
Article:
Howdon, Daniel David Henry, Lomas, James Richard Scott orcid.org/0000-0002-2478-7018 and Paulden, Mike (2019) Implications of non-marginal budgetary impacts in health technology assessment: a conceptual model. Value in Health. pp. 891-897. ISSN 1524-4733

https://doi.org/10.1016/j.jval.2019.04.001

Implications of non-marginal budgetary impacts in health technology assessment: a conceptual model

Daniel Howdon*[1,2], James Lomas[3], and Mike Paulden[4]

[1]Academic Unit of Health Economics, Leeds Institute of Health Sciences, Worsley Building, University of Leeds, Clarendon Way, Leeds, LS2 9NL, UK.
[2]Department of Economics, Econometrics and Finance, University of Groningen, Duisenberg Building, Nettelbosje 2, 9747AE Groningen, Netherlands.
[3]Centre for Health Economics, University of York, Heslington, York, YO10 5DD, UK.
[4]School of Public Health, University of Alberta, Edmonton, Canada.

April 16, 2019 (author accepted version)

Abstract

Objectives
This paper introduces a framework with which to conceptualise the decision-making process in health technology assessment for new interventions with high budgetary impacts. In such circumstances, the use of a single threshold based on the marginal productivity of the health care system is inappropriate. The implications of this for potential partial implementation, horizontal equity and pharmaceutical pricing are illustrated using this framework.
Results
Under the condition of perfect divisibility and given an objective of maximising health, a large budgetary impact of a new treatment may imply that optimal implementation is partial rather than full, even at a given incremental cost-effectiveness ratio that would nevertheless mean the decision to accept the treatment in full would not lead to a net reduction in health. In a one-shot price-setting game, this seems to give rise to potential horizontal equity concerns. When the assumption of fixity of the ICER (arising from the assumed exogeneity of the manufacturer's price) is relaxed, it can be shown that the threat of partial implementation may be sufficient to give rise to an ICER at which cost the entire potential population is treated, maximising health at an increased level, and with no contravention of the horizontal equity principle.

Keywords: cost-effectiveness, equity, pharmaceutical price regulation.

*Corresponding author: email d.howdon@leeds.ac.uk

1 Introduction

New and expensive health technologies with large eligible patient populations can impose significant budget impacts on healthcare systems[1]. A prominent recent example of new pharmacological treatments for hepatitis C caused issues for health technology assessment agencies around the world[2]. In England, where the National Institute for Health and Care Excellence (NICE) has historically emphasised the importance of cost-effectiveness as opposed to budget impact, approval by NICE for these drugs caused NHS England to break from routine by requiring a delay on implementation of the NICE guidance on account of the large projected budget impact[3]. NICE has since introduced a "budget impact test" that will result in high budget impact technologies being "slow tracked" to allow for further negotiations between NHS England and the manufacturer[4]. Critics have argued that this is a "flawed attempt" to solve the problem and is not rooted within NICE's existing opportunity cost framework[5]. Australia's Pharmaceutical Benefits Advisory Committee (PBAC) adopted a different approach when evaluating new hepatitis C drugs by using a lower cost per QALY 'threshold', justified with reference to the large expected budget impact[2]. Further writing on the topic has suggested that experiences with technologies such as these require more emphasis to be placed on affordability[6–8].

Conventional decision-making for health technology assessment relies upon accepting or rejecting, in full, new interventions based upon their incremental cost-effectiveness ratio (ICER) compared to some benchmark threshold, λ[9a]. The assumption of a constant threshold may be unproblematic in the case of decision-making which does not use up a large proportion of the budget of the health service, faced with healthcare producing health relatively flat-of-the-curve with respect to expenditure at the margin. The use of a benchmark threshold in such a first-best world (see Section 1.1) is held to be a simplification of a mathematical programming problem which in its full form would rely on information that is not available to decision-makers, and require impractical constant funding and defunding of services, with associated costs of so doing[11]. Furthermore, in such situations, decision-making at the margin does not provoke concerns regarding horizontal equity. In cases where both incremental health gain and incremental costs are positive, the comparison of a constant threshold with a constant

[a]Equivalently, a net benefit rule[10] can be used. Without loss of generality, this paper will discuss decision-making solely using the method of ICER/threshold comparison

ICER means that a treatment is either provided (where ICER $< \lambda$) or not provided (where ICER $> \lambda$) in its entirety for all patients with equal capacity to benefit from treatment.

Given the likely diminishing marginal effectiveness of spending on healthcare in improving health and associated non-linearity of the relationship between healthcare expenditure and health outcomes, larger budgetary impacts are likely to imply larger health losses per pound spent on healthcare[2,7,12–14]. Empirical estimation of this relationship has suggested that the use of a constant threshold at the margin of the health system's expenditure is likely to lead to substantial forgone population health where the total budgetary impact is large. With regard to decision-making by NICE for the English NHS, the marginal threshold of £12,936 empirically estimated by Claxton *et al.*[15] was deemed to be around 7% too high in the case of a new treatment with a budgetary impact of £2.5bn[6, b].

