Chapter 5: Cost-effectiveness analysis

Jeremy A. Lauer[1], Alec Morton[2] and Melanie Bertram[3]

1 Introduction

Cost-effectiveness analysis (CEA) is a form of economic evaluation concerned with efficiency: that is, with achieving the most for the resources ("value for money"). For example, imagine that you have billions of dollars to allocate to global health and have to decide how to spend it. Or, you are a minister of health who wants to rationalize the use of your budget. Or imagine you are the head of an agency mandated to improve human health, and you need to know what strategies to recommend. The primary aim of this chapter is to show that, in each of these cases, you ought to know something about CEA if you want to achieve your objectives. Fortunately, a number of excellent standard accounts are available (Jamison, 2009; Sculpher et al., 2017). So rather than retrace well-trodden ground, this chapter offers a complementary approach intended to respond to the needs of non-economists. It also offers a novel perspective on CEA that should be of interest to specialists. A related aim of the chapter is to explain why – in spite of its relevance – CEA remains underused for problems like those mentioned above, and misused in many cases where it is applied. We

[1] Economist, Health Systems Governance and Financing, World Health Organization; lauerj@who.int.

[2] Professor, Management Science, University of Strathclyde Glasgow; alec.morton@strath.ac.uk.

[3] Technical Officer, Health Systems Governance and Financing, World Health Organization; bertramm@who.int.

attempt to show therefore both why CEA is often appealed to and why its basic principles remain opaque.

CEA is related to ethics, since it helps to choose between competing approaches to achieving a given objective. In other words, it helps decision makers with a particular aim in mind to determine what states of the world they have most reason to bring into being, given the inputs required. CEA has no intrinsic link to health as an outcome, yet the community of people engaged in health policy and practice has a longstanding tradition of relying on CEA, which is less prevalent in other areas of public policy. The widespread use of CEA in health (as opposed to, say, in transport or energy policy) might be attributed to the fact that it embodies the idea that health is a benefit in its own terms, whereas many other policy domains are merely instrumental to human well-being. This distinguishes CEA from its main competitor in economic evaluation, benefit-cost analysis (see Chapter 7), since BCA measures health and other benefits[4] in terms of a common unit of account. Of course, there are reasons other than its measuring outcomes in health terms that encourage the use of CEA for decision-making in health.

In public policy, choosing which things to do (and by the same token which not to) inevitably involves politics, as well as ethics and economics. Decision-making in health therefore requires the coordination of multiple actors, with differing interests and objectives, around a

[4] In economic evaluation, the term "benefits" is often reserved to refer to the economic value of outcomes (cf. BCA), whereas "effects" is usually used to refer to those same outcomes when denominated in natural units; as a convenient shorthand, here we occasionally refer to health outcomes measured in natural units as "benefits" in the generic sense of "desirable outcomes".

particular action or set of actions. The coordination mechanism most familiar to economists is the market: faced with price signals, individuals make independent choices but the logic governing their actions results in a certain level of cooperation around, for example, which services to demand and which to provide. CEA plays a similar role. However, unlike prices, which provide direct financial incentives, as a non-market coordination mechanism CEA needs to persuade in other ways.

In *The nature of the firm* (1937) Ronald Coase observed that, while the use of the price system has advantages such as those described in the theory of competitive equilibrium, it also imposes costs. He hypothesized that non-market coordination mechanisms, for example, employment contracts, might arise when the costs of using the market are high. Current industrial-organization theory subsumes Coase's observation as a special case of the broader "problem of vertical integration", i.e. How do diverse actors in a common production process coordinate their activities when there are significant transaction costs, asset-specific (i.e. non-transferable) investments are required, uncertainty is prevalent and asymmetries of information are the rule? The solution noted by Coase is for the supplier and buyer to enter into a contract that aligns their individual rewards and incentives around the relevant production process. According to Coase, this solution (now understood as part of a more general coordination problem) gives rise to the firm.

Like the firm, other organizing features of the modern health system (e.g. pay-for-performance schemes, results-based financing, gatekeeping mechanisms, and public-private partnerships) are contractual − and, by extension, legal, institutional and governance − solutions to problems of vertical integration. Such problems are built in to health systems at the root, since they arise from the age-old precept that one should not be one's own doctor.

Yet, at the same time, a doctor (or other health worker) may not have her patient's best interests at heart, due to another aspect of the coordination problem called principal-agent discordance.

CEA proposes a coordination mechanism to address problems of vertical integration and principal-agent discordance in delivering healthcare in a complex system, that is, one composed of multiple layers of market and non-market relationships; however, CEA is perhaps the only coordination mechanism in the health system that is not based on contractual, institutional, governance, or – alternatively – on market arrangements. Like the price system, CEA provides a code by which actors can transmit and receive information; like contracts and related institutions, CEA reduces transactional and operational risks for decision makers. CEA constitutes a non-market and non-contractual solution to the problem of how to coordinate actions, taken at different levels and by actors with potentially different interests, around the common goal of producing the best health outcomes. Ultimately, CEA is a partial answer to a yet-unsolved problem, so some of the other chapters in this book address complementary or competing approaches.

