Comment

Does the economics of moral hazard need to be revisited? A comment on the paper by John Nyman

Åke Blomqvist*

Department of Economics, University of Western Ontario, London, ON, Canada N6A 5C2

Received 1 April 2000; accepted 3 May 2000

Keywords: Moral hazard; Health insurance; Degree of insurance coverage; Optimal degree of insurance

In a recent paper, John Nyman (Nyman, 1999b) revisits an issue that has been central to much of the literature in health economics in recent years, namely the trade-off between the gains from insurance, on the one hand, and the efficiency losses from the moral hazard effect that arises as a result of the implicit subsidy to health services utilization under conventional insurance plans, on the other. Specifically, he argues that in the existing literature, the significance of the latter has been exaggerated, while conventional approaches tend to understate the importance of the former. The result, according to Nyman, has been a tendency in the health policy debate to put too much emphasis on insurance plan features such as consumer cost sharing and provisions to control utilization in managed-care plans.

Although several of Nyman’s criticisms of conventional methodology are correct and valid, in this comment, I will attempt to show that they are largely irrelevant to the question that has been the focus of the literature on health insurance ever since the seminal paper by Zeckhauser (1970), namely that of the optimal degree of demand-side cost sharing. The reason is that, implicitly, Nyman’s paper addresses the issues of what would happen if conventional approaches would be applied to the problem of estimating the net gains from insurance against the alternative of no insurance at all. Most of the recent literature in this area, in contrast, has concerned a question at the margin: For a
given degree of insurance protection, would consumers be better off with a somewhat higher (or lower) degree of insurance? By extension, what is the optimal degree of insurance protection? As I will show below, the potential biases that Nyman points to in existing approaches are not of major significance to the recent literature on these issues.

The distinction between the total and marginal gains from insurance is analogous to that which is made in the public finance literature between the aggregate and marginal excess burden of a system of distortionary taxation. Earlier literature tried to estimate the total (or average) excess burden. More recently, however, the focus has been on the marginal excess burden: Since every society requires at least some government spending, and since non-distortionary taxation is not feasible, the relevant issue is the balance between the benefits of public spending and the real cost of public funds at the margin. The logic with respect to health insurance is similar. For example, if access to certain kinds of very costly but vitally necessary health services can be guaranteed through payment of a small insurance premium, it obviously makes sense for a risk-averse individual to have a contract that accomplishes this. However, for less serious illness and less costly health services (which the individual could afford to pay for even if their insurance did not cover them), and for which the insurance premium would be relatively high because they are used frequently, the issue is less clear. It is the question where to draw the line in this regard that has been the focus of the post-Zeckhauser literature.

One of the principal points in Nyman’s paper is that the conventional approach underestimates the gains from insurance because it fails to take into account what can be termed the “affordability” effect: that certain interventions are so costly that, in the absence of insurance, the individual would have to go without them, even if they potentially would have very large health benefits (in the limit, they may be necessary for the individual’s survival). To what extent is this argument relevant to the question of what constitutes an optimal degree of insurance coverage?

The answer depends in part on how the latter concept is defined. In Zeckhauser’s paper (Zeckhauser, 1970), the degree of insurance coverage was defined by a single, constant, co-insurance parameter specifying what fraction of total health care costs would be paid by the patient out of pocket. If insurance is defined in this way, the affordability motive may have considerable relevance in the sense that the optimal choice of co-insurance parameter might depend heavily on the benefits from not precluding access to certain very costly but vitally necessary services. ¹

However, it has long been recognized that insurance coverage through a constant co-insurance parameter is highly inefficient in comparison with plans that provide a higher level of coverage (a lower co-insurance rate) for large expenditures. This was implicitly allowed for in the design of the Rand Health Insurance Study, the results of which have provided by far the most important empirical basis for the contemporary debate about optimal insurance. While the central component of the study was a comparison of health care spending among individuals and families in insurance plans with different co-insurance rates, all families were subject to a maximum annual expenditure limit which was the lesser of a specified

¹ Nyman has explored this issue at length in another paper (Nyman, 1999a).
percentage of the family’s income (5%, 10%, or 15%), or US$1000 (equivalent to about US$2500 in 1995 dollars). Since this kind of limit certainly would be lower than what most people would consider a reasonable definition of “affordability” for urgently needed health services, the potential bias introduced by neglecting the affordability motive is unlikely to be of major significance in comparison with the relative efficiency of the plans in this set, or plans with lower deductibles.2

With respect to the efficiency loss from moral hazard, Nyman makes an important point regarding the difference between the compensated and Marshallian demand curve when the magnitude of the loss is estimated as an area above a demand curve for (implicitly) subsidized health services. Since empirical estimates typically yield Marshallian demand curves while it is the compensated one that should be the basis for welfare loss estimates, it is important to have an accurate method for accounting for the income effects that give rise to the difference between them when estimating the compensated one from the Marshallian one. According to Nyman, the conventional method for doing this is likely to yield misleading results for the case of health services because of the highly skewed distribution of health expenditures across the population.

