Journal of Health Economics 27 (2008) 1224–1236

Contents lists available at ScienceDirect

Journal of Health Economics

journal homepage: www.elsevier.com/locate/econbase

Cost-effectiveness analysis and innovation

Anupam B. Jena, Tomas J. Philipson*,[1]

The University of Chicago, United States

ARTICLE INFO

Article history:
Received 25 January 2006
Received in revised form 27 May 2008
Accepted 28 May 2008
Available online 5 June 2008

JEL classification:
I10
I11
I18

Keywords:
Cost effectiveness
Dynamic efficiency
R&D
Surplus appropriation

ABSTRACT

While cost-effectiveness (CE) analysis has provided a guide to allocating often scarce resources spent on medical technologies, less emphasis has been placed on the effect of such criteria on the behavior of innovators who make health care technologies available in the first place. A better understanding of the link between innovation and cost-effectiveness analysis is particularly important given the large role of technological change in the growth in health care spending and the growing interest of explicit use of CE thresholds in leading technology adoption in several Westernized countries. We analyze CE analysis in a standard market context, and stress that a technology's cost-effectiveness is closely related to the consumer surplus it generates. Improved CE therefore often clashes with interventions to stimulate producer surplus, such as patents. We derive the inconsistency between technology adoption based on CE analysis and economic efficiency. Indeed, static efficiency, dynamic efficiency, and improved patient health may all be induced by the cost-effectiveness of the technology being at its *worst* level. As producer appropriation of the social surplus of an innovation is central to the dynamic efficiency that should guide CE adoption criteria, we exemplify how appropriation can be inferred from existing CE estimates. For an illustrative sample of technologies considered, we find that the median technology has an appropriation of about 15%. To the extent that such incentives are deemed either too low or too high compared to dynamically efficient levels, CE thresholds may be appropriately raised or lowered to improve dynamic efficiency.

1. Introduction

Technological change is often argued to be a central force behind the growth in health care spending.[2] There is a long-standing and vast health economics literature that attempts to assess the value of spending on such new technologies by use of cost-effectiveness, cost-utility, or cost–benefit analysis, hereafter referred to collectively as CE analysis.[3] There is a

* Corresponding author. Tel.: +1 773 502 7773; fax: +1 773 702 0926.
E-mail address: t-philipson@uchicago.edu (T.J. Philipson).

[1] A previous version of this paper was circulated as an NBER Working Paper titled "Surplus Appropriation from R&D and Technology Assessment Procedures." We are thankful for helpful comments from the Editor Richard Frank, Gary Becker, Angus Deaton, Isaac Ehrlich, Darius Lakdawalla, Casey Mulligan, and Kevin Murphy as well as seminar participants at The University of Chicago, Princeton University, UC Berkeley, SUNY Buffalo, Federal Trade Commission, the 2006 ASHE Meetings, the 2006 Kiel Institute Conference on New Technology and National Health Systems, and the 2005 NIH-Director Conference *Biomedical Innovation and the Economy*. Jena received fellowship support from the NIH through the University of Chicago Medical Scientist Training Program and from the Bing Center at RAND.

[2] See e.g. Newhouse (1992).

[3] The literature is vast, but, for example, see Weinstein and Stason (1977), Phelps and Parente (1990), Weinstein and Manning (1997), Johannesson and Weinstein (1993), Gold et al. (1996), Meltzer (1997), Drummond et al. (1997), Garber and Phelps (1997), Garber (1999), Cutler and McClellan (2001), and Cutler (2004). Our analysis departs from this previous analysis by considering the implications for innovation incentives.

0167-6296/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.jhealeco.2008.05.010

growing emphasis on using such analysis to guide new technology adoption and manage its impact on long term health care spending. As the name suggests, CE analysis can offer governments and private payers an important means to allocate often scarce health care resources based on the costs and effectiveness of available medical technologies.

