Non-linearity in the cost-effectiveness frontier

Joanne Lorda,*, George Lakingb and Alastair Fischerc
a National Institute for Health and Clinical Excellence (NICE), London, UK
b Cancer Research-UK PET Oncology Group, Manchester Molecular Imaging Centre, Christie Hospital, UK
c Department of Community Health Sciences, St. George's University of London, London, UK

Summary

Conventional cost-effectiveness decision rules rely on the assumptions that all health care programmes are divisible and exhibit constant returns to scale for a homogeneous population; hence, the path between adjacent programmes on a cost-effectiveness frontier must be linear. In this paper we build a framework to analyse non-linear ‘expansion’ paths. We model the impact of two key sources of non-linearity: economies of scale or scope in the production of health care; and prioritisation of patients who are most likely to benefit from more expensive and more effective treatments. We conclude that the expansion path might be linear, convex or concave, depending on the situation. The path might also exhibit vertical discontinuity due to fixed costs or horizontal discontinuity due to indivisibility. The efficiency of resource allocation might be improved by empirical estimation of expansion paths. We discuss the advantages and disadvantages of this approach compared with a standard stratified analysis. Copyright © 2006 John Wiley & Sons, Ltd.

Keywords cost-effectiveness analysis; resource allocation; non-linearity; decision rules; expansion path

Introduction

The conventional decision rules of cost-effectiveness analysis (CEA) [1] rest on the assumption that health care programmes are all divisible with constant returns to scale for a homogeneous population. Thus, the cost-effectiveness frontier, which is the locus of non-dominated programmes in cost-effectiveness space for some given patient population, must be piecewise linear. Provided that options are excluded by ‘extended dominance’ as well as by ‘simple dominance’, the frontier must also be convex to the effectiveness axis. The amount that the healthcare system is able or willing to pay for a unit of effectiveness (the ‘threshold’ cost-effectiveness ratio) determines how far along this frontier it is efficient to go, and hence which programme or linear combination of adjacent programmes should be offered to patients. With knowledge of the threshold, we can appraise individual programmes by consideration of their incremental cost-effectiveness ratio (ICER), calculated with reference to the previous programme on the frontier. Thus, a programme would be recommended only if its ICER were less than the threshold.

This ‘threshold rule’ greatly simplifies the task of resource allocation based on cost-effectiveness information. However, some commentators [2–7] have criticised this approach. They object to all three assumptions – (1) knowledge of the correct level of the threshold to achieve optimality, (2) divisibility and (3) constant returns – arguing that these are unlikely in practice, and that relaxing

*Correspondence to: NICE, MidCity Place, 71 High Holborn, London WC1V 6NA, UK. E-mail: joanne.lord@nice.org.uk

Received 22 April 2003
Accepted 27 October 2005

Copyright © 2006 John Wiley & Sons, Ltd.
them will have important repercussions. The question of how to calibrate the threshold is clearly vital, but this is receiving attention elsewhere [8]. The aim of this paper is to address the third assumption of linearity. We develop a framework to analyse the potential causes of non-linearity in cost-effectiveness frontiers and to evaluate their impact. We also briefly touch on the second assumption of divisibility, noting that this can be treated as a special case of non-linearity.

Sources of non-linearity may be divided into two main categories: (1) economies or diseconomies of scale or scope in the provision of health care [9]; and (2) prioritisation of patients for treatment [10]. We review the first of these categories in the following section, and use a simple model to examine the impact of non-linear cost functions. We note that fixed costs can introduce vertical discontinuities to a cost-effectiveness frontier, and that the expansion path between adjacent options on a frontier will be concave with diminishing marginal costs and convex with increasing marginal costs. Economies or diseconomies of scale might also affect the curvature of the expansion path.

The second major source of non-linearity, patient prioritisation, is then discussed. Individual patients are likely to differ with regard to the expected costs and health consequences of any given intervention. Even so, if patients were to be selected in random order to be treated with the more expensive but more effective option, the locus of expected costs and effects would still be linear. However, if patients with a greater expected net benefit were to be prioritised, we show that the cost-effectiveness expansion path between two alternative programmes will be convex. Conversely, with ‘perverse’ selection criteria (prioritisation of patients with a lower expected net benefit) the expansion path could be concave.

This idea goes against a basic principle of cost-effectiveness analysis, which is that frontiers should always reflect treatment options for a given homogeneous group of patients [1]. Conventionally, health economists have argued for stratified analysis in which discrete subgroups are shown on different cost-effectiveness frontiers [11, 12]. It might be straightforward to segment major disease categories, but within these categories there is likely to be a gradation of expected costs and effects, conditional on a wide variety of diagnostic indicators. We build a model to examine the effect of ‘residual’ patient heterogeneity (that is, heterogeneity that remains after division into subgroups), subject to defined patient selection criteria.