Where the budgetary impact is greater, partial (but not full) acceptance may prove cost-effective, due to the non-linearity of the relationship between expenditure and health outcomes. Previously, such issues have appeared to be, or been held to be, irrelevant for two reasons. First, in the absence of an estimate of this relationship beyond, but close to, the margin, partial acceptance can be justified only with resort to a full mathematical programming model. However, in the light of recent empirical estimation of such a relationship, a reduced form model which takes account of this non-linearity is now possible without the need to resort to a full mathematical programming solution. A second objection has been that the partial funding of services is held to contravene principles of horizontal equity, where individuals in equal need (with equal capacity to benefit) should be given equal healthcare treatment[11]. This paper argues that the claim that partial provision would impose a unique contravention of the principle of horizontal equity is inaccurate, and furthermore that any apparent such problem is potentially obviated when the price of treatment is endogenous rather than assumed to be exogenously given.

This paper is structured as follows. Section 2 presents a four-quadrant graphical framework that proposes a reduced form account of decision-making when such situations arise, without resorting to a full mathematical programming solution. Section 3 discusses an allocation problem within the framework of the model, treating the ICER as exogenous. Section 4 considers the

[b] Note that this marginal estimate itself is substantially lower than the threshold of around £30,000 used by NICE in practice

horizontal equity implications of partial provision when the ICER is treated as exogenous, and provides a framework with which to illustrate the costs in terms of population health of the imposition of a strict constraint enforcing horizontal equity in acceptance only. Section 5 considers how a situation where partial acceptance is optimal in a static problem might be resolved dynamically given the optimal reactions of both the decision-maker and the manufacturer when the ICER is treated as endogenous (arising from the price chosen by the manufacturer given the objective function of the decision-maker) rather than exogenous. Section 6 concludes.

1.1 What does the threshold represent?

The meaning attached to the benchmark threshold requires some exposition. A comprehensive summary of the various normative and technical interpretations attached to its value, and the implications thereof, is provided by Culyer[13]. We initially consider a budget-constrained healthcare system operating on its production possibility frontier (PPF), where (as detailed in Paulden[16]) any healthcare provided by the acceptance of a new treatment displaces (parts of) existing treatment programme-populations, with the least cost-effective (parts of) existing treatment programme-populations displaced first, thus displacing as little as health as possible for any given level of displaced spending. In such a system, the threshold we first here consider for marginal decision-making represents the value implied by the efficiency of the system at the margin, representing the opportunity cost of health displaced by a new treatment with a marginal budgetary impact or the reciprocal of the shadow price of the budget constraint (a 'first best' threshold, as characterised by Culyer[13])[c]. While this is a particularly restrictive assumed model and unlikely to entirely accurately characterise anything beyond an ideal type healthcare system, it is highly useful for expositionary purposes. Many departures from these assumptions can be incorporated within the model, and we later highlight one relaxation of this assumption and the corresponding change in interpretation arising from it.

[c]This value is commonly termed k in existing literature when referring to this specific conception of the generic threshold λ. For consistency and clarity, we term this value $\lambda_{\text{marginal}}$

2 Model

Assume that a health service is able to perfectly divisibly assign healthcare inputs, with any existing resources in use able to be reswitched to alternative purposes costlessly and immediately. Health, provided by the health system, is a function of inputs used for treatment, with this relationship specifying a health production function. The decision-making process faced by the relevant authority requires the maximisation of population health, producing according to this production function, subject to its budget constraint.

$$\max \left(\sum_{l=1}^{L} \sum_{j=1}^{J_l} \sum_{i=1}^{I_l} H_{ijl} x_{ijl} \right)$$

subject to:

$$\sum_{l=1}^{L} \sum_{j=1}^{J_l} \sum_{i=1}^{I_l} c^H_{ijl} x_{ijl} \leq b$$

and

$$0 \leq x_{ijl} \leq 1,\ i = 1...I_l\ j = 1...J_l,\ l = 1...L$$

where H is health produced by treatment j with cost c within programme l, for fraction x of population i, and b is the total budget available. Given that this entails selecting the most cost-effective treatment first, followed by the next-most cost-effective treatment and so on, this implies a falling marginal effectiveness of treatment as the budget increases or, by the same token, as more health is produced. The solution of such a problem in its full form requires greater information than is available to, and greater flexibility than is practical for, the decision-making authority and, given these constraints, decision-making generally takes the form of comparing the ICER arising from the new intervention to a constant threshold, $\lambda_{\text{marginal}}$.