In practice, CE analysis has so far guided policy decisions in the form of adoption based on CE thresholds, which dictate that a given technology will be reimbursed only if the incremental costs per quality-adjusted life year (QALY) they provide are below a given threshold. Currently, this type of analysis already plays a role in public reimbursement decisions outside the US. For example, both the UK's National Institute for Clinical Excellence (NICE) and Australia's Pharmaceutical Benefits Advisory Committee have been reported to follow CE thresholds in technology adoption decisions. In Australia, for example, only 2 out of 26 submissions were accepted for reimbursement whose cost per-life year saved exceeded US$ 57,000–similarly, only 1 out of 26 submissions was rejected whose cost per life-year saved was less than US$ 32,000 (Bethan et al., 2001). Similarly, in a review of NICE determinations for which the cost per QALY saved was stated, Raftery (2001) finds that, with the exception of one drug, all recommended technologies had a cost per QALY saved less than £30,000.[4] Such explicit thresholds for adopting medical technologies are not used in public coverage and reimbursement decisions by the Centers for Medicare and Medicaid Services (CMS) in the US. However, their use has been discussed extensively and it nevertheless appears as a reasonable prediction that, *de facto*, technologies that cost more and offer fewer health benefits are more closely scrutinized before they are adopted, if not adopted less. Given the widespread and growing use of explicit or implicit CE analysis for guiding technology adoption, it is therefore important to understand the implications of such criteria for patients who consume new medical technologies and innovators who discover them.

In this paper, we interpret CE-based technology adoption as closely resembling others forms of (supply) price regulation, such as price controls and rate-of-return regulations, and therefore having similar implications for economic efficiency. CE thresholds are price controls in the sense that if price determines the costs to purchasers and the health effects determine effectiveness, adoption policies based on cost-effectiveness are adoption policies based on price. Although not explicitly stated as such, we argue that CE thresholds utilized in practice are implicitly concerned with maximizing the surplus available to consumers at the cost of reduced producer surplus or profits to innovators. In particular, many technology assessments attempt to quantify the health impacts of new technologies for patients by comparing patient benefits from a given technology with spending on that technology. Examples include cost-effectiveness using spending per quality- or disability adjusted life years, as is common by public buyers outside the US, or cost–benefit analysis monetizing mortality reductions through value-of-life estimates, as is common in studies assessing the gains of increased health care spending. The central theme of such standard CE assessments performed in practice seems to be to measure *static* consumer surplus or net consumer benefits—technologies are deemed more valuable the larger are the patient health benefits above what is spent on them.

However, when new technologies are brought to life from costly R&D, consumer surplus may be a poor guide to inducing optimal R&D investments. Rather, the degree to which producer surplus captures social surplus, often at the expense of consumer surplus, becomes the central issue that determines dynamic efficiency. This, of course, is the rationale for the patent system, which substitutes producer surplus for consumer surplus in order to stimulate more efficient R&D investment. Therefore, we argue that for the same reason that patents are preferred even though they lower consumer surplus after technologies are discovered, technology adoption criteria are preferred that do not only focus on consumer surplus as CE criteria do. Put differently, even though measured levels of CE would be larger without patents, since patients or health plans would spend less to get the same technology, dynamic efficiency would presumably be lowered. An illustrative case of this may be vaccines, which, due to government monopsony power, many times have been estimated to be extremely cost-effective yet lack any appreciable R&D investments.[5]

As consumer surplus or cost-effectiveness determines static efficiency and innovator appropriation determines dynamic efficiency, we analyze how CE adoption criteria affect the two. In the simplest case of monopoly R&D, we arrive at the stark implication that static efficiency, dynamic efficiency, as well as patient health are maximized when incremental CE ratios are *maximized*; that is, when the incremental cost per QALY is at its highest possible level. We show how this implication is altered under public R&D subsidies (such as those by the National Institutes of Health (NIH) in the US), competition in R&D (leading to patent racing that may duplicate R&D efforts), and moral hazard in insurance markets, all of which may lead to overinvestment into R&D.