The general shape of the expansion path cannot be known a priori. It depends on the health care production and cost functions, the extent of residual patient heterogeneity and the expected order in which patients are prioritised for treatment.

We start the Discussion section by summarising the expected impact of these various factors on the shape of expansion paths. We then discuss the implications of non-linear paths for resource allocation decisions. In the case of convexity, information about the shape of the expansion path could help to improve on the efficiency of simple comparison of an ICER with a cost-effectiveness threshold. Finally, we discuss the potential advantages and disadvantages of expansion path estimation relative to stratified analysis.

### Economies of scale and scope

#### Cost function model

We start by considering the possible impact of supply-side factors. Suppose that each member of a population of N patients must receive one and only one of two interventions A or B, which are consecutive interventions on a cost-effectiveness frontier. Let the mean costs and mean effects of B be greater than those of A. Interventions A and B are said to be ‘mutually exclusive’, since individual patients cannot receive both. Let n be the number of patients who receive B, so (N–n) must receive A. We assume that A and B are ‘divisible’, in the sense that n can take on any integer value between 0 and N. For the moment we also assume that there is no variation between patients in the costs or effects of the treatments.

The total cost of providing intervention B, TC_B(n), is a positive and increasing function of n. Initially, we suppose that there are no fixed costs for B (TC_B(0) = 0) and that marginal costs are constant, so that TC_B(n) is linear. The total cost of A is a positive increasing function of (N–n), and hence a positive decreasing function of n: TC_A(n). Again, we assume that there are no fixed costs (TC_A(N) = 0), and that marginal costs are constant (TC_A(n) is linear). Full implementation of B
is more expensive than full implementation of A so \( TC_B(N) > TC_A(0) \). These simple linear cost functions are illustrated in the left-hand side of Figure 1(a).

The total cost of treating \( n \) patients with B and \((N-n)\) patients with A is \( TC(n) \), which is the vertical sum of \( TC_A(n) \) and \( TC_B(n) \). Since both of these functions are linear, \( TC(n) \) must also be linear. The slope of the total cost function is determined by the net impact of a marginal increase in B coupled with a marginal decrease in A, which must be positive given our assumptions.

Similarly, we define a total effect function \( TE(n) \), which gives the total health gain (measured in QALYs, say) obtained from treating \( n \) patients with B and \((N-n)\) patients with A. Since B is more effective on average, \( TE \) increases with \( n \). For now, we also assume that \( TE(n) \) is linear.

Finally, we define the ‘expansion path’ as the locus of total costs and total effects as \( n \) varies from 0 to \( N \), hence \( TC(n) = f(TE(n)), 0 \leq n \leq N \). Clearly, if the total cost and effect functions are both linear, then the expansion path must also be linear, as illustrated in the right-hand side of Figure 1(a). The ICER for B in relation to A is the slope of this line.

**Fixed costs and irreversibilities**

Now consider the possibility of fixed costs of production for one or both of our interventions. If A and B share a common need for some fixed infrastructure, the whole total cost functions \( TC_A(n), TC_B(n) \) and \( TC(n) \) will just be shifted vertically by the same distance. Hence, the expansion path will still be linear and continuous, with the same slope as in Figure 1(a).

However, suppose that there are fixed costs for intervention A that are not common to intervention B, or vice versa. In this case, the fixed cost of providing a mixture of both interventions will be greater than that of providing either intervention on its own. Suppose that there are fixed costs \( FC_A \) and \( FC_B \) for interventions A and B, respectively, that are not shared. The total cost functions are
shifted vertically, as illustrated in the left-hand side of Figure 1(b): \( TC_A(n) \) shifts up by \( FC_A \), \( TC_B(n) \) shifts up by \( FC_B \) and \( TC(n) \) shifts up by \( FC_A + FC_B \), except for the two end points \( (n = 0 \) and \( n = N) \), where only one of the fixed costs is incurred \( (TC(0) = TC_A(0) \) and \( TC(N) = TC_B(N)) \). In Figure 1(b) we have assumed that \( FC_A > FC_B \), but that \( TC \) still increases with \( n \). (If the fixed costs of \( B \) were very large in relation to those of \( A \), it is possible that \( TC \) could be constant or decline with increasing \( n \). If \( FC_A > FC_B \), however, \( TC \) must increase with \( n \).

