Volume 9 • Number 4 • 2006
VALUE IN HEALTH

Indirect Costs and Cost-Effectiveness Analysis

Richard Ernst, PhD

Los Angeles, CA, USA

ABSTRACT

This article examines the methodological implications of the societal perspective in cost-effectiveness analyses in which the costs of health-care interventions are defined as the sums of direct and indirect costs. In the model of cost-effectiveness analysis in which the planner distributes patients among many treatments for many illnesses, the definition requires that total indirect costs be constrained, and the article proposes an iterative computational procedure for choosing a constraint under the assumption that the planner maintains a target trade-off rate between losses of health benefits and reductions in indirect costs. In the more common model in which the planner decides which of two treatments for one illness to provide to patients, the adoption of a societal perspective introduces ambiguities into the welfare properties of the decision rule, and in general the conclusion that a treatment is cost-effective is valid only if switching or assigning patients to the cost-effective treatment does not increase the planner's total direct costs.

Keywords: cost-effectiveness analysis, costs, economics, methodology.

Indirect Costs and Cost-Effectiveness Analysis

The standard models of cost-effectiveness analysis (CEA) posit a public or private planner of some kind that commands, or behaves as if it commands, fixed supplies of health-care resources with which it cares for a community having certain health problems. In Weinstein's well-known dictum, "[t]he underlying premise of cost-effectiveness analysis in health problems is that, for any given level of resources available, the decision-maker wishes to maximize the aggregate health benefits conferred to its population of concern" [1]. Health benefits are commonly regarded as utilities associated with states of health, so that the "aggregate health benefits" conferred on the planner's community is a health-related (Benthamite) social welfare function—the sum of health-related utilities over all individuals in the community. Thus, the premise that the planner wishes to maximize this function subject to a resource constraint is a behavioral axiom with roots in economic welfare theory (e.g., [2]). In fact, suggested further on, Weinstein's premise does not quite fit all of the current models of CEA, but in each of them the planner assigns patients to treatments for their health problems in a way that either maximizes its community's total health benefits subject to a cost constraint, or else improves its community's overall well-being without necessarily maximizing it.

Address correspondence to: Richard Ernst, 1712 S. Bedford Street, Los Angeles, CA 90035, USA. E-mail: Rilernst@aol.com

10.1111/j.1524-4733.2006.00114.x

Although there seem to be no substantive definitions of what constitutes "directness" and "indirectness" in the cost-related literature, the costs of interventions in health-care problems are commonly divided into two types, direct and indirect. It is generally agreed that direct costs are the opportunity costs of formal health-care goods and services such as hospital, physician, and nursing home care, drugs, and so on, but there is less than full agreement about the nature of the events and resource usage whose opportunity costs can be called indirect. The imputed value of foregone labor product when patients' labor services become inefficient or are withdrawn from production on account of morbidity or premature mortality is widely regarded as an indirect cost. But while most authors classify the imputed value of unpaid caregivers' labor (informal care, community care, or unpaid caregivers' time) as an indirect cost (e.g., [3–6]), a few prefer to label it a direct cost (e.g., [7]). Patients' "time costs"—that is, the imputed money value of unpaid household labor services and the purported money value of the nonproductive leisure time that patients give up during treatment and convalescence—have also been defined variously as direct and indirect costs [5,8], although the claim that the utility of forgone leisure activity can be monetized has not been generally endorsed [9], and is almost certainly tenuous.

For example, there are least three different problems with the notion of patients' time costs. First, the claim that the utility of nonproductive leisure activities has a money value contradicts the assumptions of standard microeconomic theory. Even if one pays for a pleasant recreational activity, the payment does not

 1098-3015/06/253 253–261

measure a quantity of utility. Second, if nonhealth-related utility has a money value, it is reasonable to ask why health-related utility does not also have a money value, and that possibility is expressly precluded in CEA. Third, the proponents of time costs price all hours of leisure time forgone at the market wage rate (e.g., [7]), and this in turn implies that the individual would supply any number of hours of labor services up to the maximum possible number at the market rate—that is, it implies that the individual and market supply functions of labor are infinitely wage-elastic at the going market rate.

Despite the grayish area that separates direct and indirect costs, this article begins by proposing one characteristic that distinguishes the two types of costs. In particular, the productive resources whose costs are ordinarily classified as direct are subject to observable, well-defined capacity constraints, but the resources whose costs are ordinarily classified as indirect are not. For example, the volume of formal health care that can be produced for a community is constrained either by the planner's budget or by the physical quantities of formal medical goods and services available for treating that community, but it makes no sense to think that a similar capacity constraint can be defined for the quantities of goods and services lost to the community when workers become sick or die and are unable to work. Increasing the community's supply of formal health-care goods and services makes it possible to increase health-benefit production. Increasing the quantity of other goods and services forgone because of illness and death does not.