Since the optimal level of innovator appropriation depends heavily on the model of R&D and any ex-post inefficiencies that may arise in health care markets, we illustrate how CE thresholds can be used to achieve any given level of appropriation. Intuitively, public CE adoption policies separate the demand price (paid by consumers) from the supply price (paid by the government to manufacturers) and therefore have the potential to maximize *both* static and dynamic efficiency by ensuring access and R&D incentives simultaneously. Without a separation between the demand- and supply-price, access and innovation cannot both be at their efficient level as is well known from the second-best nature of patents that inefficiently restricts access ex-post. To the extent that current incentives for innovation are deemed either too low or too high compared

[4] More generally, while prior to 1993 no European countries formally required economic assessments of new technologies for pricing and reimbursement decisions (Drummond et al., 1993), by 1999 most of the 13 European countries evaluated by Drummond (1999) had or were in the process of developing formal agencies responsible for such assessments.

[5] In addition to monopsony power, product liability has been argued to also play a role in limited R&D into vaccines (see e.g. Manning, 1994).

to dynamically efficient levels, CE thresholds may be appropriately raised or lowered to improve dynamic efficiency without compromising static efficiency.

As the ability of innovators to appropriate the surplus of their innovations is central to dynamic efficiency, we illustrate how common estimates of CE and the degree of appropriation relate for over 200 technologies contained in the Harvard Cost-Effectiveness Registry. While the drugs included in the Registry are by no means randomly selected, they provide an illustrative benchmark on how to use CE estimates to infer the extent to which producers appropriate the social surplus generated by their innovations. We derive conditions under which the measured level of CE of a technology may be used in conjunction with other estimable parameters to identify the share of social surplus appropriated by producers of that technology; the CE of a given technology *reveals* information about the demand parameters. When such identification is feasible, the existing and vast CE literature informs us about the degree of innovator appropriation. Given our parameter assumptions, we calculate that 25% of the interventions considered have estimated levels of appropriations of less than 7%, while 75% have appropriations less than 25%. These findings can be usefully contrasted to estimated levels of surplus appropriation by producers of HIV/AIDS drugs on the order of 5–10% (Philipson and Jena, 2005).[6]

In the simplest model of monopoly R&D investment, such levels of appropriation would be too low and CE thresholds which equal the average willingness-to-pay for a QALY (say e.g. US$ 100,000/QALY) would improve dynamic efficiency. But, other models of innovation which incorporate important aspects of healthcare substantiate lower thresholds. The main point we stress is not that existing incentives for innovation are too high or too low, but that existing policies to guide technology adoption do have implications for dynamic welfare and that these implications should be considered in assessing the merits of such policies.

The paper may be briefly outlined as follows. Section 2 discusses the relationship between CE measures, thresholds, and static and dynamic efficiency. Section 3 illustrates how existing CE estimates can be used to infer levels of surplus appropriation by producers of medical technologies. Lastly, Section 4 concludes.

2. Technology assessment and dynamic versus static efficiency

2.1. Basic framework

2.1.1. Definitions

In order to discuss how CE measures relate to standard notions of static and dynamic efficiency, we first consider a market without insurance or other forms of demand subsidization in which the supply price equals the demand price. Later, we relax this assumption to illustrate how the application of CE measures differs when demand is subsidized and patients pay only a fraction of the supply price (e.g. with third-party reimbursement). Let q denote the quantity of consumers utilizing a single discrete output, $p(q)$ the inverse demand curve or the willingness-to-pay of the qth user, and $C(q)$ the total cost function. We consider a single monopoly producer of the product whose producer surplus (profits) is given by:

$$\Pi(q) = p(q)q - C(q) \tag{1}$$

The surplus of the consumers engaged in consumption is written as:

$$Z(q) = \int_0^q [p(y) - p(q)]\,\mathrm{d}y = G(q) - p(q)q \tag{2}$$

where $G(q)$ is the total gross consumer benefit when there are q users of the good. The static social welfare $W(q)$ is then defined by consumer and producer surplus together:

$$W(q) = Z(q) + \Pi(q) = G(q) - C(q) \tag{3}$$

The statically efficient level of output q_{W} maximizes W. When there is a single supply and demand price, as would be true when demand is not subsidized, static welfare is of course maximized when price equals marginal cost.