This discontinuous total cost function gives rise to a discontinuous expansion curve in cost-effectiveness space, as shown in the right-hand side of Figure 1(b). If we start at point \( A \), where all \( N \) patients have intervention \( A \), we do not incur the fixed cost for intervention \( B \). But as soon as one patient is transferred to \( B \), we shift up to point \( A^+ \). As more patients are transferred, we move along the line from \( A^+ \) to \( B^- \). When the last patient is transferred, we no longer need to pay the fixed costs for intervention \( A \), and hence jump down to point \( B \). However, the move down from \( B^- \) to \( B \) might be delayed, as it may take some time to release assets or to retrain staff previously used to produce \( A \). So in the shorter term, we may remain close to \( B^- \). Similarly, moving in the other direction, the shift back from \( A^+ \) to \( A \) might be subject to a time delay. If so, then we should take this possible cost into consideration before deciding to introduce technology \( B \) in the first place. Prior expected costs of such irreversibilities could be incorporated by taking a 'real options' approach [13, 14].

**Diminishing/increasing marginal costs**

So far we have only considered cost functions with constant marginal costs. However, marginal costs may vary with the scale of production. Suppose that marginal costs were diminishing for both interventions over the whole relevant range of production, as illustrated by the concave cost curves in the left-hand side of Figure 1(c). Note that the slope of \( TC_B(n) \) declines as \( n \) increases, but that the (negative) slope of \( TC_A(n) \) increases with \( n \). In this situation, the total cost function \( TC(n) \), and hence the cost-effectiveness expansion curve, must be concave, as illustrated in Figure 1(c) (still assuming linear effects and fixed costs as before). (If the curvature of \( TC_A(n) \) and/or \( TC_B(n) \) is sufficient, the total cost function might rise to a maximum at some \( 0 < n < N \). If so, the cost-effectiveness expansion curve will also be non-monotonic, with a cost maximum at the same value of \( n \).)

Conversely, increasing marginal costs would give rise to convex cost curves and a convex expansion curve (which may or may not be monotonic). An obvious extension of this analysis would allow economies of scale for low levels of production, and diseconomies for high levels of production.

**Economies of scope**

If the production processes for \( A \) and \( B \) are inter-related, there are said to be economies (or diseconomies) of scope. Economies of scope would tend to decrease the curvature of concave cost functions and expansion paths, such as those illustrated in Figure 1(c). Pronounced economies of scope could even compensate for what would otherwise be diminishing marginal costs, yielding convex cost and expansion curves. Economies of scope coupled with increasing returns to scale must result in convex cost and expansion curves.

**Patient heterogeneity and selection**

**Patient selection model**

We now extend our model to examine the impact of heterogeneity in the patient population coupled with non-random prioritisation. The costs and effects of treatments inevitably vary between individuals. Some patients would attain a greater health gain with treatment \( B \) than would other patients. Suppose that doctors knew that only some patients could be given the (on average) more effective but more expensive option \( B \). The doctors would be likely to select those patients who they thought would benefit most from receiving \( B \). If they were better than random in their selection, the overall health gain would be greater. Hence, the total effects, and similarly the total costs, depend not just on the proportions of patients treated with \( A \) or with \( B \), but also on which individuals are selected for which treatment. In essence, all populations are heterogeneous, and
can be ranked for prioritisation of treatment according to triage rules. These prioritisation rules impact on the shape of the cost-effectiveness expansion path.

Suppose that $E_A$ and $E_B$ are random variables reflecting health outcomes for individual patients under intervention $A$ or $B$, respectively. Similarly, let $C_A$ and $C_B$ be random variables reflecting patients' costs with these interventions. We assume that the effect and cost variables are continuous (measured in QALYs and currency units, respectively). The correlation between the effects of the two treatments may be positive or negative, depending on the nature of the population and the interventions. Thus, it might be that patients who do well on $A$ also do well on $B$, or that patients who do well on $A$ do less well on $B$. Similarly, correlations between the costs and effects of each treatment, $\text{Corr}(E_A, C_A)$ and $\text{Corr}(E_B, C_B)$, may be positive or negative. They will be negative where patients with a poor health outcome tend to make greater subsequent use of health services. However, the causation could sometimes run the other way, where patients who make more use of health services gain a greater health improvement, resulting in a positive correlation between costs and effects. The direction and size of the correlations depend on the characteristics of the disease, the population, the treatments and the wider health system.

Further, suppose that there is some diagnostic index $D$, which reflects the expected order in which patients will be selected for the more expensive intervention $B$. Suppose that patients with a higher value of $D$ will be transferred to $B$ first. $D$ might be discrete or continuous. It might be a measure of a single patient characteristic, such as age or blood pressure, or it might be an index reflecting a complex combination of demographic, socio-economic, physiological, clinical and/or genetic indicators. In the latter case, $D$ might be derived from a statistical analysis or holistic judgement by clinicians. Note that the cost of collecting and analysing information used to construct $D$ should be included in the cost of the interventions.