The actors in the health system who are potentially influenced by CEA include not only those mentioned in the first paragraph but also clinical staff, such as doctors deciding on treatments, and hospital managers or public health administrators deciding on the organization of departments or the distribution of budgets. They also include insurance companies deciding on services to reimburse, healthcare providers and clinicians deciding on inputs to use (or outcomes to purchase), and policy makers deciding on programmes to fund, recommend, regulate, or legislate. Finally, the relevant actors can include individuals making choices affecting their own health status. CEA by itself may not provide financial incentives; yet, like

the market, it is a coordination mechanism that persuades individuals to act in certain ways of their own free will rather than through the threat of coercion, such as in contractual and legal mechanisms.

CEA is related to theories of welfarism, utilitarianism and consequentialism: CEA is welfarist because its notion of benefit (i.e. health) is integral to well-being; it is utilitarian to the extent that it orders outcomes by the magnitude of benefit; it is consequentialist because it ranks choices on the basis of outcomes. CEA represents a simple approach to accounting for benefits (health) and harms (costs) that is broadly acceptable to utilitarian consequentialists concerned with social welfare. CEA also respects individual autonomy in a way consistent with the tradition of liberal thought. Although it may not be a perfect expression of these ideas, the broad coherence of CEA with some mainstream views in moral and political philosophy represents yet another reason explaining its appeal for the establishment of priorities in health, as well as its acceptability to wider publics. Ethicists and effective altruists[5] are, after health economists, probably the largest users and promoters of CEA for decision-making (MacAskill, 2015).

The standard form of CEA is indifferent to the distribution of outcomes. Put simply, CEA does not care who benefits. This is the main respect in which CEA differs from mainstream moral philosophy. Concerns about differing claims of separate persons are at the root of concepts of fairness (Adler, 2012), and CEA might therefore be characterized as unfair. Indeed, CEA has frequently been criticized by proponents of social-choice theories that emphasize distributional concerns (see Chapters 9–11, this volume), for example, giving

[5] See The Stanford Encyclopedia of Philosophy (http://plato.stanford.edu/entries/impartiality) for a definition of effective altruism.

priority to the worse off. CEA might therefore be described, albeit facetiously, as the longest-lasting, most popular – and yet also the most thoroughly discredited – idea in the history of modern public policy.

We have so far offered four main reasons to explain the appeal and relevance of CEA:

1. It denominates benefits in health terms.
2. It addresses vertical integration and other coordination problems in the health system.
3. It respects individual autonomy.
4. It is broadly consistent with main streams in the philosophy of social choice.

Finally, the enduring appeal of CEA is presumably also due to its relatively simple technical formulation (see *The primal*, below). Whatever its difficulties, in comparison with other social-choice theories, CEA is easy both to implement and to explain. CEA is therefore suited to the needs of governments, donors and policy makers to demonstrate accountability and transparency in decision-making. In sum, CEA remains a potent expression of important social concerns that also addresses many problems faced by decision makers. For this reason, the use of CEA is not likely to disappear from priority setting in global health.

Two kinds of decision-making

The standard presentation of decision-making in health supposes a decision maker needs to choose whether to implement a given proposal, or an alternative. We refer to this as localized decision-making, and we distinguish it from what we term strategic decision-making. In health, both kinds of decision-making employ CEA, and both are familiar from ordinary experience. For example, when moving into a new home, one typically establishes a budget and identifies a complete set of furnishings and equipment to purchase: this is strategic decision-making. However, during subsequent occupancy one purchases new furnishings or

pieces of equipment in response to breakage, obsolescence, or as fresh opportunities arise: this is localized decision-making.

When CEA is used to support strategic decisions, we call it *generalized* CEA; when it is used to support localized decisions, we term it *marginal* CEA. The defining feature of generalized CEA – in the terminology used here – is that it determines, at least potentially, a large number of (marginal) decisions. Generalized CEA is therefore relevant to the establishment of strategic health-system objectives (what we call here "priorities"), whereas marginal CEA is relevant mainly to the realization of objectives (or priorities) that have already been clearly established. There is no hard boundary, however, between strategic and localized decision-making, nor therefore between generalized and marginal CEA.

2 Generalized versus marginal CEA

Two complementary approaches

CEA originates with the mathematics of constrained optimization, where the cost-effectiveness problem is presented as a "linear program". In that problem, the decision maker's objective is to maximize benefit (i.e. achieve the highest feasible level of health) while staying within the available budget. This problem is called a "program" because there are established recipes (algorithms) for obtaining solutions; it is "linear" because its equations can be represented by straight lines.