His point can be simply illustrated by considering a market in which all potential buyers have the same income and in which, on average, consumers spend 5% of their income on the good traded in it. Suppose, however, that the average conceals a very skewed distribution of purchases, so that, for example, 80% of consumers spend nothing on the good, while the remaining 20% spend 1/4 of their income on it. Suppose further that the elasticity of the Marshallian demand curve has been estimated at −0.5, while the income elasticity has been estimated at +1.0. The conventional formula to obtain the elasticity of the individual’s compensated demand curve is

\[ \epsilon_c = \epsilon_m + \eta_y s \]

where \( \epsilon_c, \epsilon_m, \eta_y \) and \( s \) are the elasticities of the compensated and Marshallian demand curves with respect to price, the income elasticity of demand, and the share of the consumer’s budget spent on the commodity in question, respectively. If one wants to apply this formula to the market demand curve, the relevant budget share clearly is that of the 20% minority of the population which purchases health care, not the average share for the population as a whole, so that in the example, the compensated elasticity is \(-0.5 + 0.25 = -0.25\); erroneously using the average budget share of 5% would yield a compensated elasticity of \(-0.5 + 0.05 = -0.45\).

Clearly, the logic in this example is applicable to the case of health care. For example, in the Rand study, a substantial proportion of families had no medical expenditure at all in each given year, and for those with some expenditures, the distribution of spending was highly skewed. In calculations based on the price elasticity of demand estimated from the Rand study, Nyman suggests that the true welfare losses from the moral hazard effect may be only a third of the values that result from calculations that do not properly correct for income effects (Nyman 1999b, p. 818).

However, while the point is correct in general, it is of limited relevance to the comparison of the relative moral hazard losses in the different plans in the Rand study that have been

---

2 Manning et al. (1987) are explicit on this point. They note that the results from the Rand study cannot be directly compared with estimates of the welfare loss based on models that represent cost-sharing as a constant co-insurance rate, but “make no apologies for this intentional non-comparability; a constant co-insurance rate, . . . , is not an insurance policy that theory suggests would be optimal”.

undertaken in the literature. First, the relevant budget share for the correction would be the share of each family’s out-of-pocket spending on health care, not total spending. Under the design of the Rand experiment, the former was never more than the upper limit referred to above. Because the limit was defined as the lesser of 5%, 10%, or 15% of family income or US$1000, it was less than 10% of family income on average and for most families with positive spending it was of course less than the limit. In the example below, I will conservatively set the share variable at 8% of family income. Moreover, Nyman uses an estimate of the income elasticity of demand for health services of 0.38, a number derived from a study based on data for individual tax filers 55 years of age or older. The income elasticities estimated from the Rand experiment itself were generally lower, ranging from 0.2 to 0.4 (Newhouse et al., 1993, p. 47). Using the mid-point of 0.3 and an uncompensated price elasticity of demand of −0.186 (corresponding to the arc elasticity of demand in the example used by Manning et al. (1987, p. 270), produces an estimated elasticity of compensated demand of −0.162, or 87% of the uncompensated one (rather than 33%, as in Nyman’s example).

Moreover, the relation between the elasticities used in Nyman’s calculation is based on point elasticities, whereas the price elasticity estimated from the Rand study is an arc elasticity. Taking this into account further reduces the significance of the income effect. To see this, consider the example in Manning et al., (1987, p. 270). In this example, the average annual medical (and dental) expenditure under the “free plan” (with 0 cost-sharing) are US$1038, while they are US$713 under the 95% cost-sharing plan. The arc elasticity based on these two expenditure levels, at prices of 0 and 0.95, respectively, is −0.186. Taking the free plan as the reference case, how large is the income effect of moving to a 95% plan? Assuming that US$713 represents 8% of the consumer’s income, the compensation necessary to leave the consumer as well off as under the free plan would be 0.95×8=7.6% of the consumer’s income. With an income elasticity of +0.3, this would lead to an increase in medical expenditure of about 2.3% of US$713, or US$16.4. The welfare loss associated with a price decrease from 0.95 to 0 in using the uncompensated demand curve is US$170.6, (1038−713) (1−(0.95/2)), assuming the demand curve is linear. Accounting for the income effect requires replacing 713 by 730; with this value the welfare loss becomes US$161.7, or 95% of the estimate based on the uncompensated demand curve, again well above the 33% figure that Nyman suggests.