The five random variables in our model may be characterised by some joint probability density function $g$ with parameter set $\theta$. $\{D, E_A, E_B, C_A, C_B\} \sim g(\theta)$. The vector $(d_i, e_{Ai}, e_{Bi}, c_{Ai}, c_{Bi})$ is a realisation of this set of variables for an individual patient $i$ ($i = 1, 2, \ldots, N$), where the patients have been ranked in order of decreasing $D$. Hence, patients with a lower index $i$, and higher value of $D$, will be transferred from $A$ to $B$ first. We assume that patients are randomly ordered in the case of a tie ($d_i = d_l$).

The total costs and effects of treating $n$ patients with $B$ are given by the cumulative sums $\text{TC}_B(n) = \sum_{i=1}^n e_{Bi}$ and $\text{TE}_B(n) = \sum_{i=1}^n e_{Bi}$, respectively. These functions are both positive and increase with $n$, so long as $e_{Bi}$ and $e_{Bi}$ are positive for all $i$. However, if costs or effects are negative for some patients, $\text{TC}_B$ and/or $\text{TE}_B$ are not necessarily monotonic. Similarly, the total costs and effects of treating the remaining $(N-n)$ patients with $A$ are $\text{TC}_A(n) = \sum_{i=n+1}^N e_{Ai}$ and $\text{TE}_A(n) = \sum_{i=n+1}^N e_{Ai}$, respectively. These are positive decreasing functions of $n$ if $e_{Ai}$ and $e_{Ai}$ are positive for all patients. We know that $\text{TC}_B(N) > \text{TC}_A(0)$ since the total cost of treating all $N$ patients with $B$ is greater than that of treating all $N$ patients with $A$, and similarly $\text{TE}_B(N) > \text{TE}_A(0)$.

The total costs and effects for the whole group of $N$ patients are given by the sums $\text{TC}(n) = \text{TC}_A(n) + \text{TC}_B(n)$ and $\text{TE}(n) = \text{TE}_A(n) + \text{TE}_B(n)$, respectively. These total cost and effect functions trace out an expansion curve in cost-effectiveness space as before.

Interestingly, this model may also be used to determine the optimum diagnostic threshold based on a receiver operating characteristic (ROC) curve [15–17]. We develop this possibility in another paper [Laking G, Lord J, Fischer A. The economics of diagnosis. Health Econ 2006; submitted for publication].

In the following sections, we explore the implications of this general model. For simplicity, we start by taking $D$ to be a binary variable $D = \{0, 1\}$. Thus, the total population of $N$ patients is divided into two subgroups: group $X$ with $D = 1$ and group $Y$ with $D = 0$. Suppose that $n_X$ is the number of patients in group $X$. The $X$ patients will be transferred to intervention $B$ first, since they have a higher value of $D$. Only when all the type-$X$ patients are on $B$, will the $(N-n_X)$ type-$Y$ patients be transferred. We also assume that $N$ is large, so that the various cost and effect functions become smooth in the limiting case.

**Linear cost functions**

Suppose that the cost curves $\text{TC}_A(n)$, $\text{TC}_B(n)$ and $\text{TC}(n)$ are all continuous and linear as in Figure 1(a). For this, there must be no economies of scale or scope on the supply side, and no trends in
individual patient costs as $n$ increases ($C_A$ and $C_B$ are not correlated with the patient selection variable $D$).

If patients are completely homogeneous in terms of outcomes, the effect functions and the expansion curve must also be linear, as in Figure 2(a). Even with outcome heterogeneity, if there is no correlation between the patient selection variable $D$ and one or both of the effect variables $E_A$ and $E_B$ the total effect and expansion curves will still be linear (or at least be approximately so for large $N$).

With a positive correlation between $D$ and $E_B$ and no correlation between $D$ and $E_A$, the situation is as illustrated in Figure 2(b). Here, the type-X patients, who are prioritised for $B$, are expected to do better than the type-Y patients on treatment $B$. This means that the slope of the TE$_B(n)$ curve must be greater for $n \leq n_X$ than for $n > n_X$. This curve is piecewise linear, since patients are selected in random order within each of the two subgroups. The type-X patients do not necessarily all derive the same costs and effects from treatment $B$, but we have no information to indicate which of them should be treated first, since they all share the same value of $D$. Hence, the segment of TE$_B(n)$ for $n \leq n_X$ is linear. Similarly, the inability to prioritise within group Y means that TE$_B(n)$ is linear for $n > n_X$. Both groups are expected to do equally well on treatment $A$, so the TE$_A(n)$ curve has constant slope throughout the whole range of $n$. Consequently, the total effect curve TE($n$) must be piecewise linear and concave, and the expansion curve must be piecewise linear and convex, as in Figure 2(b).