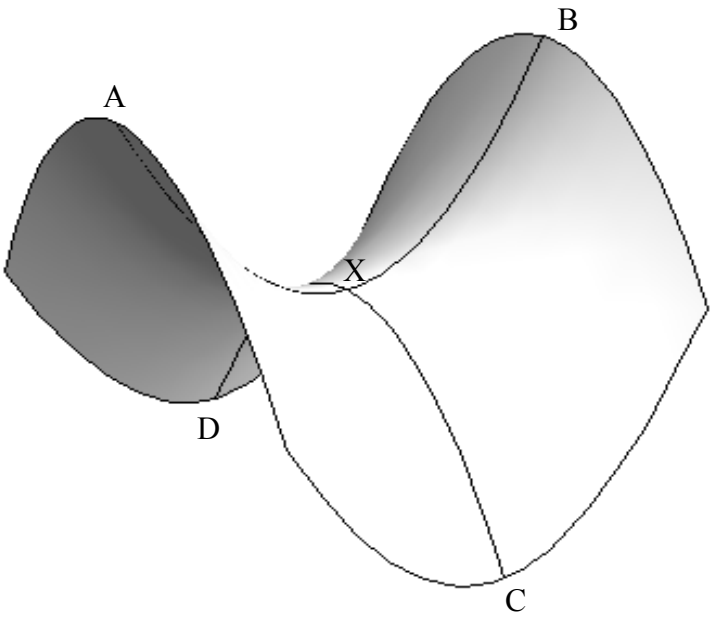

Figure 1. Duality and the saddle point solution, X.

Central to the mathematics of constrained optimization is the theory of duality. According to duality, there are always two ways of tackling a constrained optimization problem:

1. the Primal and
2. the Dual.

Duality theory shows that if the primal approach involves the maximization of some objective, then its dual involves the minimization of a complementary objective. Put visually (Figure 1), duality says that, if point X on line CD represents the point of maximum population health, then, taking into account constraints such as the budget, there is a unique line (say, AB) for which the same point X is a minimum. The solution (X) is called a "saddle point" because of this dual role, as both a maximum and a minimum. Duality further says that, given line CD and its constraints, without any additional information we can always construct the line AB and its constraints simply by following a recipe (algorithm).

It turns out that in CEA the primal and the dual correspond to generalized and marginal CEA, respectively. An early, canonical, formulation of CEA (Weinstein and Zeckhauser, 1973) proposes to maximize health subject to a budget constraint; this formulation corresponds to generalized CEA (and strategic decision-making). With this as our primal, duality says that marginal CEA (corresponding to localized decision-making) must require the minimization of some quantity, and it must yield the same solution (i.e. X).

In everyday terms, suppose we were to buy a fully furnished house: duality seems to imply that, replacing furnishings one at a time (localized decision-making), we should arrive in due course at the same set of furnishings that we would have purchased, all at once, on moving into an empty home (strategic decision-making). Practitioners of CEA typically rely, either implicitly or explicitly, on this equivalence. Their reliance is apparently supported by the theory of duality; however, as we shall see, a simplistic faith in the claims of duality theory constitutes the central fallacy in the practice of CEA.

The primal and strategic decision making

Let us imagine that there are a fixed number of health technologies, and that each technology addresses a single health problem. In mathematical notation, the primal cost-effectiveness problem is written:

$$\text{maximize } w \cdot x$$

$$\text{subject to } c \cdot x \leq \text{B}.$$

These symbols have straightforward meanings, so the primal can be read nearly as a natural English sentence: “maximize population health subject to costs being less than or equal to the budget”. The variable x represents the number of people who benefit from technologies, w represents the per-person health benefit (i.e. effectiveness) of the technologies, and c

represents their unit price. For each potential value of x the decision maker might choose, the corresponding cost ($c \cdot x$) needs to be less than or equal to the available budget (B).

To solve the primal it is sufficient to list technologies in increasing order of (average) cost effectiveness (c/w). We then implement the first technology first, calculate the (incremental) cost effectiveness of each remaining one relative to the first, re-ordering the list and implementing in turn the next best, and repeating these steps until the budget is exhausted. When we run out of money some technologies will cover 100% of those who stand to benefit, for example, the technologies implemented first; however, some will cover 0%, namely, the technologies still in the queue. Finally, the coverage of one technology will usually be somewhere between 0% and 100% (i.e. the one being implemented when the money runs out). Once this procedure is complete the decision maker does not need to do anything further until more funds become available or one of the technologies becomes obsolete. Note that this description is highly stylized in comparison with actual practice; nevertheless, this is the fundamental construct that generalized CEA assumes as a framework for decision-making.

The dual and localized decision-making

Marginal CEA, on the other hand, corresponds to localized decision-making. In the conventional account, it takes past decisions as given and looks only at adding or replacing things one at a time. In terms of our example, marginal CEA is for deciding whether to buy a new lamp for the hallway, while leaving all the other furnishings as they are.

As the rationale for generalized CEA comes from the primal, so the rationale for marginal CEA comes from the dual version of the cost-effectiveness problem. The dual, which is set up according to a mathematical recipe (Luenberger, 1973), is written here as follows:

minimize $\lambda \cdot B + \mathbf{1} \cdot \mu$

subject to $c \cdot \lambda + \mu \geq w$.

Unlike the primal, the equation of the dual cannot be readily rendered into English, and it is not intuitive to understand; this fact is central to our argument. Nevertheless, we offer an interpretation, below (see *The fable of the dual*).

