NBER WORKING PAPER SERIES

COST EFFECTIVENESS ANALYSIS AND THE DESIGN OF COST-SHARING IN INSURANCE: SOLVING A PUZZLE

Mark Pauly

Working Paper 18481
http://www.nber.org/papers/w18481

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2012

Cost Effectiveness Analysis and the Design of Cost-Sharing in Insurance: Solving a Puzzle
Mark Pauly
NBER Working Paper No. 18481
October 2012
JEL No. I11,I13

ABSTRACT

The conventional model for the use of cost effectiveness analysis for health programs involves determining whether the cost per unit of effectiveness of the program is better than some socially determined maximum acceptable cost per unit of effectiveness. If a program is better, the policy implication is that it should be implemented by full coverage of its cost by insurance; if not, no coverage should be provided and the program should not be implemented. This paper examines the unanswered question of how cost effectiveness analysis should be performed and interpreted when insurance coverage can involve non-negligible cost sharing. It explores both the question of how cost effectiveness is affected by the presence of cost sharing, and the more fundamental question of cost effectiveness when cost sharing is itself set at the cost effective level. Both a benchmark model where only "societal" preferences (embodied in a threshold value of dollars per unit of health) matter and a model where individual willingness to pay can be combined with societal values are considered. A common view that cost sharing should vary inversely with program cost effectiveness is shown to be incorrect. A key issue in correct analysis is whether there is heterogeneity either in marginal effectiveness of care or marginal values of care that cannot be perceived by the social planner but is known by the demander. The cost effectiveness of a program is shown to depend upon the level of cost sharing; it is possible that some programs that would fail the social test at both zero coverage and full coverage will be acceptable with positive cost sharing. Combining individual and social preferences affects both the choice of programs and the extent of cost sharing.

Mark Pauly
Department of Health Care Management
University of Pennsylvania
208 Colonial Penn Center
3641 Locust Walk
Philadelphia, PA 19104-6218
and NBER
pauly@wharton.upenn.edu

Introduction.

Many countries use some form of cost effectiveness analysis to inform decisions about coverage under insurance. However, there is not, to my knowledge, any conceptual model that links the findings of cost effectiveness analysis to the general question of insurance design. Specifically, given some cost-effectiveness result, what if anything does that imply about levels of coverage or cost sharing?

Formal analyses of cost effectiveness and coverage are almost all limited to considering binary social insurances, with zero cost sharing for all covered services, versus no insurance coverage or reimbursement at all. This means that, to my knowledge, they have not addressed a policy question of growing importance, as many countries move to introduce some positive cost sharing: how should cost effectiveness analysis be used in the presence of non-trivial cost sharing? As noted by Ioannidis and Garber (2011), "even in countries with national health systems that have minimized the role of individual cost sharing, individuals pay growing shares of the costs of health and health services out of pocket" (p.6). These questions are also of increased special importance in the United States as the insurance policies specified by government to be offered in Exchanges under health reform will typically involve deductibles and coinsurance over much of the range of spending. Analyses of cost effectiveness also cannot answer the question of what cost sharing should be or when it should be changed.

Perspective matters. Most cost-effectiveness analyses take an extra-welfarist or public payor perspective, but some others intend to incorporate consumer valuations into a more conventional economic welfare perspective. There have been some attempts in the literature to discuss the more general question of whether there can be a link between differing perspectives and cost effectiveness analysis as it is usually practiced. To my knowledge, there is as yet no clear

answer to this question. In this paper I will argue that broadening coverage options to consider cost sharing paradoxically makes a contribution to providing an answer to a larger question about the role of perspective that is still unresolved.

Of course, the answer to the question of the relationship between cost effectiveness values and cost sharing depends both on the perspective taken and the empirical facts. So I first outline the simple and correct application of a decision rule to treatment choices based on cost effectiveness in the "binary coverage" setting, when insurance either covers 100 percent of the cost of a given type of care or leaves it entirely uncovered, *and* the extra welfarist approach is taken. I show that this approach usually is based on two assumptions: a single value for expected improvement in health outcomes is to be applied to all patients, and a single monetary value for those expected marginal benefits prevails. This is the approach, avowedly "extra-welfarist," much favored at present in the United Kingdom by the NICE advisory body.

However, I then show that opening the door to consideration of cost sharing means that many things, including this perspective, might appropriately be modified. Modifications are needed if there is heterogeneity in either effectiveness of the treatment across patients or in the values citizens place on health outcomes, and that heterogeneity is determined to be relevant to policy. I show that the ideal level of "interior" cost sharing depends on whether consumer values are assumed to be relevant, on how much consumer values really vary, and most especially on whether the extent of variation in expected effectiveness across patients is perceived by patients but cannot be known by the insurer. I briefly consider as well the possibility that social values are variable or uncertain.

Some of the issues I discuss have recently been addressed by Basu (2011) in the context of comparative effectiveness analysis. I show how his insights can be expanded to traditional cost effectiveness analysis.