To consider the dynamic efficiency induced by common health care assessment criteria, one must consider how such criteria affect efficiency in the presence of technological change driven by endogenous R&D. If $x(r)$ is an increasing, differentiable, and strictly concave function representing the probability of discovery for a given level of R&D undertaken, r, the dynamic social welfare given R&D r and output q is:

$$D(r, q) = x(r)W(q) - r \tag{4}$$

The first-best R&D and output pair (r^*, q^*) maximizes both dynamic and static welfare. When profits induce R&D, the first-best allocation results when the monopoly firm is able to perfectly price discriminate among consumers so that consumer surplus is zero. In this case, profits equal the social surplus obtained at the competitive quantity, $W(q_{\mathrm{W}})$, and the R&D induced by these profits is first-best. In general, this allocation cannot be achieved by a single supply and demand price when consumers have heterogeneous valuations of the product, hence the second-best nature of such monopoly pricing.

[6] Our findings relate to an existing literature on the general inability of innovators to capture the social value of their inventions, see e.g., Mansfield et al. (1977), Mansfield (1985), Levin et al. (1987), Hall (1996), and Nordhaus (2004).

2.1.2. Interpreting cost-effectiveness measures

CE measures can be related to the standard economic framework above in a straightforward manner. Consider a technology that provides an incremental benefit in health h over a baseline technology and has the incremental price p. For example, h may be interpreted as the incremental extension in quality-adjusted life years due to the technology and is assumed distributed according to $F(h)$. If $g(h)$ represents the willingness to pay for an improvement in health h, only those with $g(h) \geq p$ will consume the technology. Alternatively, only those with health improvements $h_i \geq g^{-1}(p) \equiv h(p)$ will consume the technology at the price p. This leads to an overall demand for the product given by $q(p) = 1 - F(h(p))$ and an equivalent inverse demand curve given by $p(q) = g[F^{-1}(1-q)]$. The inverse demand curve, as usual, reflects the willingness to pay of the qth user of the treatment.

The overall incremental effectiveness in terms of improved health for those consuming the technology is therefore:

$$E = \int_{h(p)}^{\infty} h_i \, \mathrm{d}F(h) = q(p)E[h|h \geq h(p)] \tag{5}$$

The gross consumer benefit or monetized value of improved health among those consuming the technology is similar to that described above and is equal to:

$$G = q(p)E[g(h)|h \geq h(p)] = q(p)E[p(y)|y \leq q(p)] \tag{6}$$

The last equality simply states that the gross consumer benefit at a given incremental price p is equal to the overall demand multiplied by the average willingness to pay among consumers. Given the incremental amount spent on this technology is simply $pq(p)$, in this framework, the commonly employed incremental cost-effectiveness ratio (ICER) can be written as:

$$\mathrm{CE} = \frac{q(p)p}{E} = \frac{p}{E[h|h \geq h(p)]} \tag{7}$$

Eq. (7) states that the average incremental CE ratio is simply the ratio of incremental costs per person to the average incremental health benefit among those consuming the technology. If h is interpreted in terms of QALYs, then Eq. (7) is the incremental cost per QALY. The analogous incremental cost–benefit ratio (ICBR) simply monetizes the improvement in health:

$$\mathrm{CB} = \frac{q(p)p}{G} = \frac{p}{E\,[p(y)|y \leq q(p)]} \tag{8}$$

The denominator in the first term is the total gross consumer benefit G, i.e. the total incremental willingness to pay among those consuming the technology. The denominator in the second term is the gross consumer benefit per person.

In this framework, we argue that CE technology evaluation is implicitly related to consumer surplus, as both CE and CB ratios focus on how much patients benefit beyond what is spent on the technology after it has been developed.[7] Despite the several forms of such criteria developed to date – e.g. cost-effectiveness, cost-utility, or cost–benefit – their basic goal seems to be to determine whether increased health care spending on new technologies is justified by societal, health plan, or patient benefits in terms of improved health. In this sense then, we interpret Eq. (8) as being explicitly related to consumer surplus—CB measures relate consumer benefits as a ratio to spending while consumer surplus expresses it as a difference between the two. In particular, the CB ratio can therefore be recast in terms of consumer surplus as:

$$\mathrm{CB} = \frac{pq}{G} = \frac{pq}{Z + pq} \tag{9}$$

Importantly, consumer surplus measures based on demand also reflect the *incremental* benefits associated with a treatment since the demand curve delivers the willingness to pay given the availability and prices of other substitutes.