Because the community's supply of informal care is obviously not unbounded from above, at first glance it seems that the cost of unpaid caregiving labor does not fit the classification proposed, but the point is not that the quantity of this labor is limitless. Rather it is that in virtually all real-world circumstances no meaningful upper bound on the community's supply of unpaid caregiving labor can be defined or established, and therefore that the assumption of such a bound is not sustainable in models of health-care planning. Consider first that the maximum quantity of unpaid caregiving labor is never directly observable—especially to a cost-effectiveness analyst—and in general it can be estimated only from survey data the reliability of which is always disputable. Second, even if reliable estimates of the maximum could be obtained, the care provided by families is not usually transferable to members of other families. That is, if a family is willing to provide up to h hours per week of caregiving services to its own members, but only up to h' hours to a second family, up to h'' hours to a third family, and so on, there exists no single maximum number of hours of unpaid care that the family or community is willing to supply under all circumstances unless $h = h' = h'' = \ldots$ And finally, even if all unpaid care were transferable across families (and voluntary organizations) it is by its nature unresponsive to reimbursement incentives and the direct commands of health-care planners. As a consequence, a planner has no means of inducing or forcing the community to provide unpaid care up to the nominal maximum quantity, and, if it is only accidentally reachable by the planner's actions, the nominal maximum is irrelevant for health-care planning. These arguments seem persuasive, but if there are empirical conditions in which a well-defined, observable capacity constraint on the supply of unpaid caregiving labor can be shown, there are no compelling reasons to segregate unpaid labor from formal health-care inputs or to regard its cost as indirect.

As these remarks imply, indirect costs are the costs of illness-related resource losses and usage that are borne by the community at large, and health economists have long regarded them as proper costs of illnesses (e.g., [10]). In the literature on CEA, it has also become conventional to argue that indirect costs ought to be recognized by health-care planners in choosing treatments to be offered to or sanctioned for patients (e.g., [7,11,12]), and a planner who does account for these costs is said to hold or maintain a societal perspective. The standard means of incorporating indirect costs into CEA is to define total illness costs as the sum of direct and indirect costs (e.g., [13]), and this definitional mechanism presumably captures the consequences of the planner's decisions on total illness-related resource usage and resource losses over the community as a whole. In effect, then, the planner that maintains a societal perspective has the double obligation of caring for patients and of controlling the economic impact of illnesses on the community at large.

Although the appropriateness of imposing this twofold obligation on a planner can, perhaps, be challenged, in this article it is taken as given. Instead, it is argued that introducing indirect costs into one of the two basic models of CEA requires the planner to set a constraint on these costs, that in general the constraint forces a trade-off between indirect costs and the volume of the community's total health benefits, and that a trade-off rate between indirect costs and health benefits having reasonable welfare implications can be built into the model. But it is argued that introducing indirect costs into the second basic model does not necessarily yield meaningful welfare conclusions and that, in fact, it tends to reduce the generality of judgments that can be made about the cost-effectiveness of interventions.

Hereafter, any kind of health problem and illness and any kind of intervention in the problem will be called a treatment. For simplicity, patients' health well-being (or utility or welfare) is taken to be measurable as numbers of quality-adjusted life-years (QALYs), although other meaningful quantifiers would do as well. It is assumed that the indirect costs and QALY

productivity of treatments are not inversely related because otherwise the argument for entering indirect costs into CEAs becomes less persuasive and may be invalid altogether. If, for instance, the QALY productivity and indirect costs of treatments are inversely related, the distribution of patients among treatments that maximizes the health-care system's total QALY output, given its total direct costs, may also give a constrained minimum of total indirect costs, and in that event a QALY-maximizing planner would choose the same distribution of patients whether or not it recognized indirect costs.

Last, as suggested, two fundamental kinds of models have been employed in CEAs. In the first, here called the many-illness, many-treatment model, there are many illnesses to be cared for, possibly many alternative treatments for each illness, and the planner's mission is to assign all patients who have each illness to one or more treatments or else not to treat them. The second model, here called the one-illness, two-treatment model, is the familiar set of assumptions and decision rule in which the planner's task is to assign patients to one or both of two treatments for only one illness. The second model generalizes to any number of treatments for a single illness; but because it makes no difference in this discussion, the number of treatment alternatives to two is restricted. In each model, a cost-effective treatment is one to which patients are assigned according to the decision rules.

The Many-Illness, Many-Treatment Model

Although it has rarely been used in practice—one version of it was employed in the design of the Oregon Medicaid program during the late 1980s [14–16]—the many-illness, many-treatment model is usually characterized as the "league-table" or "prioritizing" form of CEA. In it the planner assigns patients to treatments for all illnesses to maximize the community's total output of QALYs subject to a total cost or budget constraint [17–19]. Hence the model exactly fits Weinstein's principle of CEA. Nevertheless, it is now generally understood that the same model can be formulated as a simple problem in linear or integer programming [20–22]. So suppose there are I illnesses in the planner's community and J_i treatments for the i-th illness. Assume that the planner can spend no more than the money amount C on the total number of treatments for all illnesses. Let c_{ij} be the cost per patient of the j-th treatment for the i-th illness, and let q_{ij} be the number of QALYs per patient produced by the ij-th treatment. Let N_i be the number of patients having the i-th illness, and let n_{ij} be the number of patients—which for simplicity is assumed hereafter to be continuously variable—to whom the planner provides the ij-th treatment. There are variations in the way that the programming formulation of the CEA can be stated, but essentially it is this: choose the n_{ij}, $j = 1, 2, \ldots, J_i$, $i = 1, 2, \ldots, I$, to maximize $\sum_{i=1}^{I}\sum_{j=1}^{J_i} q_{ij}n_{ij} = Q$ subject to I epidemiological constraints and a cost constraint: $\sum_{j=1}^{J_i} n_{ij} \le N_i$, $i = 1, 2, \ldots, I$ and $\sum_{i=1}^{I}\sum_{j=1}^{J_i} q_{ij}n_{ij} \le C$. A solution of the problem is consequently a distribution of patients among treatments that yields a cost-constrained maximum total QALY output for the planner's community.