While the functional form of a health production function (HPF), of the type specified above, may be unknown, we can consider some general solution of this maximisation problem which implies an indirect health production function (IHPF) of the form $H = f(b, \mathbf{c})$, where $\mathbf{c}$ is a vector of treatment costs[d]. Recent work has produced empirical estimates of the slope of such an indirect health production function beyond the margin, taking account of its non-linearity[6], enabling decision-making to potentially go beyond such a naive assumption. Although such empirical estimates do not recover the explicit production function drawn out above in full and

[d] In line with the use of 'indirect production function' in the conventional theory of the firm, we use the term 'indirect health production function' to distinguish this, as a function of prices and a budget, from a health production function, whose arguments are inputs explicitly.

the optimal mix of treatments for programme-populations, the values derived by such estimation do imply an IHPF, holding individual treatment costs constant and varying b.

Assume a new healthcare intervention is now proposed, with some large budgetary impact m if fully implemented, for a condition which is currently not treated, with no cross-condition associations with other existing or potential treatments and for which capacity to benefit for potentially treatable patients is equal[e]. This implies that any calculated ICER is based on a comparison to no treatment, and the ICER is equal to the cost per QALY gain over no treatment. We initially assume constant QALY returns to spending on this intervention. Health (H_1) can be produced by use of this new intervention, or (H_2) by the use of existing treatments already employed. The decision-making authority is tasked with maximising overall population health ($H_1 + H_2$), and decides upon whether to approve or reject the new intervention on this basis. The monetary price of the new treatment, and the associated consequent ICER, is initially assumed to be given and fixed. This implies the displacement of treatment with total expenditure of up to m, depending on the fullness of implementation.

Consider Figure 1, a graphical illustration of this situation.

[Figure 1 here]

Quadrant I (top-right) presents the final objective function and constraint: maximising total health ($H_1 + H_2$) subject to technological and budgetary constraints. This implies selecting the highest possible level of health (the highest possible isohealth curve, representing $H_1 + H_2$=constant and with slope -1) attainable from given current production technologies, prices, and a fixed budget, b (as illustrated by the given red production possibility frontier (PPF)). Up to H_{2max} can be produced from existing treatment (point A), and up to H_{2max} can be produced from new treatment (point B). This quadrant is derived from the three remaining quadrants.

[e]This could represent either all possible patients to be treated by the new intervention or the most expensive subgroup thereof, as per Claxton *et al.*[17]. While potentially each individual patient could form a separate subgroup with a different capacity to benefit, as discussed in, *inter alia* Espinoza *et al.*[18] and Gavan *et al.*[19], in general such levels of precision will be neither possible nor economically optimal, and it will be necessary to use some degree of average treatment effect in estimating the ICER of a given treatment. Indeed, our earlier cited example[3] involved an intervention that had a high budgetary impact even after being recommended only in varying doses for subgroups of patients with certain genotypes of hepatitis C, further conditional on previous treatments received.

Quadrant IV (bottom-right) displays how much health can be created by given amounts of spending on the new treatment, and has a line with a slope of the negative of the ICER.

Quadrant III (bottom-left) represents the healthcare provider's budget constraint, with the solid section of this representing the section that could potentially be used under conditions of full implementation of the new technology (m, assumed to be a large proportion of b).

Quadrant II (top-left) represents two versions of the IHPF, with the solid sections representing health that is currently generated from spending b on existing treatment[f]. The non-marginal IHPF for this range of spending (black line) exhibits diminishing returns, reflecting the changes in the cost-effectiveness of health produced by the system when moving away from the margin. The assumed marginal IHPF (grey line), however, is linear, with a slope equal to that of the non-marginal IHPF at its local maximum here ($-1/\lambda$). This linearity implies an assumption that healthcare is displaced at a constant opportunity cost equal to the threshold at the margin. Spending the full budget b on the existing treatment implies a QALY production of H_{2max}, consistent with position A in Quadrant I. The health assumed to be generated and lost at all other points on this IHPF differs according to whether the pure marginal or non-marginal IHPF is used: specifically, the pure marginal IHPF implies a lower opportunity cost in terms of QALYs from existing treatment when additional spending is made on the new treatment, for all levels of budgetary impact strictly greater than 0 and less than or equal to m.

3 Allocation problem

[Figure 2 here]

We now consider an allocation problem in which: 1) full implementation of the new intervention would reduce population health by displacing health produced in more cost-effective ways, 2) some degree of partial implementation, if possible, would lead to a rise in population

[f]Our assumption that displacement of existing healthcare provided occurs in order of least health displaced (i.e., from least to most cost-effective) can be replaced by an assumption that displacement occurs in order of increasing cost-effectiveness, even if this does not represent displacing strictly the least-cost-effective treatment first – as is reported empirically by Lomas *et al.*[6]. This would be one version of a second-best type threshold as characterised by Culyer[13]. In such a situation, the relationship plotted in this quadrant would cease to represent an indirect health production function, and instead represent what might be termed a health displacement function. All further conclusions derived in this paper remain intact, but should be reframed in this context.

health, 3) consequently the use of our non-marginal IHPF shows a fall in population health if fully implemented, 4) however, if compared to the assumed pure marginal threshold (consistent with use of the pure marginal IHPF) a rise in population health appears to be the case at full implementation. We elaborate on this in the below.