The main implication of relating CE or CB measures to consumer surplus is that standard economic analysis can be brought to bear on how those measures relate to economists' standard measures of welfare. Consider, for example, the implication for static efficiency of a reduction in the price of a treatment following the introduction of generic competitors. Standard notions of ex-post efficiency would imply that such decreases in price would be met with increases in consumer surplus and improvements in static efficiency, as price is driven closer to marginal cost and output is increased. We argue, however, that such decreases in price and improvements in static efficiency may not be monotonically related to either measures of CE or CB. For example, if $\varepsilon > 0$ is the elasticity of demand, one can easily show that:

$$\frac{\mathrm{dCB}}{\mathrm{d}p} > 0 \quad \text{iff} \quad \left[1 - \frac{1}{\varepsilon}\right] < \frac{pq}{G} = \frac{1}{\mathrm{CB}} \tag{10}$$

Eq. (10) states that decreases in price will only surely lead to decreases in the CB ratio when demand is inelastic ($\varepsilon < 1$). When demand is elastic, decreases in CB will be consistent with improvements in static efficiency only when spending comprises

[7] The implicit consumer surplus estimation of CE analysis differs from traditional economic analysis—the latter typically attempts to assess consumer surplus by estimation of demand schedules, by observing changes in demand during supply-induced price changes. Importantly, the demand curve for a good summarizes the value to consumers of both its observed and unobserved attributes. On the contrary, estimates of consumer surplus based on cost-effectiveness or cost–benefit analysis are typically formed indirectly by monetizing *observable* consumer benefits, e.g. by use of value of life estimates to estimate the gross consumer benefit from mortality reductions.

a sufficiently large share of the gross value associated with a technology. The main intuition is that reductions in price will, of course, lead to decreases in CE and CB when the average benefit to those using the treatment are held fixed. However, reductions in price increase access and therefore may lower the average level of benefits among users, depending on how elastic use is to price and how much less effective (either in health or monetized terms) the treatment is among the new users. In this case, reductions in price would improve static efficiency but might *increase* CE and CB ratios. The main implication of this result, then, is that improvements in static welfare may or may not be related to improvements in CE or CB (i.e. declines in the ratio). To the extent that the demand for health care is relatively inelastic, however, improvements in the CE or CB ratio will be consistent with improvements in static efficiency.

2.2. Public technology adoption using cost-effectiveness

The previous section illustrated that when the supply and demand price are equal (e.g. when there are no subsidies to demand), standard technology assessment measures may or may not be consistent with standard notions of static economic efficiency, depending on the elasticity of demand and the share of gross surplus comprised by spending. This section illustrates how third-party reimbursement (or other forms of demand subsidization) can be incorporated into the CE framework derived earlier and demonstrates how CE measures may impact dynamic welfare as well.

2.2.1. CE in public reimbursement

First note that the market context in which CE analysis is conducted, and the results applied, matters greatly. For example, in a standard market economy, it would be extremely surprising if correctly measured CB ratios were found to be above unity. As an illustration, consider a private market for health care without public or private insurance, as might exist for certain elective surgeries in the US, e.g. plastic surgery. A new plastic surgery technology would have a CB ratio below unity (if estimated correctly) if individuals bought the product only when their valuation of it exceeded the price. This, of course, would always be predicted under standard demand analysis.

The natural context, then, in which CE-based technology adoption has been and should be applied is in settings of third-party reimbursement, where patients bear only a fraction of the price of treatment with the remainder being reimbursed by health plans or public payers. Perhaps the most common example is CE thresholds which dictate that only those technologies will be adopted whose incremental cost per QALY falls below a given level. As discussed, such thresholds have been argued to be implicitly important in technology adoption in many countries, though so far less so in public reimbursement decisions in the US.