First, though, we note some important practical points. The dual has nothing to do with "cost minimization", as is sometimes claimed. Cost minimization (i.e. minimizing $c \cdot x$, subject to achieving a fixed quantity of health $w \cdot x \geq H$) is, like health maximization, a perfectly valid mathematical approach to CEA (cf. Weinstein and Zeckhauser, 1973).[6] However, cost minimization and health maximization are both primal formulations; moreover, the cost-minimization and the health-maximization versions do not in general give the same solution (see *Optimal pathways*, below). A given primal and its dual, however, invariably do. Furthermore, unlike their dual expressions, both primals have straightforward interpretations in natural language. Finally, solving either of the primals requires knowing the cost effectiveness of all technologies and following a recipe like that given above. Solving the dual requires higher mathematics.

Second, something is minimized in the dual, but this quantity ($\lambda \cdot B + \mathbf{1} \cdot \mu$) has no natural name; the symbols used express merely mathematical rather than real features of the problem (e.g. λ and μ; **1** is the unit vector). An important feature from the primal (i.e. the number of people benefitting, x) has completely disappeared. Nor do we line up technologies in order of cost effectiveness: instead, using calculus, we solve for the optimal value of the new variable λ. In

[6] Here we take health maximization, subject to a budget constraint, as the natural formulation of the decision maker's problem, rather than cost minimization, subject to a health constraint.

mathematical terms, λ is a "shadow price" ("shadow" meaning "not directly observable"). The variable λ indicates the health benefit obtainable by increasing the budget by a single unit, a quantity which differs according to the size of the budget and the effectiveness of available technologies. The optimal value for λ for a particular cost-effectiveness problem is conventionally called the "cost-effectiveness threshold". Note that the concept of a threshold is completely absent in generalized CEA; however, both the threshold idea, and its particular value, are absolutely central to marginal CEA.

The threshold is the key to making one-off decisions. Suppose we have distinguished λ^* as a solution value (threshold) for the dual, above: armed with this information, now the homeowner no longer has to check her bank balance or determine the cost effectiveness of all technologies; she merely has to calculate the efficiency of the, say, lamp in effectiveness-cost terms.[7] If the lamp scores above λ^* she should purchase it, but not otherwise. So once we know the value of λ^*, we can easily add or replace technologies one at a time.

[7] For a health-maximizing primal, the λ^* of the corresponding dual is an "effectiveness-cost threshold" (with units w/c) and shows the *minimum* value the decision maker should accept. For a cost-minimizing primal, however, λ^* is a cost-effectiveness threshold (with units c/w) and shows the *maximum* value she should accept. By convention, however, thresholds are always presented in cost-effectiveness terms. Our discussion of the dual refers to an effectiveness-cost threshold (w/c), corresponding to the mathematical set up. Elsewhere in the chapter we follow convention and refer to cost-effectiveness thresholds (c/w).

The fable of the dual

Although the threshold rule is easy to apply, the interpretation of the dual is not straightforward, and this difficulty conceals pitfalls for localized decision-making. The story of the primal is one of "maximizing health", or alternatively of "minimizing costs"; what then is the dual about? In a fanciful yet precise rendering, the "fable of the dual" can be told as follows.

Suppose the decision maker is facing an opponent trying to minimize population health: the dual represents this opponent's objective. The opponent is the "owner" of the dual, just as the decision maker was the owner of the primal. Yet the opponent cannot simply reduce health to zero. The decision maker controls a budget, and for each budgetary unit she can secure λ units of health. Thus, the decision maker is guaranteed to achieve population health of at least $\lambda \cdot B$. However, without higher mathematics she doesn't know the value of λ, which for now is under the opponent's control. She and her opponent therefore conduct a game. The opponent moves first by naming a threshold. Suppose he claims the value of λ is equal to the effectiveness-cost ratio of a particular technology (perhaps one he wishes to promote). Other technologies offer better value for money, and for these the decision maker would obtain an additional health gain of μ (which is the meaning of the other new variable in the dual equation; the mathematical name for μ is "complementary slackness"). As the decision maker considers the bid, her opponent tries to keep λ at his preferred value by concealing superior technologies. As much as he tries to do this, however, the constraint ($c \cdot \lambda + \mu \geq w$) forces the value of μ to become large, revealing the potential for additional gains inherent in the effectiveness (w) of existing technologies. Given his objective of minimizing health, but facing a countervailing constraint alerting the decision maker to the untapped potential of technologies, after some – possibly many – rounds of the contest it becomes clear to the

opponent that the best he can manage is to allow λ to equal the effectiveness-cost ratio of the technology the decision maker was implementing when the money ran out. The careful decision maker has refused every other bid. This is what higher mathematics confirm: the contest between the decision maker and her opponent deadlocks only when the opponent's proposed value for λ yields the same position (X) that was found in the primal; until this point is reached, the decision maker can always drive λ up, and μ down, by diligently seeking out superior technologies. In general, the decision maker may have to examine all the technologies to be sure her opponent has not thwarted her in getting the best possible value for money. The struggle with her opponent has, however, given the decision maker an important side benefit: henceforth, she can use λ^* as the sole criterion for making one-off choices.