Point A in Figure 2 represents the initial position: no spending can be made on the new, potentially high budgetary impact, treatment before its introduction. The implications of a decision to use the threshold implied by the pure marginal or non-marginal IHPF are here illustrated. Assume first that a decision is to be made between accepting the treatment and spending the full amount m on its provision, implying spending $b - m$ on existing treatment, or remaining at A, with no change in spending – a situation equivalent to a choice involving a potential treatment programme with perfect indivisibility. Under the assumption of pure marginality, acceptance would entail a move from point A to point B', consistent with an apparently higher isohealth curve and an associated higher level of overall health, and the acceptance of the technology on these grounds, due to an apparent increase in total health generated. Under the more realistic assumption of non-marginality, however, this entails a move from point A to point B, consistent with a lower isohealth curve, an associated lower level of overall health, and the rejection of this technology on these grounds, due to a decrease in total health generated.

More interesting cases arise when the treatment is perfectly divisible, and decisions can be made to accept a proportion of spending. Consider the decision to reject this treatment, or accept its full or partial provision. At any point to the right of A and the left of D (where the PPF cuts the original isohealth line), population health is increased by the partial provision of the new treatment. A discrete choice between rejection and acceptance up to a level just to the left of D, for instance E, would cause the decision-maker to accept the new treatment and displace AF QALYs, gaining a greater FE QALYs, and thus increasing population health by $FE - AF$. While this would represent an increase in overall health, it also represents a suboptimal outcome when the objective is the maximisation of total health alone.

Consider point C. As we move right from point A along the PPF towards point C, the slope of the PPF is shallower than that of the isohealth line (-1), representing the relative cost-effectiveness of the new treatment compared to existing treatments, and (diminishing

marginal) increases in population health from additional spending on the new treatment. As we move right beyond point C, the converse is true: the slope is steeper than -1, existing treatment is relatively more cost-effective, and total health is reduced by further use of the new treatment and concomitant displacement of health produced by existing treatments. An optimum in terms of total health is therefore achieved at C, where the ratio of the marginal health gains of the old and new treatments are equal to 1, the (negative of the) slope of the isohealth line. This implies that, for decisions involving non-marginal impacts on the healthcare system's budget, decision-makers maximising total health should consider the partial acceptance of a new health technology, up to the point that increasing spending on its provision ceases to produce more health than it displaces, rather than necessarily entirely accepting or entirely rejecting the treatment based on its total budgetary impact.

4 Implications for horizontal equity

Assuming a constant ICER for the entire population with capacity to benefit, objections to outcome C may be made on the grounds of a contravention of the principle of horizontal equity – that individuals with equal capacity to benefit should be treated equally, and specifically that individuals with an equal health condition should receive equal healthcare treatment[g]. On such an account, the existence of point C is irrelevant in this situation, representing a partial implementation of the new treatment that is ruled out on horizontal inequity grounds. However, it is important to recognise two sources of horizontal inequity: that of partial implementation, and that of partial displacement.

When a new healthcare intervention is approved, a fixed budget necessitates that some other healthcare provided elsewhere in the system is displaced. This means that – unless a new healthcare intervention displaces the entirety of one or more interventions already provided – that partial displacement, contravening a principle of horizontal equity, will occur when any new treatment at all is provided. The acceptance of any new treatment in its entirety, even when its total spending forms only a small proportion of the healthcare budget, will inevitably displace a

[g]We could, in principle, suggest a conception of horizontal equity that provided each individual with an equal chance of obtaining treatment provided by, for instance, some lottery mechanism which would not be contravened by such an allocation mechanism.

proportion of spending on healthcare programmes elsewhere in the system. For interventions that are divisible, limitations on the amount of provision in any given time period can arise in displacement through the extension of waiting lists for that intervention[14].

While Appleby *et al.*[20] remark that local NHS commissioners may place restrictions on patient access on grounds of clinical outcome, they further note that 'waiting list initiatives' are also employed when displacement of existing provision occurs (see also Claxton *et al.*[15]). Daniels *et al.*[21] note that in the English NHS 'when services are retracted, access criteria are often not specified and patients continue to access services in the same way; for patients, the noticeable effects of retraction are often limited to lengthening waiting times'. While not directly examining partial displacement, Chen *et al.*[22] note a number of highly cost-effective conditions in the Irish public health system that are subject to large waiting lists. While it is for law- and policy-makers to decide whether partial provision is more unjust than partial displacement, it must be recognised that, where found to occur, both are sources of the same type of horizontal inequity. Further research is warranted on the existence of, and on preferences regarding, partial displacement within publicly-funded healthcare systems, in order to inform both statutory guidance and decisions in practice.