To formalize how third-party reimbursement relates to our CE measures developed earlier, suppose that those patients deemed clinically eligible for treatment may utilize the treatment at a pre-determined demand price $p_{\mathrm{d}} < p$, where p is the supply price paid to producers. For simplicity, assume that the demand price is equivalent to the marginal cost of production, $p_{\mathrm{d}} = c$. For a given CE threshold T, this technology would be adopted if the incremental price per QALY among those using the treatment were less than the threshold:

$$\mathrm{CE} = \frac{p}{E[h|h \geq h(c)]} < T \tag{11}$$

The denominator in expression (11) is the average incremental health benefit per person under the demand price c. The numerator is simply the per-person supply price paid by the third party payer (in this case, the government) to the manufacturer of the treatment.

In the context of third-party reimbursement, technologies whose prices are lower are deemed more cost-effective and are therefore more likely to get reimbursed. Moreover, expression (11) reveals that when the threshold rule is given, firms will have an incentive to raise price so that the threshold binds:

$$p = TE[h|h \geq h(c)] \tag{12}$$

This has several implications. First, the threshold level can and does determine the price charged by manufacturers for a treatment, i.e. the supply price. In this sense, it is closely related to other mechanisms to fix price, e.g. price controls which fix price directly or rate-of-return regulations which limit price through limiting excess profits. CE thresholds are supply price controls in the sense that if price determines costs to government payers and the health effects determine effectiveness, adoption policies based on cost-effectiveness are adoption policies based on price.[8] Second, reductions in the threshold level will necessarily lead to the adoption of more cost-effective technologies. This may come at a cost to either manufacturers, who face reduced profits from lower prices, or at a cost to patients, who must on average demonstrate larger health benefits than before. Therefore, when CE thresholds are set low, prices may often have to be reduced if the treatment is to be competitively provided after adoption—if prices are not reduced, patient costs must be increased (or clinical eligibility more restricted) so that the average health benefit among those using treatment is raised to maintain the threshold.

[8] Importantly, the extent to which the price of a treatment can be raised to meet a given threshold depends on whether services that accompany the treatment are incorporated into the CE analysis. For example, in CE evaluations, the incremental cost of a drug may often incorporate the cost of the drug itself, as well as the cost of physician visits, routine blood tests to monitor side effects (e.g. liver function tests for statin medications), and so on. To the extent that these non-drug costs are important in reaching the threshold, manufactures would have less of an ability to raise price.

The main mechanism by which CE criteria may impact dynamic welfare is by affecting the rewards to innovation. Reductions in CE thresholds would naturally lead to the adoption of more cost-effective technologies (i.e. lower CE ratio) by favoring reductions in prices that would allow previously cost-ineffective technologies to become cost-effective relative to the threshold. To the extent that such reductions in price will alter the incentives to innovate, dynamic efficiency may be lowered or even enhanced depending on whether the optimal incentives to innovate already exist. Therefore, while lowering CE thresholds may lead to improvements in static welfare through lower prices, we argue that the impact on dynamic welfare is ambiguous and will depend heavily on the model of innovation and the model's implications for whether optimal incentives for innovation already exist.

2.2.2. CE and dynamic efficiency

Consider the case of a single monopolist investing in R&D who receives a share, a, of the social surplus W, where $0 \leq a \leq 1$. Let $r(k)$ be the level of R&D that maximizes expected payoffs for any hypothetical ex-post prize, k:

$$r(k) = \arg\max_{r}[x(r)k - r] \tag{13}$$

Then, $r(aW)$ represents the R&D undertaken when those investing in R&D maximize expected profits. If profits drive R&D investments, the induced expected social surplus is:

$$D(a, W) = x[r(aW)]W - r(aW) \tag{14}$$

This expression directly highlights the well-known implication that first-best dynamic efficiency occurs when those undertaking the costs of R&D have incentives that are properly aligned with society, which is true when social surplus is entirely appropriated as profits, i.e. $a = 1$ (see e.g. Arrow, 1961; Tirole, 1988). In other words, the key factor driving dynamic inefficiency in this model is that profits are less than social surplus. More importantly, the size of the consumer surplus, focused on by CE criteria, is what drives a wedge between profits and social surplus and hence leads to under-investment in R&D—this is analogous to the argument that patents hurt static efficiency but raise dynamic efficiency by engaging in similar substitution.

