, and wealth transfers to the gravely ill should anticipate the effects of that loss.

The many defects of copayment-style health insurance suggest palliatives such as deductibles, variable copayment rates, and stop-loss provisions, which are important in practice. Nevertheless, the bulk of the literature seems to assume that health insurance is ruled by copayment-style incentives with insignificant modifications. Perhaps a better idea is to view it as modifications with insignificant elements of copayment incentives. More recent work on overspending has examined the incentives of physicians, which could be helpful in clarifying the behavior of the patients. Such clarification is urgently needed because the standard story of copayment insurance no longer holds up.

Copayment has been used as a policy model most often to suggest excessive expenditures. That is incorrect as a theoretical matter, and writers of real-world health insurance seem to understand the problem. Most past and present medical plans have a "stop-loss" feature that reduces to zero the copayment in the most expensive illnesses. Thus actual insurance contracts recognize the possibility of inefficient under spending.

## References

- [1] Arrow, Kenneth J., "Uncertainty and the Welfare Economics of Medical Care," *American Economic Review*, December 1963, 53(5), pp. 941-973.
- [2] \_\_\_\_\_, "The Economics of Moral Hazard: Further

- Comment," *American Economic Review*, June 1968, 58(3), pp. 537-538.
- [3] \_\_\_\_\_, "Le rôle des valeurs boursières pour la répartition la meilleure des risques," *Colloques Internationaux du CNRS, XL*, 41-48. English translation (1964) "The role of securities in the optimal allocation of risk-bearing," *Review of Economic Studies* 31, 91-96.
- [4] Bergstrom, Theodore. "Soldiers of Fortune?" in Walter P. Heller, Ross M. Starr and David A. Starrett, eds., *Essays in Honor of Kenneth J. Arrow, Volume II: Equilibrium Analysis*. New York: Cambridge University Press, 1986, pp. 57-80.
- [5] Berk, M.L., and Monheit, A.C., "The concentration of health expenditures: an update," *Health Affairs*, 1992, 11(4), 145-149.
- [6] Cutler, David M., and Richard Zeckhauser, "The Anatomy of Health Insurance," *Handbook of Health Economics, Vol.1A*, Cuyler, A. and Newhouse, J. eds., Amsterdam: Elsevier, 2000, pp 563-643.
- [7] de Meza, David, "Health Insurance and the Demand for Medical Care," *Journal of Health Economics* 2, North-Holland Publishing Company, 1983, 47-54.
- [8] Friedman, Milton and Savage, Leonard J. "The Utility Analysis of Choices Involving Risk." *Journal of Political Economy*, August 1948, 56, pp. 279-304.
- [9] Garratt, Rod and Marshall, John M. "Public Finance of Private Goods: The Case of College Education." *Journal of Political Economy*, June 1994, 102(3), pp. 566-82.
- [10] Garratt, Rod and Marshall, John M. "Optimum Tuitions, Taxes and Fees when Completion is Uncertain." *Education Economics*, December 1995, 3(3), pp. 219-234.
- [11] Gollier, Christian. "Economic Theory of Risk Exchanges: A Review." in Georges Dionne, ed., *Contributions to Insurance Economics*, Boston: Kluwer Academic Publishers, 1992, pp. 3-23.
- [12] Havighurst, Clark C. *Health Care Choices: Private Contracts as Instruments of Health Reform*, The AEI Press, Washington, D.C. 1995.
- [13] Himmelstein, David U.; Warren, Elizabeth; Thorne, Deborah; and Woohandler, Steffie, "MarketWatch: Illness and Injury As Contributors To Bankruptcy" *Health Affairs*, February 2, 2005. <http://content.healthaffairs.org/cgi/content/full/hlthaff.w5.63/DC1>
- [14] Marshall, John M. "Gambles and the Shadow Price of Death." *American Economic Review*, March 1984, 74(1), pp. 73-86.
- [15] \_\_\_\_\_, "Insurance Theory: Reserves versus Mutuality." *Economic Inquiry*, December 1974, 12(4), pp. 476-92.

- [16] \_\_\_\_\_, "Moral Hazard", *American Economic Review*, December 1976, 66(5), 880-890.
- [17] Pauly, Mark V. "The Economics of Moral Hazard: Comment." *American Economic Review*, June 1968, 58(3) part 1, pp. 531-537.
- [18] \_\_\_\_\_, "Insurance Reimbursement," in *Handbook of Health Economics, Vol.1A*, Cuyler, A. and Newhouse, J. eds., Amsterdam: Elsevier, 2000, 537-560.
- [19] Nyman, John A. "The economics of moral hazard revisited," *Journal of Health Economics*, 18(6) , December 1999, pp. 811-824
- [20] Raviv, Artur. "The Design of an Optimal Insurance Policy." *American Economic Review*, March 1979, 69(1), pp. 84-96.
- [21] Thaler, Richard and Rosen, Sherwin, "The Value of Saving a Life," in Nestor E. Terleckyj, ed., *Household Production and Consumption*, New York, National Bureau of Economic Research, 1976, 265-302.
- [22] Zeckhauser, Richard J. "Medical Insurance: A Case Study of the Tradeoff Between Risk Spreading and Appropriate Incentives," *Journal of Economic Theory*, March 1970, 2, pp. 10-26.
- [23] Zweifel, Peter, and Manning, Willard G., "Moral Hazard and Consumer Incentives in Health Care," in *Handbook of Health Economics, Vol.1A*, Cuyler, A. and Newhouse, J. eds., Amsterdam: Elsevier, 2000, pp 409-459.