The CEA is plainly sensible if C and the unit costs c_{ij} are defined as direct costs. But suppose instead that C is redefined as the sum of total direct and total indirect costs and the c_{ij} are redefined as the sums of direct and indirect costs per treatment. Then first of all, the program must be revised to accommodate the resource constraint on total direct costs because the planner can never provide more formal treatment care than its resource base or budget allows. But indirect costs cannot be entered into the objective function without changing the nature of the CEA model, and if total indirect costs are not constrained, the solution of the program is the same whether or not the planner recognizes indirect costs. Hence the assumption of a societal perspective will affect the assignment of patients to treatments only if total indirect costs are constrained, and this can be done either by setting a single constraint on total indirect costs or by setting separate constraints on the components of these costs such as the costs of lost productivity, informal care, etc.

Because it is the more obvious of the two alternatives, consider placing a constraint on total indirect costs. Then the constraint must be binding because otherwise it has no effect on the solution of the program when only total direct costs are constrained. But if the constraint is binding, it constricts the space of admissible n_{ij} defined by the program when only total direct costs are constrained, so that in general a binding constraint on total indirect costs forces a smaller maximal Q than the planner could obtain by constraining total direct costs alone. In short, as soon as indirect costs are introduced into the model, the planner faces a conflict between producing QALYs and constraining total indirect costs, and except fortuitously the planner that adopts a societal perspective is unlikely to produce as large a total output of QALYs as the planner that does not.

Because there is no self-evident upper bound on total indirect costs, the planner is free to set a constraint on these costs according to any principle it chooses. But the rational planner will know in advance of choosing the constraint that the more restrictive it is, the larger is the total number of QALYs that its

patients will be compelled to forego. Thus, the rational planner will not set the constraint arbitrarily, but will select it only after weighing the societal worth of losing (gaining) QALYs in exchange for reducing (increasing) the community's total indirect costs. That is, the rational planner will (or should) choose a socially acceptable marginal trade-off rate between QALYs and direct costs—the largest number of QALYs it will give up for a given unit reduction in total indirect costs—and set the upper bound on total indirect costs to achieve that rate. Otherwise, constraining indirect costs satisfies no particular welfare objective at all.

Although it requires some computational effort, such a socially acceptable upper bound on total indirect costs can, or can approximately, be obtained by iteratively solving the dual of the primal linear program revised to incorporate a constraint on total indirect costs. The revised primal program is: choose the $n_{ij} \geq 0$ to maximize $\sum_{i=1}^{I}\sum_{j=1}^{J_i} q_{ij}n_{ij} \leq Q$ subject to the $N+2$ constraints $\sum_{j=1}^{J_i} n_{ij} \leq N_i$, $i = 1, 2, \ldots, I$; $\sum_{i=1}^{I}\sum_{j=1}^{J_i} c_{ij}^{D} n_{ij} \leq C^{D}$; $\sum_{i=1}^{I}\sum_{j=1}^{J_i} c_{ij}^{IN} n_{ij} \leq C^{IN}$; where the superscripts D and IN denote direct and indirect costs, respectively. The dual program is: choose the $I+2$ variables $\pi_i \geq 0$, $i = 1, 2, \ldots, I$, $\pi^{D} \geq 0$, $\pi^{IN} \geq 0$ that minimize

$$\sum_{i=1}^{I} N_i\pi_i + C^{D}\pi^{D} + C^{IN}\pi^{IN} = V$$

subject to the $\sum_{i=1}^{I} J_i$ constraints

$$\pi_i + c_{ij}^{D}\pi^{D} + c_{ij}^{IN}\pi^{IN} \geq q_{ij},\ j = 1,2, \ldots, J_i;\ i = 1,2, \ldots, I$$

The π_i are the shadow prices of QALYs with respect to the patient population sizes N_i, and $\partial(\max Q)/\partial C^{D} \equiv \pi^{D}$ and $\partial(\max Q)/\partial C^{IN} \equiv \pi^{IN}$ are the shadow prices of QALYs with respect to total direct costs and total indirect costs (e.g., [23,24]). The maximum QALY output $\max Q$ is a concave, piecewise linear function of the upper bound C^{IN} on total indirect costs (e.g., [23,25]), and on account of the concavity and piecewise linearity of $\max Q$ in C^{IN}, $\partial(\max Q)/\partial C^{IN} \equiv \pi^{IN}$ is a nonincreasing step function of C^{IN}. Notice that $\partial(\max Q)/\partial C^{IN}$ is the reciprocal of the implicit marginal money price (= marginal cost) of QALYs at the optimum. More to the point, though, it is precisely the marginal trade-off rate between QALYs and total indirect costs at the optimum. That is, given C^{IN} and assuming that the assignment of patients to treatments remains optimal, π^{IN} is the rate of loss (gain) in total QALY output per marginal reduction (increase) in C^{IN}.