3 Approximately equal number of families were assigned to each percentage limit category, and median income of those enrolled was in the range US$6400–11,800 in 1969 dollars (Newhouse et al., 1993 pp. 16 and 415). By the time the data were collected in second half of the 1970s, these income figures had risen substantially (by more than 80%), reducing the value of the US$1000 limit (which was constant in nominal terms) as a share of income. 4 According to Newhouse et al. (1993, p. 358), 65% of families in the 95% cost-sharing plan (see below) had expenditures less than the upper limit in any given year. 5 The study is done by Feenberg and Skinner (1995). As the authors note, a major item of health care spending for this group is nursing home care, an item for which the income elasticity of demand is likely to be higher than for the kinds of medical expenditure incurred by those in the Rand study (which excluded the elderly). Another reason for the large income effect correction in Nyman’s calculation is that as his share parameter he uses the percentage of medical care in personal consumption expenditure (16% in 1997). But this share presumably includes health insurance premiums, an item that is irrelevant to estimating the income effect of changing an insurance plan’s co-insurance rate.
The calculations above pertain to what Manning et al. refer to as their minimum estimate of the welfare loss. Their maximum estimate (US$274.6 rather than US$161.7 on average) is derived by assuming that the marginal value of the last unit of health services used under the 95% plan is US$0.31 rather than US$0.95 as assumed in the minimum estimate. The uncertainty regarding which is the correct one arises because some individuals (those that have exceeded their spending limits) under the 95% plan continue to face an effective marginal price of 0 (rather than 0.95) for health services and will therefore utilize services beyond the point where their marginal value corresponds to 0.95. In the minimum estimate, it is implicitly assumed that all individuals who reduce their service use when going from the free plan to the 95% plan do so until the last unit they use has a marginal benefit of 0.95. But this assumption is unlikely to be valid. From the numbers, it is clear that correcting for this effect is likely to be more important quantitatively than correcting for the income effect.

The paper by Manning et al. does not directly address the issue, first raised by Feldstein (1973), of the “welfare loss from excess health insurance” since they do not try to quantify the gains from risk reduction. However, their work is extended by Feldman and Dowd (1991) who incorporate both the moral hazard loss and the gains from risk reduction. They conclude that the Rand study 95% co-insurance plan is superior, from an efficiency point of view, to a free care plan: The estimated gain from the higher degree of insurance under the free care plan is not large enough to offset the increase in the estimated moral hazard loss.

Although neither of these papers addresses the question of what would constitute an “optimal” degree of insurance, given the evidence from the Rand study, that question is explicitly addressed in two other works: Manning and Marquis (1996), and in Newhouse et al. (1993, Chap. 4, pp. 137–139). An interesting issue is why, in the latter study, the conclusion is that the optimal plan involves a higher degree of insurance, both in the form of a lower co-insurance rate and in the form of a lower maximum out-of-pocket cost “stop-loss” provision, than in the Rand 95% co-insurance plan, in spite of the result in Feldman and Dowd that the reduction in the moral-hazard effect when moving from free care to a 95% co-insurance plan is greater than the estimated cost of increased risk-bearing. The explanation appears to lie in the non-linearity of the estimated demand curve for health services as a function of the co-insurance rate. Although it is based on the same data, the Manning and Marquis paper arrives at a substantially different optimal plan, with both a higher co-insurance rate and a higher maximum expenditure limit. Further analysis of what causes this difference would clearly be of interest.

Nyman concludes his paper by suggesting that the moral-hazard loss associated with the implicit subsidy to services utilization under conventional health insurance plans should be regarded as a transactions cost similar to the administration cost of any insurance plan. Implicitly, this is precisely the way that the moral hazard effect has been treated in the papers discussed above: In estimates of the optimal degree of insurance, the marginal moral hazard loss plays a role quite similar to that of administration costs in the conventional analysis of insurance. A well-established result in the insurance literature is that when there are positive administration costs so that actual premiums exceed the actuarially fair levels by some “load factor”, it is not optimal, even when individuals are risk-averse, to be fully insured against a given loss (see, e.g. Folland et al., 1997, pp. 235–241). The result that the
optimal degree of insurance against the cost of health care should depend on the marginal moral hazard loss is exactly analogous.

Nyman is correct in observing that the affordability issue, and the special nature of the income effect in the choice of health insurance, may be important if the purpose is to estimate the gains from some insurance as against the alternative of no insurance at all. However, the results in the recent literature on the issue of the optimal degree of patient cost sharing have been much less affected by the potential biases that may arise from these factors than Nyman’s paper suggests.

References