The result will be similar when there is a negative correlation between $D$ and the effect of intervention $A$ but no correlation between $D$ and the effect of $B$, as in Figure 2(c). In this case, the type-X patients, who are given priority for transfer, are those who are expected to do worse on $A$ and equally well on $B$. Here too, the expansion curve will be piecewise linear and convex.

In practice, it is likely that the effects of the two interventions will be correlated. If effects are negatively correlated and if the patients who do better on $B$ are selected first, the expansion curve will be convex. However, if effects are positively correlated and patients who do better on $B$ are selected first, the expansion curve might be convex or concave. Nevertheless, provided that $D$ is positively correlated with the incremental effects ($E_B - E_A$), the expansion curve must still be convex. This will be so if clinicians seek to prioritise patients with a greater expected health gain from the more expensive treatment, and are at least partly successful in this endeavour. It is possible that clinicians could adopt ‘perverse’ selection

Figure 2. Impact of patient heterogeneity and non-random selection (assuming linear costs and no correlation between the effects of intervention $A$ and $B$): (a) Random selection or no heterogeneity; (b) prioritisation of patients with better expected outcomes on intervention $B$ using a binary diagnostic marker; and (c) prioritisation of patients with worse expected outcomes on intervention $A$ using a binary diagnostic marker

criteria, based on some variation of the Rule of Rescue, say. If so, we would expect a concave expansion curve.

**Correlation between costs and effects**

The situation is more complicated if costs and effects are correlated. We continue to assume that there are no economies of scale or scope from the supply side, and that patients with a higher expected health gain are prioritised for treatment $B$. However, we now assume that costs for individual patients are related to clinical outcomes.

If costs are negatively correlated with effects, $(\text{Corr}(E_A, C_A) < 0$ and $\text{Corr}(E_B, C_B) < 0)$ the curvature of the convex cost-effectiveness expansion curve must be increased. This is illustrated in Figure 3(a) for a case where effects are negatively correlated. Here the health gain from $B$ is greater for the type-X patients than for the type-Y patients, but given our assumptions the incremental cost of $B$ must be lower for the X’s. Hence, the ICER, the slope of the expansion curve, must be lower for $n_x < n_X$ than for $n > n_X$.

Conversely, if costs and effects are positively correlated $(\text{Corr}(E_A, C_A) > 0$ and $\text{Corr}(E_B, C_B) > 0)$, the curvature of the otherwise convex expansion curve will be reduced, possibly to the extent that it could become concave. Though the health gain from $B$ is greater for the X’s than for the Y’s, the additional cost that they incur also tends to be greater. Thus, it is unclear whether the additional cost per QALY gained is higher or lower for the X’s.

In the above discussion, we have assumed that the selection criteria for the more expensive intervention are based entirely on expected clinical effects. It is possible that clinicians could also use economic selection criteria, including expectations of costs in their decision-making process. However, this seems less likely [18]. Thus, although correlations between diagnostic indicators $D$ and costs are likely to exist, they are probably best seen as a consequence of the correlations between $D$ and effects, and between the effects and costs. Similarly, any correlations between the costs of different treatments are likely to be a consequence of heterogeneity in patient characteristics, as indicated by a marker $D$, causing different responses to the treatments, and hence different costs for follow-up care.

**Continuous diagnostic variable**

Extension of the above analysis for three or more subgroups is straightforward. For each additional value of the discrete diagnostic variable $D$ there is an additional subgroup of patients, and an

---

Figure 3. Impact of patient heterogeneity and non-random selection (with negative correlation between effects of $A$ and $B$ and negative correlation between costs and effects): (a) Prioritisation of patients with better expected outcomes on intervention $B$ using a binary diagnostic marker; and (b) prioritisation of patients with better expected outcomes on intervention $B$ using a continuous diagnostic marker

additional segment on each cost, effect and expansion curve. At the extreme, as $D$ becomes a continuous variable and $N$ approaches infinity, the cost, effect and expansion curves approach smooth and continuous curves, as illustrated in Figure 3(b). We might generally expect the expansion curve to be increasing in slope as shown due to more or less efficient prioritisation. However, it is not necessarily monotonic. It could have a negative slope near $B$, for example, if there is some subgroup of patients for whom $B$ is dominated by $A$.