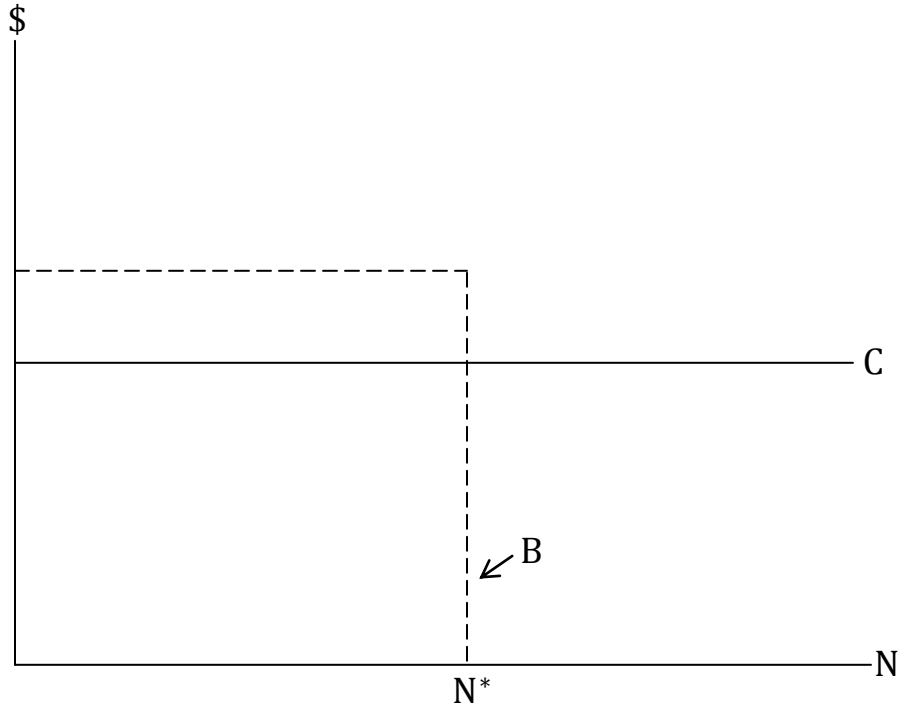
A benchmark cost effectiveness decision model: The extra-welfarist case and beyond.

The standard model of optimal decisions using cost effectiveness analysis is well specified in the literature. Agreement is reached on what aspects of resource use are to be considered as relevant costs and what measures are to be accepted for effectiveness. Each distinct type of treatment is assumed to have a uniform level of expected benefit for all patients who are targeted to or eligible to receive that treatment. (This effectiveness is potentially conditional on characteristics of different patient subpopulations that insurers can observe.) The treatment can be characterized by an incremental cost effectiveness ratio (ICER): this ratio measures (compared to some alternative next best treatment or no treatment) the incremental cost and incremental effectiveness of the treatment being analyzed for the subpopulation in question. We will describe the form of this ratio in the conventional "dollars per QALY" framework, where QALY means the use of some kind of measure of quality adjusted life years as the outcome or effectiveness measure. Importantly, both QALYs and their valuations lead to constant values for marginal benefit per unit of health.

This measure is then compared to a threshold value W for acceptable cost effectiveness, stated in terms of a maximum value for dollars per QALY that will be socially accepted or approved; this can be interpreted as a measure of society's willingness to pay for QALYs. Usually the form of insurance coverage is taken as given. In many cases, as in the National Health Service in the United Kingdom, it takes the form of full coverage of the cost or price of

the care with zero cost sharing. In this case, design and decision rules are simple: if a treatment meets the threshold it should be fully covered; if it fails to meet the threshold it should not be covered at all. There is an important corollary to this model: all programs for which the cost per QALY is less than *W* should ideally be implemented.

Figure 1



A way of illustrating this case, which will be made less trivial in what follows, is shown in Figure 1. We consider a treatment that is provided in a single unit per person (such as an angioplasty). It costs *$C* per unit, and is to be delivered to N^* persons. If the amount of expected incremental health is the same for each one of those N^* persons, we can plot a social marginal benefit curve *B* in dollars by multiplying the incremental QALYs by *W*. If the program is cost effective, that marginal (equals average) benefit curve will lie above C for treatment of the N^* people, and then will fall to zero.

The determination of N^* is still a loose end. Ideally, the condition being treated will affect N^* people in exactly the same way—so that there will be a constant large marginal health benefit up to N^*—and then provide zero marginal benefit from treating one more than N^* persons. But N^* might simply represent the people declared eligible for treatment.

Complications also begin to arise, however, when methods of cost effectiveness analysis are used in a US setting where cost sharing is advocated (under the rubric of "value based cost sharing" [VBCS]) to be used to improve health system efficiency, but the benchmark model is still (apparently) being employed. What then is the relationship between coverage and CEA?

For example, Braithwaite and Rosen (2007) say: "High-value CEA assessments could be linked to a waiver of all cost sharing (that is, no copayments or deductibles), low-value or ambiguous CEA assessments could leave cost sharing unchanged, and very-low-value CEA assessments could be linked to increased [relative to prevailing values in public and private US health insurance] cost sharing" (p. 603). Some of the terms in this statement need to be defined, like "high value" and "ambiguous," but even so, in terms of the benchmark extra-welfarist model, it is not really complete and correct. If "high value" means above the threshold, low value means moderately below the threshold, and very low value means much below the threshold, cost sharing should be zero in the first case and unity in the second two cases. If high value and low value are both above the threshold, cost sharing should be uniformly zero for both. There is no room in this theory for anything other than binary values of cost sharing. (Considerations of ambiguity or imprecision in estimates of population-level cost effectiveness raise even more complex considerations that we will consider below.) If patient demand for each of the N^* people is positive at a zero user price (full insurance coverage), that price will bring forth the socially efficient outcome.

Relevance of the benchmark model.

A necessary condition for cost sharing to be relevant is that there is positive consumer demand at positive prices. Sometimes this may not be a plausible assumption, even in the U.S. For example, consider a population of poor insureds who have very low private values (relative to cost) for the potential treatment. Then the optimal decision rule is the same as before: If the value of *C*/QALY is less than *W* the program is to be fully covered by insurance. If the value is greater than *W* the program is not to be covered by social insurance, and whatever else happens next is of no concern to the social planner.

The closest approximation to this case in the US is Medicaid. In this program, cost sharing is only nominal if it exists at all. In addition, the private values per QALY, or V_i, of those in the program are close to zero, so only the social value plausibly matters. In this case only the social determined value W is relevant, so exactly the same decision rule can be used as in the extra-welfarist case.