Localized versus strategic decision-making

In summary:

- The primal is what most practitioners are thinking of when they do CEA. The primal embodies a simple and intuitive solution strategy, which explains its conceptual appeal; however, implementing the primal requires knowing the cost effectiveness of all technologies.
- The dual is what practitioners actually rely on when they do CEA. Implementing the dual requires knowing only a single value, but solving for the threshold requires higher mathematics (or an exhaustive struggle with a determined opponent). Since thresholds summarize a lot of information, they are useful for making decisions about things individually.

Until now, however, probably no one has described the rationale of CEA in dual terms: "choosing a shadow price to minimize the value of the budget denominated in health units,

while using only the best technologies". This is because, in spite of being a succinct distillation, such language hardly facilitates the understanding of CEA, or of priority setting in global health. CEA practitioners inevitably appeal to one of the primal formulations (health maximization or cost minimization) to explain their discipline. Nevertheless, most believe that defining (or using) a threshold is an essential feature of CEA, and only rarely do they consider making choices over the whole set of technologies. In effect, cost-effectiveness practitioners suffer from a form of cognitive dissonance: they "think primal, but do dual". Unfortunately, this can result in errors.

The focus of marginal CEA is using a threshold for making discrete choices. This works as intended, however, *only if we are already at the optimal position*. It works, in other words, only when we do not need to confirm (or reverse) past decisions and can focus exclusively on the decision under consideration (see *Example*, below). In other words, we have the "correct" threshold only when we have already maximized health, either because the health system happens to have been optimized in practice, or because we've gone through the exercise of performing a generalized CEA. In practice, usually we are not at the health maximizing position, and largely because of this fact − coupled with the appeal of an apparent shortcut − the use of thresholds in CEA has become fraught with various controversies (Claxton et al. 2015b ; Marseille et al., 2015; Bertram et al., 2016).

Let's now re-frame the decision maker's problem in terms of our previous example. In fact, the typical situation is neither one of "furnishing an empty house" or of "adding or replacing things, one at a time", but rather something like moving into a flat that shows the accumulation of choices made by a succession of past tenants. Some things have been left behind, either by accident or because they were too difficult to remove, but the flat is not

what you could call “fully furnished”. And we have only a temporary lease. It’s clear in such a case that we do not have the luxury of “adding only new things”, since we may urgently need to reconsider, for example, the orange shag carpet in the kitchen. At the same time, we cannot possibly re-visit all past choices, given our limited time frame, and budget.

Encouraging identification of the current situation with what is optimal is the main source of confusion resulting from the dual. Encouraging the idea that every past decision can be revisited is the principal conundrum resulting from generalized cost-effectiveness analysis. Only when we implement the best technology first can we rely exclusively on incremental ratios and one-off decisions; otherwise we need to consider the cost effectiveness of everything in order to find out first what is best. But only when we have sufficient time, and budget, can we possibly reconsider everything already in place. Using incremental ratios from a sub-optimal starting point usually conceals relevant alternatives (just as the decision maker’s opponent wished to do), and so we “bake in” past mistakes. At the same time, performing a generalized CEA requires a lot of time and effort, and not all mistakes can be undone in practice. “Mistakes” might be relatively minor ones, such as purchasing a shipment of the wrong malaria tablets. But when they commit the health system to an entire delivery platform (such as a luxury hospital providing specialty care for a minority of the population), mistakes can take years or even decades to repair.

Relying naively on a threshold is not consistent with the decision maker’s objective of health maximization (or cost minimization). All conventional threshold-based rules share this flaw. Thresholds are moreover susceptible to being hijacked by “opponents” with, for example, commercial interests in decisions concerning the technologies offered in benefit packages. Indeed, such real-world opponents operate much like the one in our fable: they downplay

relevant alternatives and tout the cost effectiveness of preferred technologies. Making a threshold approach robust to misuse requires a diligent decision-maker with a broad and deep knowledge of technologies, as well as an unwavering commitment to the legitimate priorities of the health system.

Generalized CEAs are expensive and difficult, but are robust to the criticism of being self-defeating. Unfortunately, self-defeating practices are common in decision-making in health (Claxton et al., 2015a; Claxton et al., 2015b). Among the range of purpose-built tools for priority setting in health, generalized CEA is one of the only to encourage the explicit examination of such dilemmas.

Example: CEA of tuberculosis treatment

We illustrate how we might use CEA in considering the example of tuberculosis (TB) treatment in a population of 1,000,000 people, with a TB incidence rate of 100 cases per 100,000 population per year and an annual case fatality rate of 25% (see section preamble, p. XX).

A TB programme that reduces the case fatality of beneficiaries by 90% is currently achieving 40% coverage of those who stand to benefit, and there is a proposal in place to reach 50% coverage within the current year. The overhead costs for the programme are $10,000, and the additional costs per patient reached are $100. Using this information, we calculate the costs and effects for three scenarios represented in the columns of the Table.[8]

[8] The fixed overhead cost introduces a technical complication (i.e. "non-convexity") that is not discussed in the text. The essential concepts of CEA can nevertheless be applied.

Table. Current and projected costs and effects of a tuberculosis programme, for three scenarios.