Further, we can illustrate the QALY costs due to such a horizontal equity constraint in implementation as $GC - AG$ (Figure 2), the loss of potential health gains accrued by a failure to partially accept at the optimum. Again, it is for policy-makers to decide if this loss of population health is a price worth paying to avoid horizontal inequity in partial implementation.

5 Implications for pharmaceutical pricing

The preceding sections have assumed the exogeneity of the price of treatment and of the associated ICER. The decision-making process with regard to low budgetary impact technologies, where the relevant threshold is known or can be inferred from previous decisions, provides an incentive for the manufacturer to price at the threshold, such that no net health gain is made from the acceptance of the treatment but all surplus is captured by the manufacturer[23]. This nevertheless represents situation in which the decision-maker, with an objective function of maximising health, and the manufacturer would be unable to effect a mutually beneficial movement from this point of full implementation at a price consistent with ICER_1.

[Figure 3 here]

The story is somewhat different in the case of decision-making with large budgetary impacts, where estimates of the shape of the IHPF beyond the margin can be made. Consider now Figure 3, where full implementation does not represent an optimal outcome for a decision-maker trying to maximise total health, but results in zero net health gain (with both A and B lying on the same isohealth curve)[h]. As discussed, holding the ICER as exogenous and given, the requirement to maximise total health would necessitate (setting aside equity concerns) the approval of only a proportion of the total possible volume of use, treating some patients and not others but maximising population health, again at point C. While this partial approval would represent an optimal solution from the point of view of a decision-maker in a one-shot game where price is treated as exogenous, it also represents a point from which there may exist the potential for mutually-beneficial gains for both decision-maker and manufacturer. As per Pandey *et al.*[24] and Paulden[25], the likely distribution of such manufacturer reserve ICERs – their minimum willingness to accept in order to supply – is an empirical question requiring further research. Nevertheless, given the often low marginal cost associated with pharmaceutical products, profit at the margin at this partial-approval point is likely to be positive in at least some cases at a price that the decision-maker would be willing to pay for more units of the treatment.

The existence of point C in such a case would represent a credible threat to manufacturers: it represents an improvement in population health compared to both point A and point B. The decision-maker can improve population health by not treating all patients with capacity to benefit, and thus reducing manufacturer profits compared to those that can be gained, compared to full implementation with no health loss (B). In short, point C is preferable to point B for the decision-maker (representing a gain in terms of population health) but less preferable to point B for the manufacturer (reducing profit by reducing quality while retaining the same price). The non-optimality (in the sense that mutual gains between the manufacturer and decision-maker may be possible) of point C means that a lower price of the treatment (but higher than at the partial implementation point C) can result than that which would prevail at point B where the

[h]If the manufacturer expects the relevant threshold for a given level of budgetary impact to be set at that involving no total health loss, a profit-maximising manufacturer would, in a one-shot game, set price at this level. We therefore adopt this as a new starting point in this section.

programme is fully implemented and no net health gains are made. The process by which this may occur is illustrated in Figure 3.

By drawing the implied PPFs resulting from a change in the price (change in the ICER) of a treatment, we can consider one possible price-setting outcome. A lower price shifts the illustrated line in quadrant IV from ICER_1 to ICER_2, leading to a related outward shift in the PPF. Given that the entire patient population has a total potential benefit of H_{1max}, the PPF is again undefined to the right of this position. At some point, the ICER will fall to a low enough level such that the PPF is tangent to the isohealth curve at a position H_{1max}: i.e. at point B^*. Between B^* and A, it is possible to plot out a best response function (green line, top-right quadrant) of the decision-maker to a changing ICER, giving a locus of tangencies of the PPF with the isohealth line (including C) as this ICER varies. The manufacturer's problem now becomes choosing a price concomitant with a position on this best response curve such that profit is maximised[i]. At B^* health is maximised at a level at which mutual gains for manufacturer and decision-maker are impossible, with a higher total level of health than at C, the entire potential patient population treated, and as a result with no horizontal equity concerns arising in the provision of this new treatment.