Introducing heterogeneity.

Table 1 shows the possible combinations of heterogeneity and homogeneity I will discuss in this paper, and the models or demand curves that correspond to them.

Table 1

Monetary Value of Health \ Marginal Health Product	**Homogenous**	**Heterogenous**
Homogenous	(1) Simple Benchmark Model	(2) Preference-driven Demand
Heterogenous	(3) Health-driven Demand	(4) Combined Care

It is possible that the marginal effectiveness of care from a given treatment may vary across individuals (Cell 2). Usually one would assume that marginal benefit is positively correlated with illness severity, though this need not always be the case. Alternatively, different consumers may attach different values to a given health outcome, based on their preferences and incomes (Cell 3). Finally, both may vary (Cell 4). The presence of either kind of variation (over the relevant range) is sufficient to yield a downward sloping demand curve for medical care as a function of the coinsurance rate. The social choice model will presumably take variation in effectiveness into account, and may or may not pay attention to variation in marginal values of health in the social decision process. We now discuss cells (2) and (3) briefly, before considering more complex cases.

The simplest standard cost effectiveness model assumes that patients who will receive a treatment are homogeneous in terms of the expected marginal effectiveness of treatment; they are all of similar severity and responsiveness. If severity or responsiveness differs across patients in ways that can be identified precisely, the population of patients is then to be broken down into subgroups for which different levels of cost effectiveness are calculated. Full coverage is then provided to those subgroups with ratios below the benchmark, and no coverage to the rest (even if they would get some positive health benefits from the care).

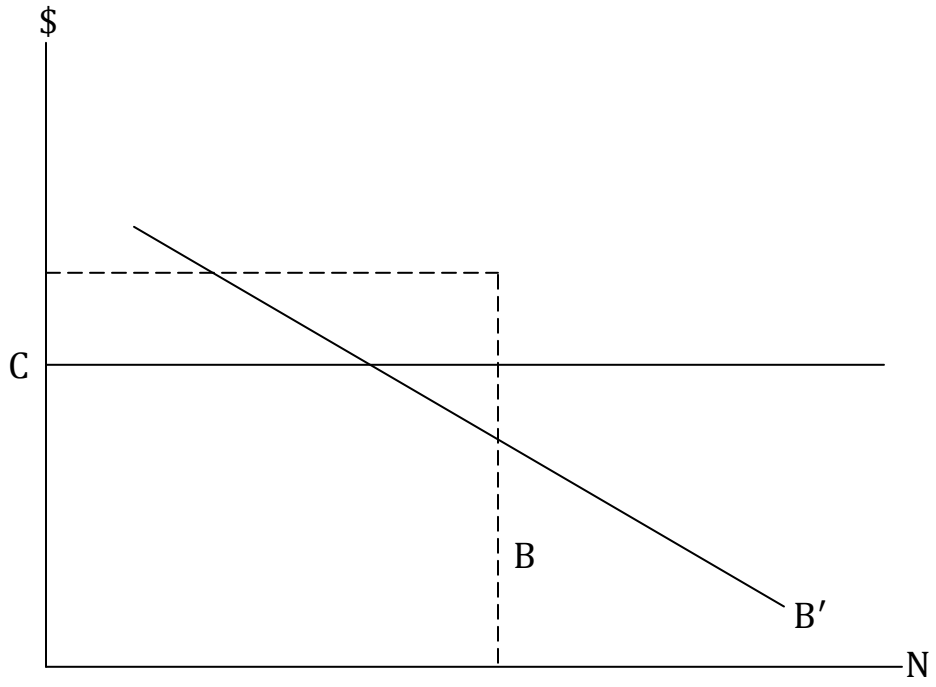
But the assumption that perfect segmentation is possible is often heroic. People may have different illness severities that they (or their physicians) can detect but insurers and planners cannot, or cannot specify criteria to target. Within an apparently similar patient population, there can then be a downward sloped demand curve as patients of low severity or marginal benefit are encouraged to use care (even if it is a covered service) only if cost sharing is low.

One immediate dividend from thinking about patient demand is a way to test whether an identified population is homogeneous in marginal health benefit from care. Take the simple case where the patient demand curve corresponds to the societal demand curve. If we have identified a population where the receipt of care is cost effective for every member, then they should all be willing to demand that care regardless of the level of coinsurance (including 100%). Obviously one key assumption here is that the value of increments in health are uniform across the population, but if we control for determinants of value like income and education and do find little response to cost sharing we can have more confidence that we have identified a clinically homogenous population than if we find high and varied responsiveness.

Heterogeneity in medical effectiveness and medical benefits.

Now suppose we consider the case of uniform monetary values per improvement in health, but do not assume that the insurer can distinguish among people who get different values of health benefits from the same treatment so as to be able to classify them into groups that are homogeneous with respect to the number of QALYs added by the treatment. That is, the insurer or social planner cannot accurately classify patients in advance as uniformly high benefit versus uniformly low benefit; they all look the same to the same when there is a claim. If patients and physicians are equally unable to make a priori distinctions, the case is one of homogeneous expected health benefits, and the analysis proceeds as above. But what if the patient (independently or with physician advice) can tell what the marginal health benefit is, or at least make a distinction between low and high benefits? This is the information asymmetry case discussed by Basu.