	0% coverage	40% coverage	50% coverage
Population	1,000,000	1,000,000	1,000,000
Incident cases of tuberculosis per year	1,000	1,000	1,000
Cases treated at coverage level	0	400	500
Deaths from tuberculosis per year	250	160	140
Programme overhead costs	$0	$10,000	$10,000
Treatment costs at coverage level	$0	$40,000	$50,000

We calculate cost-effectiveness ratios (CERs) using the equation:

$$CER = \frac{\Delta Costs}{\Delta Effects}.$$

We calculate incremental and average cost-effectiveness ratios. Incremental cost-effectiveness ratios are calculated relative to the current position, while average cost-effectiveness ratios are calculated relative to the position of doing nothing (0% coverage).

The average cost-effectiveness ratio of the current programme, compared to no programme, is $556 per death avoided (= (($10,000 – $0) + ($40,000 – $0))/(250 – 160)). However, increasing coverage has an incremental cost-effectiveness ratio of $500 per additional death avoided (= (($10,000 – $10,000) + ($50,000 – $40,000))/(160 – 140)).

If we believe that we're already at the health-maximizing position, then we simply compare the incremental cost-effectiveness ratio of increasing coverage with the threshold previously determined. Suppose the threshold for our budget level has been estimated to be $550 per death avoided: we then increase coverage since the incremental cost effectiveness of the expanded programme is below the threshold (note that the cost-effectiveness ratio of the current programme is above the estimated threshold).

However, if we suspect we are not at the optimal position, then we should then compare the average cost effectiveness of the expanded programme with that of other technologies, including the current programme. By this logic, we should also expand coverage, since at $545 per death avoided ((= (($10,000 – $0) + ($50,000 – $0))/(250 – 140)), the average cost effectiveness of the expanded programme is superior to that of the current programme, demonstrating that the current programme was sub-optimal. However, the presence of fixed costs means the more the decision maker expands the programme, the more value for money she obtains, since she spreads the fixed costs over more people: given this fact, investment in the TB programme should therefore be made at full coverage (100%) or not at all.

---

3 Strategic planning and priority setting

Real-world decision-making

Since generalized CEA is highly stylized in comparison with actual decision-making, how can we better link theory and practice? Economists give the set of technologies determined by CEA a special name: it is the position of "allocative efficiency". The main purpose of CEA can be said to be to encourage movement, usually gradual, towards the efficient set of technologies. Information about the contents of the target portfolio is ultimately what allows actors in the health system to coordinate choices in the way discussed earlier in the chapter: as a given set of technologies requires a particular type of health system, with a particular set of fixed assets, so information about the optimal portfolio can influence many decisions, including notably those about investments in human resources and facilities with profiles matched to the requirements of the target portfolio. Priority setting is by definition strategic and context setting; discrete decision-making needs to be informed by priorities even if it cannot be wholly determined by them.

Strategic-planning

How can we bridge the gap between priority-setting in theory and decision-making in practice? The problem of the decision maker can again be restated: given a set of priorities, and a set of current activities, how can decisions be organized so that, over time, activities converge towards priorities? We refer to decision-making done with this perspective as *strategic planning*. Strategic planning is decision-making where the objectives have been determined through a prior priority-setting exercise (such as a generalized CEA). Once priorities have been established, relevant constraints to implementation, such as health-system bottlenecks (e.g. in facilities or human resources), can be explicitly taken into account during strategic planning.

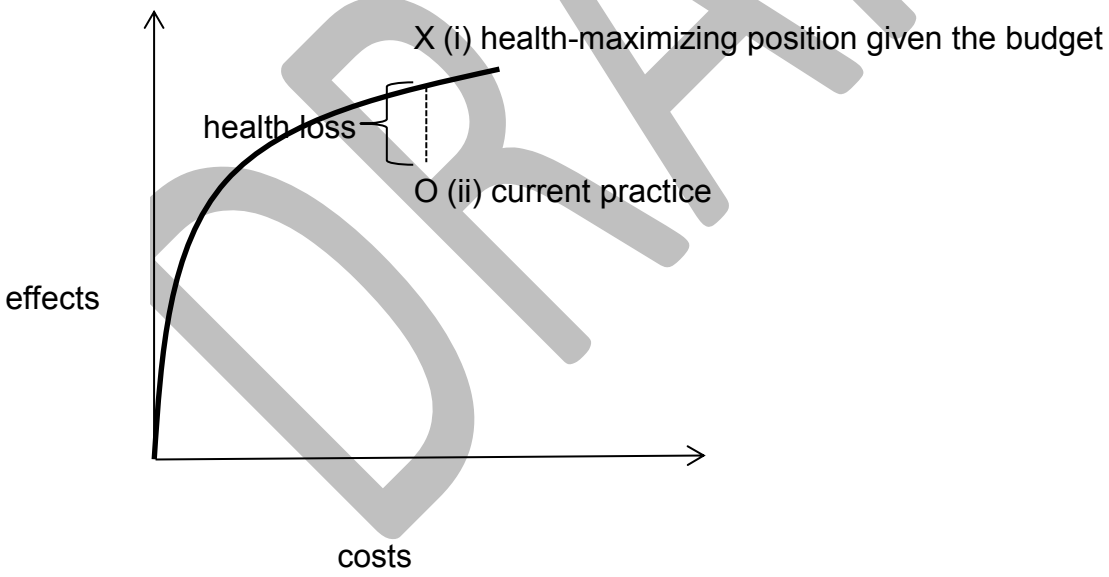

Figure 2. The health maximizing approach to efficiency

Thus, in practice, strategic planning compares two portfolios:

(i) health maximizing activities, and

(ii) current activities.