Assume then that the planner preselects a socially acceptable marginal trade-off rate between QALYs and indirect costs, say $\hat{\pi}^{IN}$, where by "socially acceptable" is meant a rate deemed acceptable by the planner's community and determined by a community vote, community survey, or by some other method that is irrelevant here. But having chosen $\hat{\pi}^{IN}$, the planner's charge is to select a C^{IN} that yields $\hat{\pi}^{IN}$. Obviously, the choice of C^{IN} cannot be made a priori, but it can be made, or made approximately, by iteratively solving the dual program. An algorithm for selecting C^{IN} is as follows if there exists a solution such that $\pi^{IN} = \hat{\pi}^{IN}$. (If no such solution exists, the planner has no way of assigning patients to treatments that achieves its welfare objectives.) First, since π^{IN} is an everywhere nonincreasing function of C^{IN}, with some care large and small values of C^{IN} can be chosen so that solutions of the dual program give values of π^{IN} that bracket $\hat{\pi}^{IN}$. Let $C^{IN,1}$ be the smaller of the two initial upper bounds on total indirect costs and $C^{IN,2}$ be the larger of the two. Then solutions of the dual program yield the shadow prices $\pi^{IN,1}$ and $\pi^{IN,2}$, and $\pi^{IN,1} \geq \hat{\pi}^{IN} \geq \pi^{IN,2}$. If either of these last two inequalities is an equality, the algorithm is terminated. Otherwise, the program is reformulated and re-solved first with a slightly larger value of C^{IN} than $C^{IN,1}$, and next with a slightly smaller value of C^{IN} than $C^{IN,2}$. Denote the two new shadow prices by $\pi^{IN,11}$ and $\pi^{IN,22}$, respectively. Accordingly, $\pi^{IN,1} > \pi^{IN,11} = \hat{\pi}^{IN}$; or $\hat{\pi}^{IN} = \pi^{IN,22} < \pi^{IN,2}$; or $\pi^{IN,11} > \hat{\pi}^{IN} > \pi^{IN,22}$. Thus, the bracketing algorithm continues until a trial value of C^{IN} yields $\pi^{IN} = \hat{\pi}^{IN}$ or until $|\pi^{IN} - \hat{\pi}^{IN}| < \delta$, where $\delta > 0$ is predetermined by the planner.

The algorithm is laborious but workable. It has been mentioned the planner may wish to set separate constraints on the components of indirect costs, but the algorithm cannot be used with multiple cost constraints because the shadow prices of QALYs with respect to the upper bounds on the cost components are functions of the upper bounds on all of the cost components. In general, changing one of the bounds changes all of the shadow prices, and there may not even exist a set of upper bounds on the cost components that yields the planner's target shadow prices. This suggests that the planning program should be formulated with only two cost constraints, one on total direct costs and one on total indirect costs.

The One-Illness, Two-Treatment Model

In what has been called the one-illness, two-treatment model—it is also known as the cutoff-point or threshold form of CEA—the planner must decide between two medically substitutable treatments for illness *i*, *ij*, and *ik*, to provide to or sanction for patients. The model is commonly used in economic evaluations of

new health-care interventions and it can be described as follows. Assume first that the planner does not maintain a societal perspective and does not recognize indirect costs. Its community therefore consists only of patients. Let the direct costs per treatment of *ij* and *ik* be denoted by c_{ij}^{D} and c_{ik}^{D}, and let the QALYs per treatment be denoted by q_{ij} and q_{ik}. Assume further that $c_{ij}^{\mathrm{D}} > c_{ik}^{\mathrm{D}}$ and $q_{ij} > q_{ik}$, because otherwise one treatment dominates the other and it is never efficient to provide a dominated treatment to patients. Next, define the incremental direct-cost-effectiveness (ICE) ratio on any pair of treatments *a* and *b*, where $c_a^{\mathrm{D}} > c_b^{\mathrm{D}} c_b$ and $q_a > q_b$, as $(c_a^{\mathrm{D}} - c_b^{\mathrm{D}})/(q_a - q_b)$, and consider all such ICE ratios in the health-care system in which treatment *a* is currently provided to or sanctioned for patients and *b* is the next most costly, next most QALY-productive treatment for the same illness. Let λ^{D} denote the largest of all of these ICE ratios, and call λ^{D} the cutoff or threshold point.

The decision rule in the one-illness, two-treatment model then states that treatment *ij* is cost-effective, meaning that one or more patients having *i* should be assigned to it rather than to treatment *ik*, if and only if the ICE ratio

$$\frac{c_{ij}^{\mathrm{D}} - c_{ik}^{\mathrm{D}}}{q_{ij} - q_{ik}} < \lambda^{\mathrm{D}} \tag{1}$$

Although the decision rule has been known for some time (e.g., [26–29]), its application does not necessarily give a cost- or resource-constrained system-wide maximum of QALYs, and for that reason its usefulness as a planning tool and its place in CEA have been disputed [30,31]. Nevertheless, it is readily demonstrated that—at least when indirect costs are ignored—the rule is a device for testing whether there are social welfare advantages in assigning patients to treatment *ij* rather than to treatment *ik*. An argument to that effect is as follows.