Point C, held to be in contravention of a particular conception of horizontal equity, does not form an optimal solution but nor does it necessarily form the ultimate solution, but rather it acts as a credible threat that the decision-maker can use when the price for treatment is to be set. If the existence of point C cannot even be considered by the decision-making body, it cannot exist as a credible threat that may be able to force a move to a lower price (lower ICER) for the treatment that can also result in the full population receiving treatment, with no potential horizontal equity concerns arising in this final outcome. While manufacturer knowledge of the decision-making process for health technology assessment with a constant threshold implied by the IHPF at the margin leads to a situation where net health gains are driven towards zero and

[i]This is a similar result to that discussed in Claxton *et al.*[17] (see footnote e). The 'menu of options' discussed here is analogous to our price-offer curve. In Claxton *et al.*[17] a discrete number of subgroups gives rise to different combinations of quantity and willingness to pay due to the different effectiveness of the same treatment within each of these subgroups, with opportunity costs assumed to be constant. In our model, different combinations of quantity and willingness to pay arise due to the diminishing efficiency of healthcare expenditure with respect to health, with the effectiveness of new spending assumed to be constant across the relevant patient population.

all surplus captured by the manufacturer, knowledge of the existence of the non-linear PPFs (and the associated best-response functions of the decision-maker) may lead to a situation where the price set by the manufacturer is that associated with the maximisation of total health, with a positive net health gain over non-implementation.

6 Conclusion

The conventional approach of rejecting or accepting a new healthcare technology by comparison of the ICER to some cost-effectiveness threshold is appropriate, as an approximation to a full mathematical programming solution to the maximisation of total health, if the impact on the healthcare service's budget is marginal. Such a case generates no concerns regarding horizontal equity in the provision of new treatments, as such treatments are either accepted or rejected in their entirety. This paper presents a simple theoretical model illustrating the choice faced by decision-makers in situations where the budgetary impact of an intervention is great enough to make the non-linearity of the relevant section of the IHPF important, and where those decision makers have knowledge of the shape of the IHPF as we move away from the margin. In such a situation, optimal acceptance of the new intervention in a static exogenous-ICER case may be partial rather than full. If implementation is partial rather than full, horizontal equity concerns may appear to arise from the fact that patients with identical health conditions and equal capacity to benefit will not receive equal treatment. It is important, however, to recognise that such horizontal equity issues exist within the actually-existing system, and are exacerbated when any new treatment is accepted and health provided by other existing treatments in other sectors is displaced. Furthermore, even if the relevance of this horizontal equity constraint is accepted in the statics, it does not necessarily exist in a price-setting game that arises from treating the ICER as endogenous rather than exogenous. Such a price-setting process may result in an ICER that is lower than that implied by a threshold involving no loss of health at full implementation, and which may maximise population health while treating the entire population with capacity to benefit.

References

[1] A. Glassman and R. Forman, *Setting Universal Health Coverage Priorities: India and Dialysis*, 2016, https://www.cgdev.org/blog/setting-universal-health-coverage-priorities-india-and-dialysis.

[2] A. H. Harris, Beyond the threshold, *Health Econ Policy Law*, 2016, **11**, 433–438, DOI: 10.1017/S1744133116000050.

[3] National Institute for Health and Care Excellence, *Ledipasvir-sofosbuvir for treating chronic hepatitis C*, 2015, https://www.nice.org.uk/guidance/ta363.

[4] National Institute for Health and Care Excellence, *Procedure for varying the funding requirement to take account of net budget impact*, 2017, https://www.nice.org.uk/Media/Default/About/what-we-do/NICE-guidance/NICE-technology-appraisals/TA-HST-procedure-varying-the-funding-direction.pdf.

[5] V. Charlton *et al.*, Cost effective but unaffordable: an emerging challenge for health systems, *BMJ*, 2017, **356**, j1402, DOI: 10.1136/bmj.j1402.

[6] J. Lomas, K. Claxton, S. Martin and M. Soares, Resolving the "Cost-Effective but Unaffordable" Paradox: Estimating the Health Opportunity Costs of Nonmarginal Budget Impacts, *Value Health*, 2018, **21**, 266–275, DOI: 10.1016/j.jval.2017.10.006.

[7] S. D. Pearson, The ICER Value Framework: Integrating Cost Effectiveness and Affordability in the Assessment of Health Care Value, *Value Health*, 2018, **21**, 258–265, DOI: 10.1016/j.jval.2017.12.017.

[8] A. Bilinski, P. Neumann, J. Cohen, T. Thorat, K. McDaniel and J. A. Salomon, When cost-effective interventions are unaffordable: Integrating cost-effectiveness and budget impact in priority setting for global health programs, *PLoS Med*, 2017, **14**, e1002397, DOI: 10.1371/journal.pmed.1002397.

[9] M. F. Drummond, M. J. Sculpher, K. Claxton, G. L. Stoddart and G. W. Torrance, *Methods for the Economic Evaluation of Health Care Programmes*, Oxford University Press, 2015.

[10] A. A. Stinnett and J. Mullahy, Net health benefits: a new framework for the analysis of uncertainty in cost-effectiveness analysis, *Med Decis Making*, 1998, **18**, S68–80, DOI: 10.1177/0272989X98018002S09.

[11] D. M. Epstein, Z. Chalabi, K. Claxton and M. Sculpher, Efficiency, Equity, and Budgetary Policies: Informing Decisions Using Mathematical Programming, *Med Decis Making*, 2007, **27**, 128–137, DOI: 10.1177/0272989X06297396.