Optimal pathways

Given (i) and (ii), natural questions to ask are:

1. How much of the health loss in Figure 2 – i.e. the difference between the health level resulting from current practice and that which would be realized with the optimal portfolio – can be explained by appealing to other health-system objectives such as fairness (e.g. priority to the worse off) or financial risk protection?
2. How much can be set down to genuine barriers to implementation, such as having the wrong fixed assets, or not enough of them?
3. How much can be attributed to the presence of sub-optimal (but non-fixed) technologies in the current portfolio?

The answer to question 1 represents potentially justifiable deviations from efficiency (legitimate competing goals); the answer to question 2 shows areas where longer-term solutions need to be formulated (strategic plans); the answer to question 3 indicates the potential for short-term gains ("quick wins", such as by better sourcing of medicines and supplies).

Barriers such as inadequate human resources and facilities, or poor logistics and supply chains, represent fixed assets that may not be replaceable, transferable or created in the short term. These are sometimes called "non-budgetary constraints". Determining which non-budgetary constraints are most binding, and developing appropriate strategies to address these barriers, should be the main focus of strategic planning. Addressing the constraints posed by fixed assets, or "asset-specific investments", is essential for determining the best pathway towards improved efficiency over the longer term (van Baal et al., 2018).

If health systems have to contract, such as following the global financial crisis of 2007–2008 and the resurgence of austerity policies, the optimal pathway to improved efficiency might be to consider the implications of cost minimization (Figure 3). Note that the cost-minimizing position (Y, Figure 3) for a given level of health does not correspond to the health-maximizing position (X, Figure 2) for a given level of budget. In general, health maximization and cost minimization will yield a different optimum.

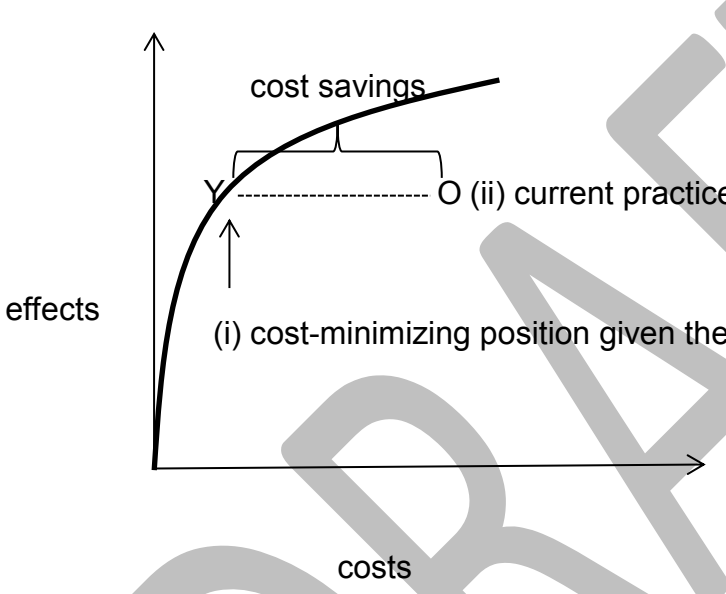

Figure 3. The cost minimizing approach to efficiency

Typically though, strategic planning looks at moving, not in one jump based on health maximization (up) or cost minimization (over, to the left), but rather, through a series of gradual steps, to positions of successively higher effects and costs (up and over, to the right). So in defining longer-term priorities such as those used for strategic planning, it is important to exclude the impact of non-budgetary constraints. In other words, at least for priority setting, the cost effectiveness of technologies should be evaluated with respect to their technically efficient implementation (i.e. without non-budgetary constraints), rather than as actually

implemented in a specific setting (i.e. with non-budgetary constraints). If non-budgetary constraints are allowed to influence the cost effectiveness of technologies, and therefore to determine the priorities for