Figure 2



Let us begin with the simple case where marginal health benefit from the treatment is positively correlated with illness severity, and let us assume that the patient (and perhaps the patient's physician) knows severity, but the insurer does not. Let us also assume as before that demand is binary: each individual patient either demands and gets the service, or does not. Finally, let us assume that the distribution of illness severity each consumer would expect is identical and continuous, but realized severity will be known before use of the treatment is determined. This assumption, combined with heterogeneity of private patient values, is sufficient to give rise to a downward sloping aggregate marginal benefit curve for both society because marginal expected benefit declines as coinsurance (and severity) fall and more of the population is treated. But this can also be a patient demand curve for the service: given that consumers know their marginal health benefit (or illness severity) and some positive average consumer value for given increments in health that is independent of severity, at lower

coinsurance rates, as illustrated by the curve *B'* in Figure 2, people with lower levels of expected net benefit will seek treatment.

If all patients attach the same monetary value to all increments in health, this demand curve is a transformation of the social marginal benefit curve described by [$(\Delta QALY)^* W$]. Consumers could have diminishing marginal values for increments in health as a function of initial health state; they may attach lower values to additional units of health when they are healthier. Then the patient demand curve could have a steeper slope than the social demand curve. But in either case, the more people who use the service, the more who will be people of lower marginal benefit. And in either case, the shape of the curve depends on the shape of the frequency distribution of severities.

The striking implication then is that, even in the benchmark social perspective, both the average and marginal cost effectiveness of the intervention will depend on the level of cost sharing; so the answer to the question of whether *X* is cost effective (has average $/QALY less than *W*) is that "it all depends." It depends on the level of cost sharing, which in turn affects which patients with which levels of marginal benefit will actually use the service. Thus, logically, one cannot use prior information on aggregate or average cost effectiveness or value of this service to tell what the insurance coverage should be; coverage and (marginal) value are determined simultaneously.

Heterogeneity over values.

Another reason to deviate from the benchmark model is if the personal value V_i is positive, deemed to be relevant, and variable across persons. Once we allow for the possibility that citizens have private values for improvements in health—in the sense that they attach positive

values in terms of willingness to pay to their own health—we need to reconsider the social decision model. We have already observed the well-known proposition that the benchmark model implies that there is social value (relative to other uses of society's resources) of health in the amount *W*. The watershed question here then is whether private values V_i are included in *W* or additional to *W*. One possibility is to say that society has decided to subsume all private values into *W*—and, to the extent that they may vary, to ignore that variation as not relevant. The other simple possibility is to have *W* represent the values that citizens place on the health of members of society other than themselves (or their families); in this case the total value of a given improvement in health, other things equal, is $W + V_i$. (This approach was taken many years ago in Pauly, *Medical Care at Public Expense* [1971]. See also Culyer [1971].)

This possibility of a positive V_i causes few problems if the cost effectiveness ratio for a given program is everywhere below *W* even when only social values are considered; the program should be implemented, and zero cost sharing would be optimal. The main issue is that zero cost sharing not the only optimum in this case. If the cost sharing were set at a positive level below V_i and all beneficiaries for whom the treatment is to be given have positive values of *V* greater than that level, the optimal outcome would still be reached—all would use—but the distribution of costs across patients versus taxpayers would be altered.

But what is the right decision in the more relevant and interesting case in which the cost effectiveness ratio is above *W*? Based on social values alone the program is not efficient. But it is possible that $V_i + W$ may exceed *C*/QALY but V_i may also be less than C/QALY. The sum of social and private values would exceed W. Having the potential beneficiary of a program pay cost sharing that is positive but less than V_i would allow the program to be implemented if the social insurance were to pay *W*. But then we have a seeming paradox: the program is not worth

its cost to society, and is not worth its cost to the beneficiary. But the sum of these values does exceed its cost. Should it be implemented?

However, then *W* will not be uniform if some people have larger private benefits than others and those are supposed to be included in the social value. If the value *W* only includes the benefits to non-users, which could arguably be uniform, then it makes sense to allow private supplementation as long as the sum of private and social benefits exceeds marginal cost. The point here is that there has to be a decision.

If the program yields uniform medical benefits to people with different private values, one might imagine setting cost sharing at the difference between *W* and *C*. Then everyone for whom $W + V_i$ exceeds C would obtain the service. However, that outcome could also be achieved if full coverage was provided to all for whom the sum of private benefits exceeds cost; if desired, there would be a premium equal to the difference between *W* and *C*. Since there would be complete protection against risk, this solution with full coverage but differential premiums might be said to dominate the cost sharing solution.

Using coinsurance for cost-effective care.

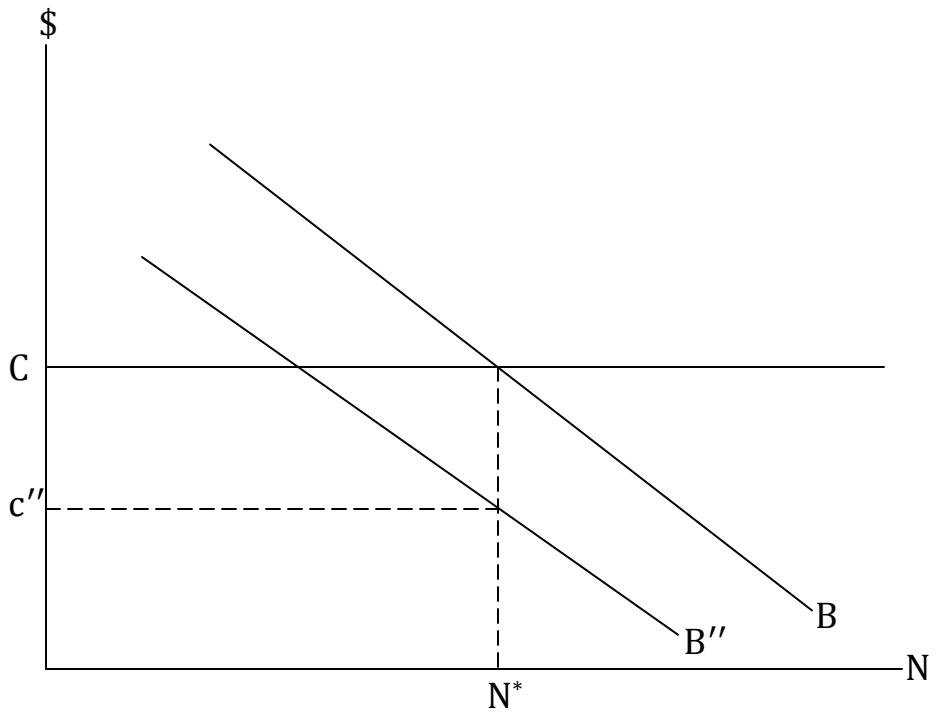
We now expand on the role of coinsurance in the case of these two kinds of heterogeneity. We consider first heterogeneity in marginal health benefit.