[12] C. McCabe, K. Claxton and A. J. Culyer, The NICE cost-effectiveness threshold: what it is and what that means, *PharmacoEconomics*, 2008, **26**, 733–744.

[13] A. J. Culyer, Cost-effectiveness thresholds in health care: a bookshelf guide to their meaning and use, *Health Econ Policy Law*, 2016, **11**, 415–432, DOI: 10.1017/S1744133116000049.

[14] M. Paulden, J. O'Mahony and C. McCabe, Determinants of Change in the Cost-effectiveness Threshold, *Med Decis Making*, 2017, **37**, 264–276, DOI: 10.1177/0272989X16662242.

[15] K. Claxton *et al.*, Methods for the estimation of the National Institute for Health and Care Excellence cost-effectiveness threshold, *Health Technol Assess*, 2015, **19**, 1–503, v–vi, DOI: 10.3310/hta19140.

[16] M. Paulden, *Opportunity cost and social values in health care resource allocation*, 2016, `https://doi.org/10.7939/R3M902D4P`.

[17] K. Claxton *et al.*, Value based pricing for NHS drugs: an opportunity not to be missed?, *BMJ*, 2008, **336**, 251–254, DOI: 10.1136/bmj.39434.500185.25.

[18] M. A. Espinoza, A. Manca, K. Claxton and M. J. Sculpher, The Value of Heterogeneity for Cost-Effectiveness Subgroup Analysis, *Med Decis Making*, 2014, **34**, 951–964, DOI: 10.1177/0272989X14538705.

[19] S. P. Gavan, A. J. Thompson and K. Payne, The economic case for precision medicine, *Expert Rev Precis Med Drug Dev*, 2018, **3**, 1–9, DOI: 10.1080/23808993.2018.1421858.

[20] J. Appleby, N. Devlin, D. Parkin, M. Buxton and K. Chalkidou, Searching for cost effectiveness thresholds in the NHS, *Health Policy*, 2009, **91**, 239–245, DOI: 10.1016/j.healthpol.2008.12.010.

[21] T. Daniels, I. Williams, S. Robinson and K. Spence, Tackling disinvestment in health care services: The views of resource allocators in the English NHS, *J of Health Org and Mgt*, 2013, **27**, 762–780, DOI: 10.1108/JHOM-11-2012-0225.

[22] T. C. Chen, D. Wanniarachige, S. Murphy, K. Lockhart and J. O'Mahony, Surveying the Cost-Effectiveness of the 20 Procedures with the Largest Public Health Services Waiting Lists in Ireland: Implications for Ireland's Cost-Effectiveness Threshold, *Value Health*, 2018, **21**, 897–904, DOI: 10.1016/j.jval.2018.02.013.

[23] A. Gafni and S. Birch, Incremental cost-effectiveness ratios (ICERs): The silence of the lambda, *Soc Sci Med*, 2006, **62**, 2091–2100, DOI: 10.1016/j.socscimed.2005.10.023.

[24] H. Pandey, M. Paulden and C. McCabe, *Theoretical models of the cost-effectiveness threshold, value assessment, and health care system sustainability*, Institute of health economics technical report, 2018.

[25] M. Paulden, *Strategic Behaviour and the Cost-Effectiveness Threshold: A New Conceptual Model*, 2018, `https://www.ihe.ca/publications/theoretical-models-of-the-cost-effectiveness-threshold-value-assessment-and-health-care`