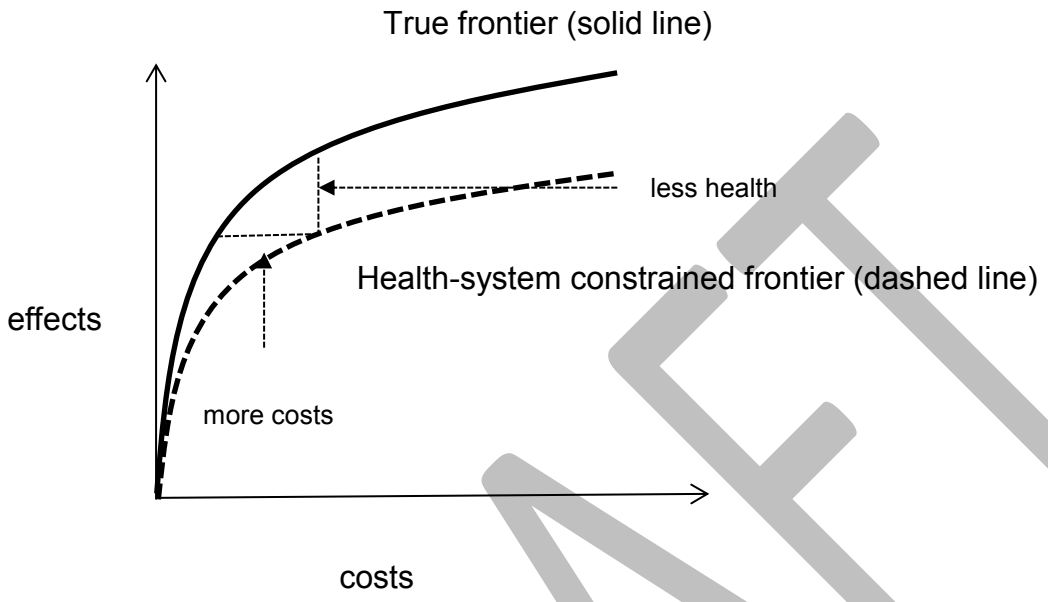

Figure 4. The self-imposed glass ceiling of health-system-constrained objectives

strategic planning, there's a risk that the health system will experience strategic failure (Figure 4). Non-budgetary constraints should be addressed through strategic planning, after priorities have been determined. Genuine priorities, not ones conditioned by poorly functioning health systems, are the proper objectives for strategic planning.

4 Conclusions

Priority setting in global health should be concerned with the efficiency of achieving health outcomes. Priority setting also needs to respect other recognized social objectives, such as fairness and financial risk protection (WHO, 2014). Generalized CEA is a means of directly

formulating the objective of maximizing health, taken here as the defining goal of health systems. Marginal CEA offers a potential shortcut but risks a "threshold trap" based on ignoring sub-optimal past practice, widespread misunderstandings about threshold rules, or the deliberate misuse ("hijacking") of thresholds by those with private interests.

It should go without saying that we should not conflate the concerns of CEA with a given set of techniques, methods, or tools. The principles and values of CEA (e.g. obtaining value for money) are more important than a specific technical approach. Finally, we have mentioned but not discussed in detail fairness and financial risk protection, nor have we elucidated how health-system goals are influenced by governance, institutional arrangements, and considerations of political economy. Some of the remaining chapters take up these subjects in more detail.

5 Acknowledgements

The authors thank many readers and reviewers for their helpful comments. Jeremy Lauer would like to thank Martin Chalkley of York University for particularly helpful conversations. Alec Morton would like to thank the staff of the University of Science and Technology of China for their hospitality while working on this paper, as well as the government of Anhui province for their support under the 100 Talents scheme.

6 References

Adler MD. Well-being and fair distribution: beyond cost-benefit analysis. Oxford, Oxford University Press, 2012.

Bertram MY, Lauer JA, De Joncheere K, Edejer T, Hutubessy R, Kieny M-P, Hill S. Cost-effectiveness thresholds: pros and cons. Bulletin of the World Health Organization, 2016; 94:925-930.

Claxton K (a), Martin S, Soares M, Rice N, Spackman E, Hinde S, et al. Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold. Health Technology Assessment, 2015;19(14):1-503, v-vi.

Claxton K (b), Sculpher M, Palmer S, Culyer AJ. Causes for concern: is NICE failing to uphold its responsibilities to all NHS patients? Health Economics, 2015;24(1):1-7.

Coase R. The Nature of the Firm. Economica, 4(16):386–405, 1937.

Jamison DT. Cost-effectiveness analysis: concepts and applications. In Detels R, McEwen J, Beaglehole R, Tanaka H (eds.) Oxford Textbook of Public Health: Volume 2, The Methods of Public Health, fifth edition. Oxford: Oxford University Press, 2009, pp. 767-782.

Luenberger DG. Introduction to linear and non-linear programming. Addison Wesley, 1973.

MacAskill W. Doing good better, Avery, 2015.

Marseille E, Larson B, Kazi DS, Kahn JG, Rosen S. Thresholds for the cost–effectiveness of interventions: alternative approaches. Bulletin of the World Health Organization, 93:118-124, 2015.

Sculpher M, Revill P, Ochalek JM, Claxton K. How Much Health for the Money? Using Cost-Effectiveness Analysis to Support Benefits Plan Decisions. In Glassman A, Ursula Giedion U, Smith PC (eds.) What's in, What's out? Designing Benefits for Universal Health Coverage, Center for Global Development, 2017.

van Baal, PH, Morton A, Severens JL. Health care input constraints and cost effectiveness analysis decision rules. Social Science and Medicine, Social Science and Medicine, 200: 59-64, 2018.

Weinstein M, Zeckhauser R. Critical ratios and efficient allocation. Journal of Public Economics, 2:147-157, 1973.

WHO. Making fair choices on the path to universal health coverage: Final report of the WHO Consultative Group on Equity and Universal Health Coverage. Geneva, World Health Organization, 2014.