There may be a few services whose social marginal benefit to all eligible patients is always high enough to meet the benchmark; the social marginal benefit or demand curve would be always above the cost line as in Figure 1, either for the entire population or for the set of people who get positive marginal benefit. There can be other services whose benefit (though positive) is always lower than the benchmark, and so is the demand curve. Then the service is never cost

effective, and should not be covered at all. But the most general cases are those of services which provide benefits in excess of the benchmark for some and less for others. This is the classic problem of optimal coinsurance: coinsurance depends on the slope of the curve that plots variation in marginal benefit across people, that is, on the demand curve.

With the (diminishing) marginal health product curve and the patient demand curve as data, what then is the optimal value of coinsurance when only societal benefits matter—that is, when the monetary value attached to those marginal health benefits is just *$W* per QALY? The answer not surprisingly turns on the position and shape of the patient demand curve, and how it relates to the social demand curve.

Figure 3



Take first the benchmark (though unrealistic) case in which the patient demand curve coincides exactly with the social demand curve. The optimal rate of use of the procedure considering only societal benefits is N^* in Figure 3, where all users with benefits greater than

cost have been treated. It is easy to see that if the patient (market) demand curve coincides with this curve, they both cross the cost line at N^*, so the optimal level of cost sharing to lead to optimal use is 100%, and the optimal level of insurance coverage is zero. Indeed, even if the market demand curve were to have a different slope from the societal marginal benefit curve but crossed the cost line at the same point, the optimal level of coinsurance will still be 100%. (This assumes there are no societal benefits from risk protection per se.) The reason is that insurance coverage is not needed to assure use at the socially cost effective level.

The more relevant cases are those in which the two demand curves cross the cost line at different points. This can happen both if patients attach different values to health benefits than the (constant) value of *W* used by the planner and *or* if patients under- or over-estimate the marginal health benefit from care at any quantity.

One case is easy. If the patient demand curve crosses at a larger quantity than N^*, and insurance coverage cannot be negative, optimal coverage is still zero. (If a tax could be levied on people who use more care than the amount which is cost effective, a literal interpretation of this model is that there should be a tax.)

But suppose the consumer demand at zero coverage (full price) is less than the socially optimal amount. Then it is clear that, as long as the patient rate of use is when care is free is not below the optimal point, coinsurance should be positive, but less than 100 percent. This is shown in the case of demand curve *B''* in the diagram; the optimal coinsurance rate in this case should be something like *c"*.

We first provide some qualitative propositions about this more general case and then show a numerical example. We can characterize potential patient demand curves by their intercept on the y axis and by their slopes. Given slope, coinsurance should be lower the lower the intercept.

And given the intercept, coinsurance should be lower the steeper the slope. In effect, coinsurance induces patients to self-ration by severity, even though the insurer does not know severity.

The main point here is that the optimal coinsurance rate depends on the patient's demand curve as well as on the societal marginal benefit curve. Hence, the average ratio from a CE assessment (as in Braithwaite and Rosen) is neither necessary nor sufficient to determine what the coinsurance rate should be, even in a world in which private values and their variations are ignored as a normative consideration. Rather, it is the marginal CE ratio and its relationship to the coinsurance rate that matters.

Table 2 shows a hypothetical numerical example of the relationship between coinsurance and cost effectiveness (still using the single-value societal approach to the latter).

Table 2

Cost Sharing and Marginal Cost Effectiveness: Example

Coinsurance Proportion	Total Spending	Fair Premium	Total QALYs	Marginal QALYs	Marginal $/QALY	Average $/QALY
1	100,000	0	2.5	2.5	40,000	40,000
0.75	110,000	27,500	2.68182	0.18182	55,000	41,000
0.50	120,000	60,000	2.807	0.125	80,000	42,750
0.25	130,000	97,500	2.876	0.0625	160,000	45,300
0	140,000	140,000	2.907	0.0313	320,000	48,275

Table 2 also provides a numerical example to illustrate the relationship of cost effectiveness analysis and cost sharing. Consider a treatment with a unit market price of $1 per treatment. With no insurance coverage, 100,000 people seek treatment at a total cost of $100,000, and the aggregate benefit to this population is 2.5 QALYs, yielding a cost effectiveness ratio of $40,000

per QALY. Assume that coinsurance rates can only take on the values of 1.0, 0.75, 0.5, 0.25, or 0. The table shows that for each 0.25 reduction in user cost, 10,000 more people seek treatment; it also shows the incremental QALYs they receive. From this menu, if the societal threshold is $50,000, the optimal level of coinsurance is 100%, because reducing cost sharing to 75% would have a ratio of $55,555, above the threshold. In contrast, if the threshold were $100,000 the optimal level of coinsurance would be 0.5.