**Combining supply side and patient selection effects**

Introduction of supply-side economies of scale and/or scope to the above model might further modify the position, curvature and/or continuity of the cost curves, and hence of the cost-effectiveness expansion path. To conceive of such factors as an extension of our patient selection model, it might be helpful to differentiate two types of cost. First, there are the upfront costs of providing the interventions, which could be assumed not to vary between individual patients. These could be modelled using a conventional cost function. Second, there are the costs of follow-up care, which vary between individual patients and are correlated with the patients’ outcomes as in the above patient selection model.

**Expansion curve estimation**

Expansion curves could be estimated using a modelling approach. This has been done by Marshall and Rouse for alternative strategies for the primary prevention of cardiovascular disease [19]. They estimated what we have called expansion curves convex to the effectiveness axis in cost-effectiveness space due to prioritisation of patients with a greater estimated risk of coronary heart disease.

Expansion curves could also be estimated directly from clinical trial data. If a crossover trial were possible, the vector $\{D, E_A, E_B, C_A, C_B\}$ could be directly observed for a sample of patients. More usually, the vector $\{D, E_A, C_A\}$ would be observed for a sample of ‘control’ patients and $\{D, E_B, C_B\}$ for a different sample of ‘intervention’ patients. Such two-sample data could still be used to estimate an expansion curve by effectively using $D$ as a matching variable [20].

The impact of second-order uncertainty over the location of the expansion curve could also be estimated using a Monte Carlo simulation or non-parametric bootstrapping. Instead of a single point on the cost-effectiveness plane, each iteration of the probabilistic sensitivity analysis would yield a separate estimate of the whole expansion curve. Thus, we would obtain a banana-shaped 95% probability region in cost-effectiveness space, reflecting uncertainty over the location and shape of the expansion curve. Instead of a single cost-effectiveness acceptability curve, we would have a family of such curves, one for each level of implementation of treatment $B$.

**Discussion**

**Possible shapes of expansion paths**

It is conventional to assume that an expansion path between adjacent options on a cost-effectiveness frontier is continuous and linear (as in Figure 4(a)). We have argued that this is not necessarily the case. Some possible sources of non-linearity and discontinuity in cost-effectiveness expansion curves are summarised in Table 1.

An expansion path could be concave (as in Figure 4(b)) if there are diminishing marginal costs for one or both of the technologies. In theory, concavity could also result from diseconomies of scope, or from the use of ‘perverse’ patient selection criteria. The most likely reason for convexity of the expansion path (as in Figure 4(c)) is prioritisation of patients with a greater expected incremental net benefit. However, a convex curve could also result from increasing marginal costs for one or both interventions, and/or from economies of scope between the interventions.

The cost-effectiveness expansion paths illustrated in Figures 4(a)–(c) are continuous and smooth with constant, continuously increasing or continuously diminishing slope over the interval $[A,B]$. However, expansion paths are not necessarily so well behaved. For example, they may sometimes have points of inflexion or points (other than $A$ or $B$) where costs or effects reach a maximum or minimum. Discontinuity is also a possibility, as discussed below.
In the presence of fixed costs of production, which are not shared by both interventions, vertical discontinuities may be introduced at the ends of the expansion path (as in Figures 4(d)–(f)). If fixed costs are irreversible, the finishing point might depend on the start point (a form of hysteresis). Thus, it might not be possible to reach point B starting from point A, or vice versa.

We have not considered ‘pure’ indivisibility (horizontal discontinuity of the expansion curve) in this paper, but this is also a possibility. For instance, it might be thought inequitable to give similar patients different interventions, so that we are restricted to point A or B. In reality, clinical practice and treatment recommendations often do result in the use of different treatments.

Table 1. Summary of factors that could influence the shape of the expansion curve

<table>
<thead>
<tr>
<th>Impact on expansion curve</th>
<th>The nature of production and cost functions</th>
<th>Patient heterogeneity and non-random selection</th>
</tr>
</thead>
<tbody>
<tr>
<td>Concavity</td>
<td>(1) Diminishing marginal costs</td>
<td>(3) Prioritisation of patients with lower incremental net benefit (perverse)</td>
</tr>
<tr>
<td></td>
<td>(2) Diseconomies of scope</td>
<td></td>
</tr>
<tr>
<td>Convexity</td>
<td>(4) Increasing marginal costs</td>
<td>(6) Prioritisation of patients with higher incremental net benefit</td>
</tr>
<tr>
<td></td>
<td>(5) Economies of scope</td>
<td></td>
</tr>
<tr>
<td>Vertical discontinuity</td>
<td>(7) Fixed costs that are not common to programmes</td>
<td></td>
</tr>
<tr>
<td>Horizontal discontinuity</td>
<td>(8) Production indivisibilities</td>
<td>(9) Equity considerations</td>
</tr>
</tbody>
</table>