First of all, if (1) holds there exists at least one pair of treatments for another illness *h*, say *ht*(*h*) and *hs*(*h*) such that

$$\frac{c_{ij}^{\mathrm{D}} - c_{ik}^{\mathrm{D}}}{q_{ij} - q_{ik}} < \frac{c_{ht(h)}^{\mathrm{D}} - c_{hs(h)}^{\mathrm{D}}}{q_{ht(h)} - q_{hs(h)}} \le \lambda^{\mathrm{D}} \tag{2}$$

where *ht*(*h*) is the currently provided treatment for patients having *h*, $c_{ht(h)}^{\mathrm{D}} > c_{hs(h)}^{\mathrm{D}}$, $q_{ht(h)} > q_{hs(h)}$, and next to *ht*(*h*), *hs*(*h*) is the most costly, most QALY-productive treatment for *h*. Then assuming for simplicity that the numbers of patients can be varied continuously, it is always possible to switch one or more patients $n_{i(h)}$ ($\le N_i$) having illness *i* from treatment *ik* to treatment *ij* and simultaneously to switch one or more patients n_h ($\le N_h$) having illness *h* from *ht*(*h*) to *hs*(*h*) so that

$$(c_{ij}^{\mathrm{D}} - c_{ik}^{\mathrm{D}})n_{i(h)} = (c_{ht(h)}^{\mathrm{D}} - c_{hs(h)}^{\mathrm{D}})n_h$$

that is, so that the constancy of total direct costs is preserved. But because

$$\frac{(c_{ij}^{\mathrm{D}} - c_{ik}^{\mathrm{D}})n_{j(h)}}{(q_{ij} - q_{ik})n_{i(h)}} < \frac{(c_{ht(h)}^{\mathrm{D}} - c_{hs(h)}^{\mathrm{D}})n_h}{(q_{ht(h)} - q_{hs(h)})n_h} = \frac{(c_{ij}^{\mathrm{D}} - c_{ik}^{\mathrm{D}})n_{j(h)}}{(q_{ht(h)} - q_{hs(t)})n_h}$$

$(q_{ij} - q_{ik})n_{i(h)} > (q_{ht(h)} - q_{hs(h)})n_h$ and total QALYs increase. Now sum over all such possible switches so that $\sum_h n_{i(h)} \le N_i$, and the total change in QALYs is

$$(q_{ij} - q_{ik})\sum_h n_{i(h)} - \sum_h (q_{ht(h)} - q_{hs(h)})n_h > 0$$

when the change in the health care system's total direct costs is

$$(c_{ij}^{\mathrm{D}} - c_{ik}^{\mathrm{D}})\sum_h n_{i(h)} - \sum_h (c_{ht(h)}^{\mathrm{D}} - c_{hs(h)}^{\mathrm{D}})n_h = 0$$

That is, there always exists a reassignment of patients from *ik* to *ij* and between pairs of treatments elsewhere in the health-care system that increases total QALYs but does not change total direct costs.

The reassignment therefore improves the (patient) community's health well-being; and because it does not change total direct costs, it does not reduce the income available to the community for expenditure on nonhealth goods and services either. As a consequence, it does not reduce the community's current and future consumption of nonhealth goods and services, so that the improvement in health well-being either increases the utility patients derive from this consumption or else directly improves the community's overall well-being (shifts its social welfare function upward). In either event, the reassignment unambiguously improves the community's social welfare and justifies the provision or sanctioning of treatment *ij* for patients.

But suppose now that the planner maintains a societal perspective so that its community consists of the entire population living in some specified geographic or other area. Let the total cost of any treatment *a* be denoted by c_a, write $c_a = c_a^{\mathrm{D}} + c_a^{\mathrm{IN}}$, and let the cutoff point λ be the largest ICE ratio of the form $(c_a - c_b)/(q_a - q_b)$ anywhere in the health-care system, where $c_a > c_b$, $q_a > q_b$, treatment *a* is currently provided to or sanctioned for patients, and next to treatment *a* treatment *b* is the most costly, most QALY-productive treatment for the given illness. Then for the planner having a societal perspective treatment *ij* is cost-effective if and only if

$$\frac{c_{ij}^{\mathrm{D}} - c_{ik}^{\mathrm{D}}}{q_{ij} - q_{ik}} + \frac{c_{ij}^{\mathrm{IN}} - c_{ik}^{\mathrm{IN}}}{q_{ij} - q_{ik}} < \lambda \tag{2}$$

Reasoning similar to what has just been given can be adduced to show that if (2) holds there exists a reassignment of patients such that one or more patients can be assigned to *ij* and patients can be switched

between other pairs of treatments in a way that increases total QALYs without increasing total costs. Nevertheless, the welfare implications of this result are no longer clear because nothing precludes the possibility that the reassignment increases total direct costs, and, if it does, there can be no assurance that patients can be reassigned in a way that improves the community's social welfare.

To see this is so consider a numerical example. Suppose a planner cares for patients having only four illnesses labeled 1–4, the QALY productivity and costs of which are shown in Table 1. Patients having illness 1 are currently cared for by treatment 11, and treatment 12 is a new treatment for this illness. Patients having illness 2 are currently cared for by treatment 22, and treatment 21 is the next most QALY-productive, most costly treatment for the illness. Similarly, patients having illnesses 3 and 4 are currently cared for by treatments 32 and 42, respectively, and treatments 31 and 41 are, respectively, the next most QALY-productive, most costly treatments for these illnesses. The planner's task is to decide whether treatment 12 is cost-effective and whether therefore to provide or sanction it for patients in place of treatment 11. The total-cost and total-direct-cost ICE ratios calculated on the four treatment pairs are shown in the rightmost two columns of the table.