For each change in coinsurance, the marginal cost effectiveness is less favorable than the average cost effectiveness, as one would expect. For example, if coinsurance were 100%, implementing the program with that level of cost effectiveness would be better than doing nothing if the threshold were $50,000 or $100,000, but in each case net societal benefit is maximized at higher levels of cost sharing. If we interpret the schedule of quantities and marginal values in the example as representing the true marginal health product of the treatment, but if the market demand curve were lower (so that fewer people sought treatment at each coinsurance rate than shown), then optimal coinsurance at a social value of $50,000 per QALY might be lower.

Of course, if demanders in the market valued the health benefits at a use rate of 100,000 at a higher marginal value than $50,000, then the market demand curve at that point would be shifted to the right compared to what is shown in the table.

If the coinsurance is at the optimal level for some service, as described above, that is the ideal program to implement from a societal perspective if the service is to be provided at all. With constant marginal costs, it will also always be a cost effective program, whether the coinsurance rate is 100% or some fraction of the price and cost. This model can also be used to explore what happens if coinsurance is set at levels other than the optimal level. Consider the

case in which the social and private demand curves coincide; optimal coinsurance is 100% and the program is cost effective if there is positive demand at that price. But what if coinsurance is zero? Then the lower level of cost sharing encourages the use of care whose social value is less than its cost. This reduces the net social value of the program; if the welfare loss from moral hazard is large enough, the program can even cease to be cost effective.

If in contrast the private demand curve is to the left of the social demand curve, lowering coinsurance may increase the cost effectiveness of the program, up to the point where private demand equals the quantity at which the social demand curve crosses the price line. Thus in this case it can be cost effective to lower cost sharing—until the optimal level of coinsurance is reached.

In many cases cost effectiveness analyses are done for new products that will involve substantial amounts of lump sum expenditure for research and development in addition to the production and distribution cost of the product once it is on the market. In this case the model tells us to determine the coinsurance rate which is optimal based on marginal cost effectiveness, and then calculate the overall cost effectiveness from the treatment given the rate of use associated with that coinsurance rate, and compare it to the social benchmark. This is a simple but powerful result. It tells us both whether to invest in the product or treatment and what kind of insurance coverage it should receive.

In the illustrative example, the net benefit of the program considering only costs of production, with coinsurance at zero, is $5350. This is the difference between the 2.907 QALYs added valued at $50,000 each, or $145,350, and the production costs of $140,000. If the R&D costs were $20,000, this program would not represent good value. But if coinsurance were

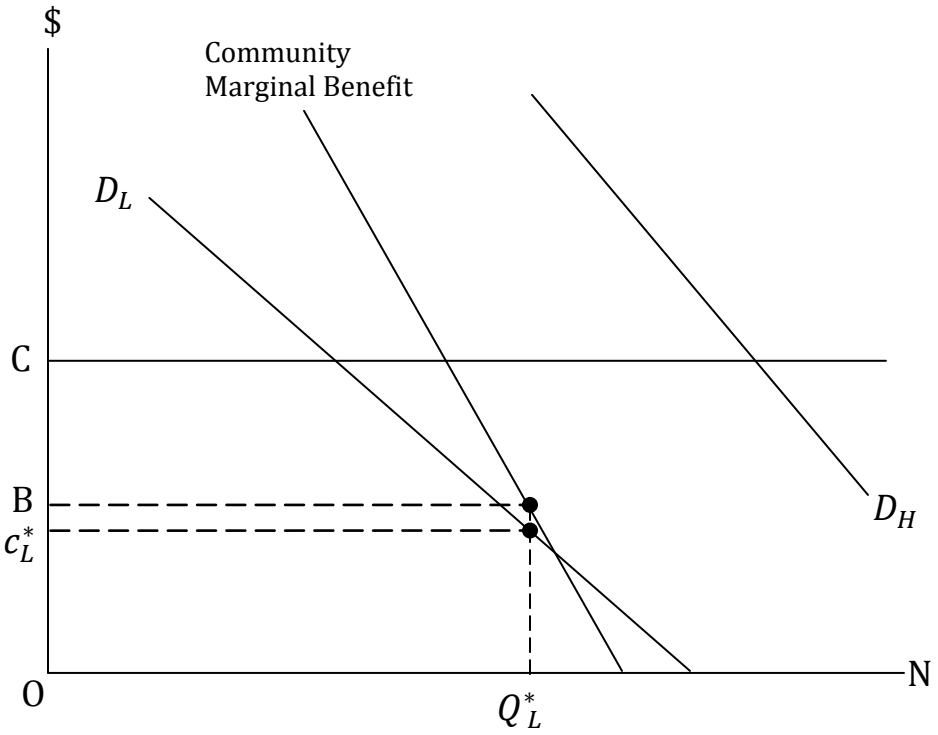
increased close to unity, the net benefit would be $25,000 ([2.5 x 50,000]-[100,000]), enough to make the program socially efficient.

Adding value to private benefits.

Now let us return to the case where there are (varying) private values for measures of benefit such as quality adjusted life years. What if these private benefits (in the sense of benefits to users that differ from social benefits) that are deemed appropriate to take into account? If private willingness to pay or value for one's own QALYs are considered as part of the full social value, but adds to that potential positive values that others than the direct user in society may place on additional health for someone, given some initial level of health, then the total marginal value of improvements in health for people with different realized levels of severity should include both.

We imagine there are both private marginal benefit curves (ICER multiplied by each individual's value of health added by the treatment) and "value-to-others" or community marginal benefit curves (ICER multiplied by the sum of values of all others). Here we ignore insurance (risk reduction) benefits which are private benefits to any risk averse person and could have some element of social benefit as well. Figure 4 shows a community marginal benefit curve; assume that the same curve applies to every person. However, personal marginal benefit curves will differ when people attach different values to incremental health. As drawn here, the social demand curve has a negative slope and presumably goes to zero at a smaller quantity or proportion than does the private benefit curve. We distinguish two (identifiable) classes of consumers that attach different values to health (based either on different incomes or tastes), their marginal benefit curves are shown as D_L and D_H.