Figures

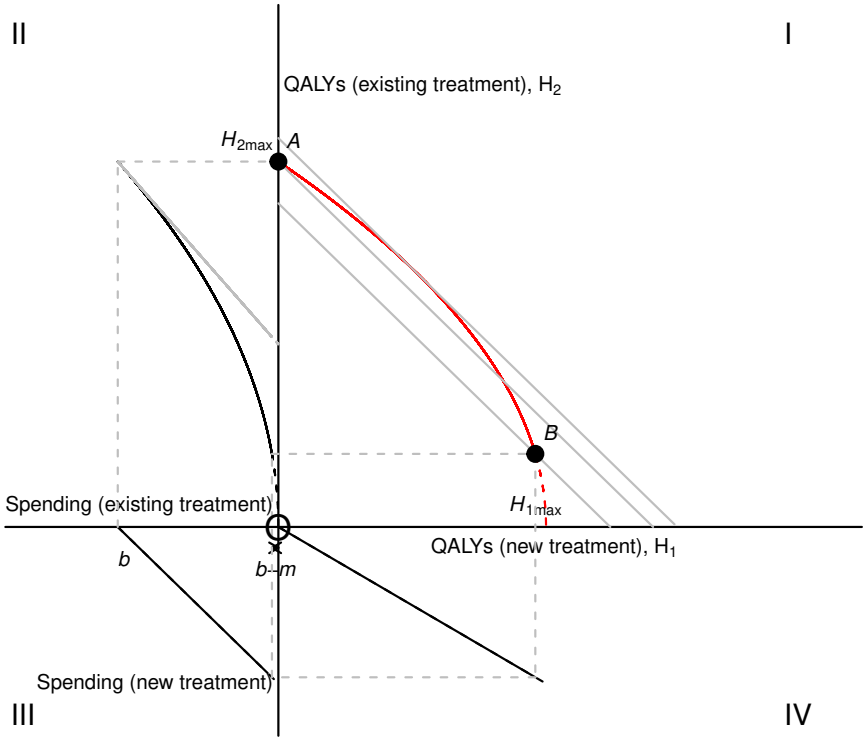

Figure 1: A four-quadrant model of health production

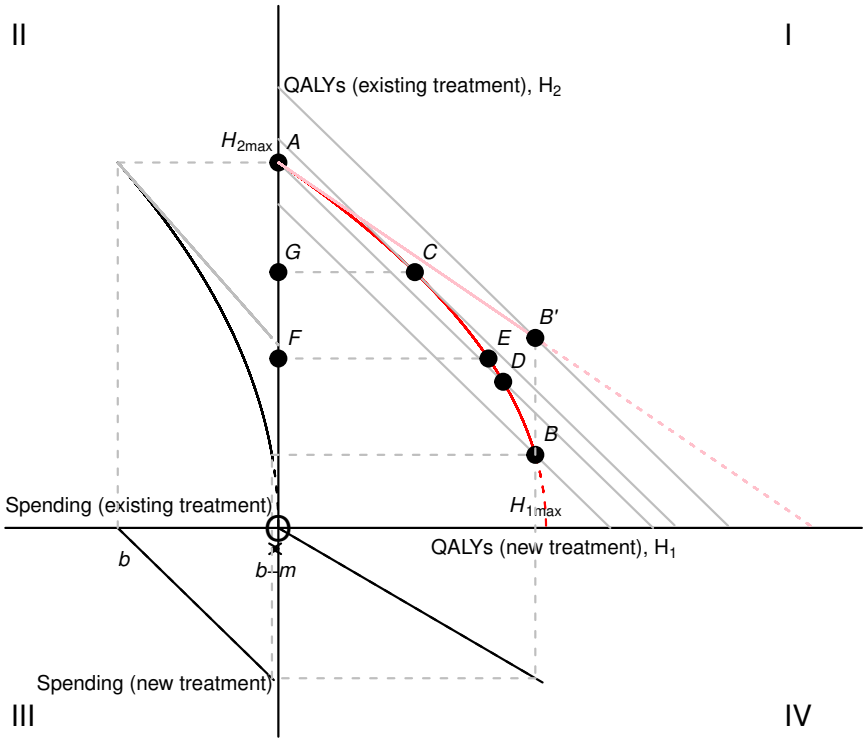

Figure 2: An allocation problem

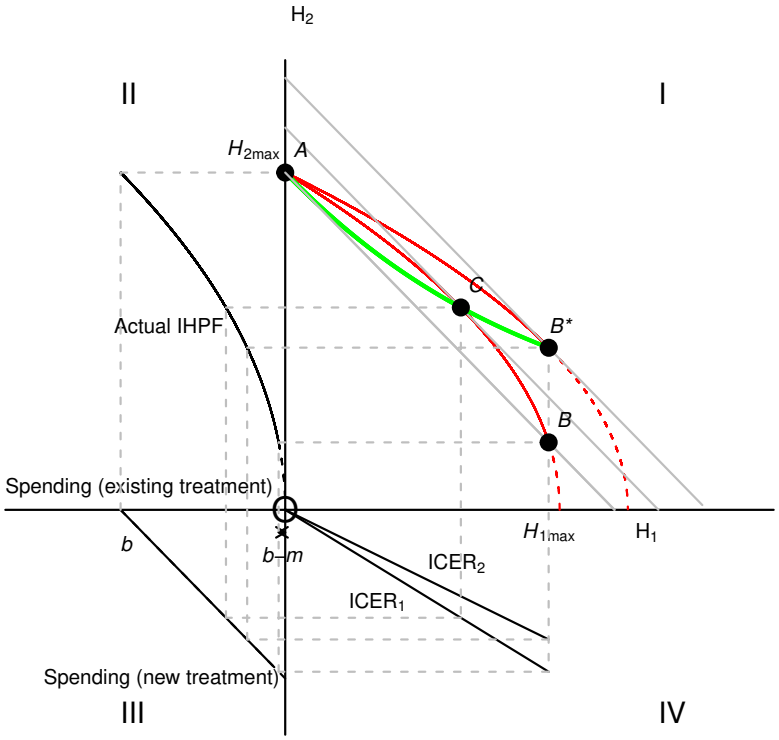

Figure 3: A best response function of the decision-maker