By inspection and in dollars per QALY, $\lambda = 8750$ and $\lambda^D = 7400$. Because $(c_{12} - c_{11})/(q_{12} - q_{11}) = 8400 < \lambda$, treatment 12 is cost-effective by the decision rule (2). Hence the planner will attempt to switch patients having illness 1 from treatment 11 to treatment 12 and to switch patients from more costly to less costly treatments for other illnesses so as to increase total QALYs without increasing total costs. In the example the only possible such switch is between treatments 11 and 12 and treatments 22 and 21, because $(c_{12} - c_{11})/(q_{12} - q_{11})$ is larger than both $(c_{32} - c_{31})/(q_{32} - q_{31})$ and $(c_{42} - c_{41})/(q_{42} - q_{41})$, and any switch of patients from 11 to 12 and from 32 to 31 or from 42 to 41 that preserves the constancy of total costs reduces or does not change total QALYs. (For a proof of the claim see Appendix.) Accordingly, the planner can switch patients from treatment 11 to treatment 12 and from treatment 22 to treatment 21 in the ratio of 5 : 6, and each such switch increases total QALYs by 0.01 without changing total costs. If it is possible to perform the reassignment, the community's overall social welfare is therefore unambiguously improved. Nevertheless, every switch of five patients from 11 to 12 and six patients from 22 to 21 also increases total direct costs by $490. Indeed, because $(c_{12}^D - c_{11}^D)/(q_{12} - q_{11}) > (c_{22}^D - c_{21}^D)/(q_{22} - q_{21})$, any switch of patients from treatment 11 to treatment 12 and from treatment 22 to treatment 21 that increases total QALYs increases total direct costs. (A proof of the assertion is given in the Appendix.) The issue then is how the increase in total direct costs affects the planner's ability to assign patients to treatment 12 and how it bears on the conclusion that treatment 12 is cost-effective.

If the planner elects to pay the added direct cost by increasing its income—by raising its insurance rates or its treatment prices, by securing additional funding from government, or in some other way—any such action absorbs income or resources that the community spends on current or future nonhealth consumption and reduces the utility that the community derives from that consumption. The reduction in utility is not measurable or not measurable as QALYs or other quantifiers of health benefits, but it offsets in whole or part of the gain in social welfare attributable to the community's improved health status, and the most that can be concluded is that treatment *ij* is possibly cost-effective. If the planner is inefficient and operates with idle resources or income, it can, of course, meet part or all of the increase in its total direct costs by employing those resources or spending its unused income. But if that is true, a CEA is unnecessary to justify providing or sanctioning *ij* for patients. The planner need only provide or sanction the treatment by employing its idle resources or spending its unused income, and it is either unnecessary or not obviously necessary to reassign patients to treatments for other illnesses.

Table 1 Hypothetical QALY productivity and costs of treatments for four illnesses and ICE ratios defined on the treatments

Illness	Treatment	QALYs per patient	Total cost per patient ($)	Direct cost per patient ($)	Indirect cost per patient ($)	ICE ratios ($/QALY)	
						Total cost	Direct cost
1	12	0.15	570	475	95	8400	7000
	<u>11</u>	0.10	150	125	25		
2	<u>22</u>	0.06	750	520	170	8750	6750
	21	0.02	400	310	90		
3	<u>32</u>	0.20	630	570	60	7600	7400
	31	0.15	250	200	50		
4	<u>42</u>	0.08	360	270	90	6250	4500
	41	0.04	110	90	20		

Currently provided treatments are underlined.
ICE, incremental direct-cost-effectiveness; QALY, quality-adjusted life-year.

It is reasonable, however, to think that planners do not usually maintain idle resources or income, that they ordinarily spend up to their resource or budget limits, and thus that they generally operate with binding constraints on their direct costs. But if the planner in the example maintains a binding constraint on its total direct costs, it cannot carry out the reassignment of patients that increases total QALY output and preserves the constancy of total costs without violating that constraint. The gain in social welfare obtainable by switching patients to the putatively cost-effective treatment *ij* is therefore merely hypothetical, and the treatment is not cost-effective regardless of the implication of (2).

Although it is unlikely, it is conceivable that a planner might spend less than the upper bound on its total direct costs on account of regulatory restrictions. It is hard to say what these restrictions might be, but if they exist they force an even tighter constraint on total direct costs than the constraint defined by the planner's resource base or budget. Hence the planner is no more able to afford the increase in total direct costs than it would be if the constraint were imposed by its budget or resource limits, and again treatment *ij* again is not cost-effective. It is also unlikely but conceivable that the planner operates with a constraint on its total indirect costs, that this constraint is binding, and that the constraint on its total direct costs is not binding because any additional expenditure of direct costs to produce or sanction more or more costly treatments would increase total indirect costs and violate the constraint on those costs. In that case, the planner could absorb the added direct cost of the patient reassignment as long as the reassignment does not increase total indirect costs. In the example, each 5:6 switch of patients between treatments 11 and 12 and between treatments 22 and 21 reduces total indirect costs by $130. Thus in this one scenario, treatment 12 is cost-effective—that is, it can be provided to patients and its adoption increases the community's social welfare—if the planner initially maintains a binding constraint on its total indirect costs and a nonbinding constraint on its total direct costs. Unfortunately, to know that these conditions are satisfied, the analyst must be familiar with the planner's managerial policies, and because these policies are not necessarily uniform across planners, treatment 12 may be cost-effective in one healthcare system and not in another, even when treatment costs and QALY productivity and the cutoff point λ are the same in both systems.