Figure 4



Given the private and social values of increments to health described by these curves, we can solve for the optimal proportion of each type of individual that should be treated by adding the demand curves vertically and finding the point of intersection of the summed demand curve (not shown) with the price line. For the person with marginal benefit D_H we note that even at no insurance coverage the marginal social benefit is zero; here the standard individual determination of optimal coinsurance would apply. For the person with private demand D_L there are positive marginal social benefits at no insurance coverage but zero such benefits at full coverage. The optimal level of coverage ignoring risk reduction benefits is the level of coinsurance that yields the quantity treated $Q^*_{L.}$ Here the person's contribution reflects his marginal benefit, and the community's contribution reflects its marginal benefit, and the difference between the person's marginal benefit (Oc^*_L) and marginal cost just equals the community's marginal benefit (OB). Alternatively, the excess of actual cost in dollars per QALY over the societal value W is just

offset by the private value. The implication then is that optimal coinsurance from a social point of view can be lower than privately optimal coinsurance if there are positive marginal social benefits at the level of use associated with the privately optimal level of coinsurance. Particularly for lower income populations or treatments of diseases primarily affecting lower income populations, this marginal social benefit may be positive.

So far we have not considered private risk spreading (insurance) benefits to the person. As is well known, such benefits can cause the individual to prefer insurance that tolerates the use of care that is not itself cost-effective but which is promoted by the lower coinsurance that provides risk protection. Adding these benefits adds another term to the value of additional care associated with lower cost sharing but the risk of the analysis proceeds as before. It may well be that the coinsurance rates some individuals would choose would be higher than the rate which leads to the socially optimal use of care; then these should be subsidies. But if the privately chosen coinsurance rate would be lower, there is no need for subsidy and that would be the socially optimal rate.

Optimal cost sharing from a combined perspective: toward a more useful approach.

Developing information on costs and effects of medical treatments for purposes of offering advice on cost sharing in insurance then requires a rather different model from any of the orthodox binary (either cost effective or not) normative models so common in the literature. Along the way toward specifying such a model, two potential alterations to the orthodox of model of optimal coinsurance may also be required. That model assumes that consumer/patients are well informed about the marginal benefit from care, and that only private values for improvements in health matter. When these assumptions do not hold, the optimal coinsurance

rate (from a societal perspective) will change, and the policy that carries that optimal rate may not be demanded voluntarily by unsubsidized consumers.

The changes in optimal coinsurance associated with consistent under or overestimation of private marginal benefit have been described by Pauly and Blavin (2008). In the case of underestimation, ideal coinsurance will be lower than would prevail under accurate information, and higher for overestimation. Information on benefits and costs is needed to determine the ideal rate. If there is also social value (value to non-users) attached to some medical service, either because of contagious disease or altruism, the relevant marginal benefit curve includes the marginal benefits to nonusers as well as to users; given this new MB curve, one can solve for optimal coinsurance as before. The main difference in the both the underestimation of private demand and presence of social demand cases is that the consumer may need a subsidy to get him to purchase the socially optimal policy. Here again information on benefits and costs at each level of cost sharing is needed, but there is no single cost effectiveness calculation that is useful.

Instead, the ideal analysis would analyze costs, effects on health, *and* risk reduction benefits at every level of coinsurance for some medical service, determine the level of coinsurance at which the sum of marginal risk reduction benefits plus health improvements just equals the cost, calculate the net benefits at that ideal point, and determine whether they are positive or negative. If they are positive, the insurance and care program should be implemented; if they are still negative, it should not.

Incorporating insurance administrative costs and price responsiveness.

Some of the discussion of the use of value based cost sharing expresses disappointment when lowering cost sharing to zero produces only small impacts on the rate of use of undervalued

services. Assuming that such services have cost effectiveness ratios above the threshold; does such disappointment correspond to negligible impacts on welfare? The answer is: not necessarily.

The reason is that even if lowering cost sharing has only a small impact on use, it provides financial benefits to those who are insured. The welfare evaluation depends on whether those financial/income redistribution benefits are to be counted and, if they are, the size of those benefits and as well as on the marginal benefits from use. Coverage may increase insurance premiums, but insureds will get all the money back. The answer is different if there is (as there always is) an administrative cost to covering the service with insurance. Then the answer depends on the administrative cost of insurance and the degree of risk attached to the services being covered. In some examples the "insurer" was the Veterans Administration which may or may not be concerned with financial benefits from lower cost sharing to veterans above and beyond health benefits. If in contrast we are looking at private insurance for which risk averse individuals pay the premium, lowering coinsurance raises premiums but, even if use is only slightly affected, all purchasers may expect to get higher insurance benefits to offset the higher premiums. The downside to an "ineffective" reduction in cost sharing is the higher insurer administrative cost as in Held and Pauly. But if that cost is not too high relative to the risk reduction associated with lower cost sharing, doing so may well improve social welfare even if it does not much improve health.

However, this case is clearly well beyond the standard benchmark model of cost effectiveness in health that uses dollars per QALY and a societal threshold. Perhaps if the (expected) quality of life also incorporated considerations of risk reduction the stories could be

made consistent, but standard measures of quality of life do not take this consideration into account unless lower risk affects mental health.

Take the extreme case in which a service was formerly not covered at all (coinsurance proportion equals unity), was used by a relatively large share of the population and paid out of pocket , and has a zero cost offset, and has a positive demand that is very unresponsive to price. But suppose use only increases by one unit when coverage is provided that reduces the user price of a moderately priced service to zero for the population of insureds. Then one must trade off the incremental administrative costs of covering the formerly self-paid services for all those insured, against the benefit from that service: the cost offset and the health benefits for that one unit, plus any risk reduction benefits; the change might still well be negative. If, in contrast, use expanded substantially, welfare would rise if the incremental health benefits to those brought in, plus the incremental risk premium for coverage for all users, exceeded the incremental administrative costs for insurance coverage.