Figure 4. Impact of the shape of the expansion path on the optimal decision: (a) Linear, no fixed costs; (b) concave, no fixed costs; (c) convex, no fixed costs; (d) linear, fixed costs; (e) concave, fixed costs; and (f) convex, fixed costs.
for different subgroups of patients. For instance, the National Institute for Health and Clinical Excellence (NICE) frequently advocates restriction of technologies according to clinical or other patient characteristics [21]. Thus, ‘vertical equity’ often trumps ‘horizontal equity’. Pure indivisibility is unlikely to arise from restrictions on the supply side. For any thing other than public goods, it is difficult to conceive of a situation where it is not possible to provide an intervention for one individual without providing it for all individuals. It may well be inefficient to do so, due to high fixed costs for example, but this does not constitute a true indivisibility, and can be analysed using our expansion curve concept as discussed below.

**Implications for resource allocation**

We now consider the value of expansion path estimation for resource allocation.

We assume that the cost-effectiveness threshold is known and ‘correctly calibrated’, in the sense that it reflects the opportunity cost of health care resources within the health sector: it is equal to the ICER for the ‘marginal’ project. Thus, we are working within an ‘extra-welfarist’ paradigm, in which the objective is to maximise the health effects subject to a fixed budget constraint [22–24]. Also note that we are working from a ‘utilitarian’ position, in the sense that we only care about the sum of health effects, and not their distribution. We further assume that resource allocation decisions are made purely on the basis of the expected values of costs and effects [25,26].

Given a choice of two mutually exclusive, non-dominated options, the ‘threshold rule’ says to implement the more expensive more costly option if and only if its ICER is less than or equal to the threshold. As noted above, Johannesson and Weinstein demonstrated that this is optimal, assuming divisibility and linearity [1].

Application of the threshold rule can be illustrated graphically using net benefit isoquants (NB₁, NB₂ and NB₃ in Figure 4) [27]. These show combinations of cost and effect that yield a fixed level of net benefit. The slope of these lines is the cost-effectiveness threshold, and those that lie further to the southeast yield a greater level of net benefit and are preferred. We assume that the net benefit isoquants are approximately linear. This is appropriate, at least in the short term, for a large payer organisation if the cost and health gain of each individual programme is relatively small in relation to its total ‘portfolio’ of programmes.

With a linear frontier segment (Figure 4(a)), as assumed by Johannesson and Weinstein, the highest attainable net benefit will almost always be achieved at one or other extreme (in this case at B). The only exception might occur when the slope of the expansion path is equal to the threshold: B is the marginal programme, which may not be possible to fully fund.

Now suppose that the expansion path is concave (as in Figures 4(b), (d) or (e)). Here the optimal point must always lie at an end point (at B in the illustrated examples). Partial implementation will not be cost-effective, and hence estimation of the expansion path will not be productive. The result will usually be the same as that identified by the simple ICER threshold rule. An exception might occur if there are irrecoverable costs, as in Figures 4(d) and (e), in which case the best attainable point might lie somewhere between A and A⁰, or between B and B⁰.

Estimation of an expansion path is likely to be useful, however, where it is convex (as in Figures 4(c) or (f)). Here, maximum net benefits might be attained with a mix of the two programmes – as at B⁰ in Figures 4(c) and (f), where the slope of the expansion path is equal to the cost-effectiveness threshold [Lord J, Laking G, Fischer A. Health care resource allocation: Is the threshold rule good enough? J Health Serv Res Policy 2006; in press]. Simple application of the ICER threshold rule in this situation might be misleading. In the cases illustrated in Figures 4(c) and (f), it would lead us to point B, with a lower net benefit than B⁰.

In summary, expansion path estimation may be worthwhile where there is scope for substantial prioritisation of patients with higher incremental net benefits, causing convexity in the expansion path. This convexity, however, could potentially be outweighed by supply-side factors, such as diminishing marginal costs or large fixed costs.

**Expansion path estimation vs stratified analysis**

Explicit estimation of non-linear expansion/contraction curves might thus be used to improve the efficiency of resource allocation rules. But is this necessary? Could we achieve the same end by a
more careful segmentation of our patient populations?

The conventional approach when faced with two identifiable groups of patients, X and Y, who are expected to respond differently to alternative interventions, is to consider each group on a separate cost-effectiveness frontier. Thus, each segment of the frontier in Figure 3(a) (\(AB_X\) and \(B_YB\)) would be treated as a separate, linear frontier. In fact, this would yield exactly the same result as taking both groups together in the same frontier.