None of this is to say that the cutoff-point decision rule always or even regularly gives false or misleading information when the planner's perspective is societal. It will be evident, for example, that with only a few changes in the figures in Table 1, treatment 12 becomes cost-effective (or not) beyond argument. All the same, whenever the planner holds a societal perspective, a new treatment is judged cost-effective by the cutoff-point decision rule, and the required reassignment of patients among treatments required by the cutoff-point methodology increases total direct costs. It is hard to think of plausible scenarios in which the planner can afford the added cost or, if it can, in which its actions unequivocally increase its community's social welfare. And not only is the conclusion that a new treatment is cost-effective not necessarily reliable; but to assess the reliability of the conclusion, the analyst must be familiar with the planner's cost structure and management. Simple applications of the cutoff-point rule are not sufficient. (Notice that in the example, treatment 12 is cost-effective by the rule even if the planner does not take a societal perspective because $(c_{12}^{D} - c_{11}^{D})/(q_{12} - q_{11}) = 7000 < \lambda^{D}$, and the result contributes nothing to a resolution of the reliability problem.) This suggests that any finding that a new treatment is cost-effective when a societal perspective is assumed should be accompanied by the qualification that it is valid only if the planner can and does execute the required reassignment of patients among treatments without increasing its total direct costs.

Conclusion

The premise that adding indirect costs to direct costs corrects or adjusts the outcomes of CEAs for the incidental effects of formal health care on the community at large is intuitively appealing, but it also introduces complications into models of CEA. It can be incorporated in the many-illness, many-treatment model only by imposing a constraint on the planner's total indirect costs, but this constraint cannot be specified arbitrarily because if it is, it implies unrealistically that the planner must be indifferent between losses of QALYs and reductions in total indirect costs. This article proposes an adaptation of the model in which with some computational effort the rational planner can obtain a resource-constrained maximum of total QALY output and at the same time realize a predetermined marginal trade-off rate between QALYs and indirect costs. Other methods of building welfare objectives into the many-illness, many-treatment model may exist as well, but without a constraint on total indirect costs, the model lacks significant welfare content.

Although cost constraints have no explicit role in the standard one-illness, two-treatment model of CEA, they cannot really be disregarded when the planner elects to recognize indirect costs. If the planner sets a constraint on its total indirect costs, it presumably chooses the constraint so as to maintain a target marginal trade-off rate between QALYs and indirect costs. Nevertheless, it is hard to see how such a marginal trade-off rate can be built into the model. The most obvious candidate, the largest ICE ratio of the form $(c_a^{IN} - c_b^{IN})/(q_a - q_b)$—the largest marginal indirect cost

of QALYs—in the health-care system, denote it by λ^{IN}, plays no immediate part in the cutoff-point decision rule, and, except that it is positive, it can take on any value at all. In addition, the difference $\lambda^{IN} - (c_{ij}^{IN} - c_{ik}^{IN})/(q_{ij} - q_{ik})$ can have any algebraic sign whether or not treatment 12 is cost-effective by the cutoff-point rule (2). In view of these indeterminacies, there can be no assurance that the rule gives a conclusion consistent with the planner's choice of a target QALY-indirect cost trade-off rate by which it sets the constraint on total indirect costs.

Whether or not they actually constrain their total indirect costs, it is reasonable to think that all planners tend to face constraints on their total direct costs, either budget limits or the total costs of the limited supplies of formal health-care resources that they command or direct. When these constraints are not binding because planners are able to meet the direct costs of assigning patients to new, putatively cost-effective treatments by increasing their incomes, the welfare implications of the cutoff-point methodology are compromised. But when the constraints are binding, assigning patients to a societally cost-effective treatment can increase the health-care system's total direct costs, forcing a violation of the constraint and making the treatment not cost-effective per se, and there seem to be no ways of determining whether the constraint is violated short of making a detailed investigation of the planner's behavior and the cost and QALY structure of its health-care system. As a consequence, unless it can be shown empirically that violations of the constraint are so rare that they can be safely ignored, the reliability of the cutoff-point decision rule cannot be taken for granted whenever the planner is assumed to hold a societal perspective.