But this preference for a larger response is only true in the range where incremental use has already been determined to be cost effective (little or no moral hazard, or marginal benefit equals net (of offsets) marginal cost. If the cost effectiveness can cover the wide range indicated by a continuous declining marginal benefit curve, a larger price responsiveness is generally associated with a lower ideal cost sharing level from the perspective of the individual consumer.

Coverage of a new service.

While the theory of optimal coinsurance solves the conceptual problem for existing services, and while the research just described would solve the empirical problem, what about new services for which no data yet exists on how use would respond to coinsurance? The ideal

analysis describes the benefits and the net cost of the new service at various rates of use by different populations corresponding to the different levels of cost sharing. If there is positive demand for the service at full price (no insurance), it should receive at least some coverage. More generally, given a schedule of marginal benefits, costs, and risk premia, an optimal coinsurance rate can be found.

The problem is that it is very difficult to infer what this schedule would look like from typical clinical trial data. That is because the schedule of marginal benefits largely plots the response of sequentially *different* (in terms of initial illness severity or likely response to treatment) patient populations, whereas a well-designed trial will work as hard as it can to get a set of patients who are similar in observed illness severity and observed characteristics that might be related to response to treatment. At best the trial might help to plot the marginal benefit from additional treatment for a given population (more frequent screening or dosing, or varying doses).

If we are trying to value ex ante a potentially innovative medical product, it is often the case that there are research and development costs that must be incurred before it can be supplied. Then marginal cost is no longer constant. In such a case, Lackdawalla and Sood have shown that it may be efficient to set coinsurance at the marginal cost of production and distribution. This will (ignoring insurance benefits) maximize net benefits given that the product is produced. But evaluation of the overall program requires that aggregate benefits in excess of the coinsurance rate be larger than the cost of R&D. That is, not only must the coinsurance rate be set to induce the socially (or socially and privately) cost effective quantity of the product if it is put on the market, but it also has to be cost effective to innovate it at all. A two-part test is needed.

What might help.

Even if we cannot get things exactly right in linking cost effectiveness to cost sharing using the standard models, there are some thoughts that might help. The first point to note is that if a CE ratio incorporating both individual and social values finds that an intervention provided to a defined population facing a given level of coinsurance (e.g., 20%) does not meet the cost effectiveness test, that intervention should not be insured at that level of coinsurance. But it might still efficiently be covered at higher levels of coinsurance for the same population.

The more analytic approach would be to begin with an intervention that meets the test (that is, average CE is below the threshold), and then reduce the threshold, calculating with each decrement the additional cost and the additional health outcomes. If consumers are fully informed at the initial level of coinsurance was less than one, we should find that reducing coinsurance does not provide more private benefit than cost, but imperfect information or the presence of external benefit may cause the ratio to be low enough. At a coinsurance rate where the net benefit turns negative, one needs a measure of the benefit from risk reduction to tell how much further to go.

Conclusion.

These are somewhat discouraging conclusions. They definitely imply that there is no simple but correct way to move from findings of a typical cost effectiveness study to saying what the coinsurance rate should be for a non-poor population. The most one could hope for would be a binary decision of whether or not a particular treatment should or should not be covered by insurance with a particular predetermined cost sharing rate (usually but not necessarily zero). They also imply that the cost effectiveness of a treatment cannot be properly determined unless

coinsurance is set at the optimal level. So, unless there is perfect information to identify heterogeneity of benefits, considerations of consumer demand (in the classic economic sense of the shape of the demand curve) need to be added, regardless of the normative model.

The fundamental problem is the assumption of a uniform benefit of uniform value which is central to the societal cost effectiveness model. This assumption is presumably made for administrative and expository reasons, not because anyone believes that marginal health benefits are uniform, or that the marginal value of a health benefit (to a consumer or society) is independent of the current level of health or allocation of resources. It was hard enough to get policymakers to accept the need for considering costs and the need to establish a money value for health outcomes, however arbitrary, and in the United States neither of these concepts is as yet effective. But paradoxically it might be more feasible to get political acceptance if more attention to reasonable variation in effectiveness and value were explicit rather than suppressed in the analysis. As always, there is a case to prefer approximating the perfect rather than precisely hitting the imperfect as a method of policy analysis.

References.

Basu, A. 2011. Economics of individualization in comparative effectiveness research and a basis for a patient-centered healthcare. *Journal of Health Economics* 30(3): 549-559.

Braithwaite, R.S., and A.B. Rosen. 2007. Linking cost sharing to value: An unrivaled yet unrealized public health opportunity. *Annals of Internal Medicine* 146(8): 602-605.

Culyer, A. 1971. The nature of the commodity "healthcare" and its efficient allocation. *Oxford Economic Papers* 23(2): 189-211.

Held, P.J., and M.V. Pauly. 1990. Benign moral hazard and the cost-effectiveness analysis of insurance coverage. *Journal of Health Economics* 9(4): 447-461.

Ioannidis, J.P., and A.M. Garber. 2011. Individualized cost-effectiveness analysis. *PLoS Medicine* 8(7): e1001058. doi:10.1371/journal.pmed.1001058.

Pauly, M.V. 1971. *Medical Care at Public Expense: A Study in Applied Welfare Economics.* New York: Praeger Publishers, Inc.

Pauly, M.V., and F.E. Blavin. 2008. Moral hazard in insurance, value-based cost sharing, and the benefits of blissful ignorance. *Journal of Health Economics* 27(6): 1407-1417.