The choice between expansion curve estimation and stratified analysis is largely a matter of analytical convenience. Segmentation may be straightforward for major disease categories. Indeed, it is essential that groups who are candidates for distinct sets of tests and treatments be shown on different cost-effectiveness frontiers. If there are then a relatively small number of readily distinguished subgroups, segmented analysis will be manageable and easy to communicate. Coyle et al. have shown how a systematic stratified analysis can be used to estimate the opportunity cost of failure to take advantage of information about subgroup heterogeneity [12].

However, the identification of distinct subgroups might not always be easy. At the extreme, each of the \(N\) patients might comprise a separate subgroup. A pragmatic response to this reductio ad absurdum would be to segment the population by dividing some diagnostic index \(D\) into intervals, such that each subgroup has an approximately linear expansion path. However, it may not be evident how many segments should be used, nor how the cut-points should be established.

Empirical estimation of the expansion path could be used to identify an (ex ante) optimal cut-point for \(D\). Consider the situation illustrated in Figure 4(c). The optimal point on this convex expansion path is \(B^*\), where the slope is equal to the threshold. From this point we can identify the optimal proportion of patients \(n^*/N\) to be given intervention \(B\) (as illustrated in Figure 3(b)). Hence, the optimal cut-point \(D^*\) for the diagnostic variable can be determined.

Note that the expansion path is conditional on the diagnostic variable \(D\) used to select patients – different selection criteria will give rise to different expansion paths. Hence, we are no longer evaluating just one treatment versus another, but instead we are comparing protocols of care, including diagnostic or risk assessment criteria for patient selection. For the purposes of resource allocation, \(D\) should be chosen to reflect beliefs about the selection criteria that clinicians will use in the real world. Thus, the expansion path is intended to be predictive of expected costs and effects given different scales of provision. However, the concept of expansion paths could also be used in a normative fashion to evaluate different selection criteria – with better criteria giving a more convex path and a greater net benefit (Laking et al., 2006).

Information about economies of scale or scope in the production of the technologies is also more easily included in an expansion curve analysis. Conventional econometric techniques may be used to estimate cost functions for the interventions under review. These may then be combined with patient-level information on health outcomes and follow-up costs from clinical trial data. In contrast, integration of supply-side effects in a subgroup analysis is more complicated. Suppose that there are fixed costs for one or both of our technologies. Without information about the order in which the subgroups will be transferred to the more expensive treatment, it is not clear how to distribute the fixed costs between groups. The incremental costs for one subgroup are thus dependent on provision for the other subgroups.

Explicit estimation of an expansion path also has potential statistical advantages over a stratified analysis. As the number of strata increases, the sample sizes fall, and the reliability of the estimated points on the cost-effectiveness plane declines. This loss of power is less of a problem for the ‘matching’ method of expansion curve estimation suggested above (see [20] for details).

It should be noted, however, that our method incurs additional costs for the collection and analysis of information. In contrast with conventional subgroup analysis, information on the diagnostic variable \(D\) is required. To estimate the expansion curve we must be able to predict the order in which patients are to be prioritised for treatment. This might not always be easy, or even possible. However, similar criticisms can be levelled at subgroup analysis.

A further problem with expansion curve analysis is that it cannot be readily extended for the comparison of three or more options for a given patient group. For example, suppose that there are three interventions on a conventional linear
frontier, \( A, B \) and \( C \) (in order of increasing average costs and effects). It is possible that a convex expansion path from \( A \) to \( C \) could pass to the southeast of point \( B \). If so, then point \( B \) is dominated by a mixture of \( A \) and \( C \). The cost-effectiveness frontier for three or more alternatives should thus be defined as the outer envelope (furthest to the southeast) of the joint expansion paths for the interventions. Estimation of this envelope of permissible and efficient combinations for three or more programmes is a complicated task.

### Conclusions

The expansion path between adjacent programmes on a cost-effectiveness frontier is likely to be non-linear because of non-linearity in cost functions and/or residual heterogeneity and prioritisation of patients for treatment. In some circumstances, empirical estimation of the expansion curve might help to improve resource allocation decisions. Before proceeding, however, research is required to examine the expected value of additional information about the shape of expansion curves, given the additional information and analysis that is required to estimate them. Further work is also needed to examine the impact of second-order uncertainty on expansion curves, and how to communicate this information to decision makers.

### References

7. Oliver A. Accounting for the missing opportunity costs in incremental cost-outcome analysis. *Appl Health Econ Health Policy* 2002; **1**: 191–196.