Source of financial support: none

References

1 Weinstein MC. Clinical decisions and limited resources. In: Weinstein MC, Fineberg HV, eds. Clinical Decision Analysis. Philadelphia, PA: W.B. Saunders, 1980.
2 Laffont J-L. Fundamentals of Public Economics. Cambridge, MA: MIT Press, 1988.
3 Anderson A, Levin LA, Ertinger BG. The economic burden of informal care. Int J Technol Assess Health Care 2002;18:46–64.
4 Mullins CD, Whitelaw G, Cooke JL, Beck EJ. Indirect cost of HIV infection in England. Clin Ther 2000;11:1333–45.
5 Jacobs P, Fassbender K. The measurement of indirect costs in the health economics evaluation literature. Int J Technol Assess Health Care 1998;14:799–808.
6 Posnett J, Jan S. Indirect cost in economic evaluation: the opportunity cost of unpaid inputs. Health Econ 1996;5:13–23.
7 Luce BR, Manning WG, Siegel JE, Lipscomb J. Estimating costs in cost-effectiveness analysis. In: Gold MR, Siegle JE, Russell LB, Weinstein MC, eds. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press, 1996.
8 Weinstein MC, Siegel JE, Garber AM, et al. Productivity costs, time costs and health-related quality of life: a response to the Eransmus Group. Health Econ 1997;6:505–10.
9 Brouwer BF, Koopmanschap MA, Rutten FFH. Patient and informal caregiver time in cost-effectiveness analysis. A response to the recommendations of the Washington Panel. Int J Tech Assess Health Care 1998;14:505–13.
10 Weisbrod BA. Economics of Public Health. Philadelphia, PA: University of Pennsylvania Press, 1961.
11 Drummond M, McGuire A. Economic Evaluation in Health Care: Merging Theory with Practice. Oxford: Oxford University Press, 2001.
12 Muennig P. Designing and Conducting Cost-Effectiveness Analyses in Medicine and Health Care. San Francisco: Jossey-Bass, 2002.
13 Garber AM, Weinstein MC, Torrance GW, Kamlet MS. Theoretical foundations of cost-effectiveness analysis. In: Gold MR, Siegle JE, Russell LB, Weinstein MC, eds. Cost-Effectiveness in Health and Medicine. New York: Oxford University Press, 1996.
14 Hadorn DC. Setting health care priorities in Oregon. Cost-effectiveness meets the rule of rescue. JAMA 1991;265:2218–25.
15 Eddy DM. What's going on in Oregon? JAMA 1991;266:417–20.
16 Tengs O, Meyer G, Siegel JE, et al. Oregon's Medicaid ranking and cost-effectiveness analysis: is there any relationship? Med Dec Making 1996;16:99–107.
17 Johannnesson M, Weinstein MC. On the decision rules of cost-effectiveness analysis. J Health Econ 1993;12:459–67.
18 Karlsson G, Johannesson M. The decision rules of cost-effectiveness analysis. Pharmacoeconomics 1996; 9:113–20.
19 Laska EM, Meisner M, Siegel C, Stinnett AA. Ratio-based and net health benefit-based approaches to health care resource allocation: proofs of optimality and equivalence. Health Econ 1999;8:171–4.
20 Torrance GW, Thomas WH, Sackett DL. A utility maximization model for evaluation of health care programs. Health Serv Res 1972;7:118–33.
21 Birch S, Gafni A. Cost-effectiveness/utility analyses: do current decision rules lead us to where we want to be? J Health Econ 1992;11:279–96.
22 Stinnett AA, Paltiel AD. Mathematical programming for the efficient allocation of health care resources. J Health Econ 1996;15:641–53.
23 Dantzig GB, Thapa MN. Linear Programming. 1: Introduction. New York: Springer-Verlag, 1997.
24 Takayama A. Analytical Methods in Economics. Ann Arbor, MI: University of Michigan Press, 1993.
25 Dixit AK. Optimization in Economic Theory (2nd ed.). Oxford: Oxford University Press, 1990.

26 Kaplan RM, Bush JW, Berry CC. Health-related quality of life measurement for evaluation research and policy analysis. Health Psychol 1982;1:61–80.
27 Ganiats TG, Schneiderman LJ. Principles of cost-effectiveness research. J Family Prac 1988;27:77–84.
28 Laupacis A, Feeny D, Detsky AS, Tugwell PX. How attractive does a new technology have to be to warrant adoption and utilization? Tentative guidelines for using clinical and economic evaluations. Can Med Assoc J 1992;146:473–81.
29 Weinstein MC. From cost-effectiveness ratios to resource allocation: where to draw the line? In: Sloan FA, ed. Valuing Health Care. Cambridge: Cambridge University Press, 1995.
30 Gafni A, Birch S. NICE methodological guidelines and decision making in the National Health Service in England and Wales. Pharmacoeconomics 2003;21: 149–57.
31 Birch S, Gafni A. The "NICE" approach to technology assessment: an economics perspective. Health Care Manag Sci 2004;7:35–41.

Appendix

Suppose $ht(h)$ and $hs(h)$ are treatments for illness h, $c_{ht(h)} > c_{hs(h)}$, and $q_{ht(h)} > q_{ht(h)}$. If

$$\frac{c_{ij} - c_{ik}}{q_{ij} - q_{ik}} \geq \frac{c_{ht(h)} - c_{hs(h)}}{q_{ht(h)} - q_{hs(h)}} \tag{3}$$

and $n_{i(h)}$ patients are switched from ik to ij and n_h, patients are switched from $ht(h)$ to $hs(h)$ so that $(c_{ij} - c_{ik})n_{i(h)} = (c_{ht(h)} - c_{hs(h)})n_h$, then

$$\frac{(c_{ij} - c_{ik})n_{i(h)}}{(q_{ij} - q_{ik})n_{i(h)}} \geq \frac{(c_{ht(h)} - c_{hs(h)})n_h}{(q_{ht(h)} - q_{hs(h)})n_h} = \frac{(c_{ij} - c_{ik})n_{i(h)}}{(q_{ht(h)} - q_{hs(h)})n_h}$$

whence $(q_{ij} - q_{ik})n_{i(h)} \leq (q_{ht(h)} - q_{hs(h)})n_h$. If the switches between treatments are such that $(q_{ij} - q_{ik})n_{i(h)} > (q_{ht(h)} - q_{hs(h)})n_h$,

$$\frac{(c_{ij} - c_{ik})n_{i(h)}}{(q_{ij} - q_{ik})n_{i(h)}} \geq \frac{(c_{ht(h)} - c_{hs(h)})n_h}{(q_{ht(h)} - q_{hs(h)})n_h} = \frac{(c_{ht(h)} - c_{hs(h)})n_h}{(q_{ij} - q_{ik})n_{i(h)}}$$

which implies $(c_{ij} - c_{ik})n_{i(h)} > (c_{ht(h)} - c_{hs(h)})n_h$. In words, if (3) holds any switch of patients between ik and ij and between $ht(h)$ and $hs(h)$ that preserves the constancy of total costs reduces or does not change total QALYs, and any switch of patients that increases total QALYs increases total costs. The results are obviously also true if costs are defined as direct costs.