The main issue in this particular model of R&D, then, is that $a < 1$, which will typically be the case with linear pricing.[9,10] The important implication of this is that to the extent that linear pricing does not allow a monopolist to capture the entire social surplus arising from their product, low CE thresholds which attempt to reduce prices may have negative effects on dynamic welfare, though static welfare will be improved. This conclusion relies on two points: the first is that reductions in CE thresholds will in fact lead to improvements in static efficiency through lowering prices and the second is that dynamic welfare is increasing in innovator appropriation. The second is a finding that is consistent with this model of innovation and as we discuss later, may not apply generally.

This model illustrates that the CE associated with the ex-post market for a technology may not be clearly and monotonically related to measures of either static or dynamic efficiency. Indeed, in this model, the dynamically optimal allocation of surpluses implies that the consumer surplus should be *minimized*, as opposed to maximized under a CE criteria, to enhance dynamic efficiency. *In this case, dynamic efficiency dictates that a technology should just break even ex-post* (i.e. CB = 1). The dynamically efficient maximization of the CB ratio in this model is a direct implication of the classic problem of non-appropriation by innovators leading to under-investment in R&D. Importantly, note that setting the CB ratio to unity in this context still maximizes *patient health* though not consumer surplus. This is because full appropriation ensures that the technology is fully adopted by consumers; under price discrimination, there are no output effects of monopolies.

There are, of course, important instances in which full appropriation of social surplus by producers may not be dynamically optimal. In these cases, depending on whether optimal incentives for innovation already exist, CE thresholds may play an important role in reducing or even enhancing dynamic welfare.[11] For example, full innovator appropriation may not be optimal when R&D is characterized by so-called patent racing. Since competitive R&D leads to an equilibrium level of R&D that is determined by the average (rather than marginal) profit associated with entry, non-appropriation may enhance efficiency by taxing the over-provision of R&D. This may be particularly relevant to the debate over excessive R&D into so-called "me-too" drugs in the pharmaceutical industry. If rewards for innovation are already too high, so that firms over-invest in R&D, CE thresholds which lower these rewards through reductions in price may be warranted.

Another important instance in which non-full appropriation may be optimal is when publicly funded R&D comprises a significant portion of total R&D, as is common in US health care through the NIH. Since the dynamically optimal level of *total* R&D is still $r = r(W)$, the presence of publicly funded R&D implies that the optimal private R&D (and hence, appropriation) should be lowered accordingly. Since the marginal product of private R&D is decreasing in the level of subsidized R&D, private R&D (and hence appropriation) optimally falls as its public counterpart increases. As in the case of competitive

[9] For example, when production is characterized by constant returns to scale, it can be shown that monopolists facing either linear or constant-elasticity demand earn profits that are proportional to the *potential* social surplus, here defined as the surplus obtained at the competitive output (rather than the monopoly quantity). Specifically, $a = 1/2$ in the case of linear demand and $a = [(\varepsilon - 1)/\varepsilon)]^{\varepsilon}$ under constant elasticity of demand.

[10] Interestingly, profits may even exceed the *private* social surplus (i.e. the gross benefit to consumers net of costs of production) when there are external effects in consumption. See, for e.g., Philipson et al. (2006) who discuss R&D under altruism in health care.

[11] We summarize some of these instances below, which are considered in more detail in the NBER Working Paper version of this paper (NBER Working Paper #12016).

research investments, the less-than-full appropriation that is optimal under public subsidization of R&D implies that lower CE thresholds may be preferred in settings where total R&D relies on a large public component. And of course, increases in CE thresholds may not be practically feasible as well, particularly when fixed public budgets impose a necessary trade-off between static and dynamic efficiency that precludes the raising of CE thresholds.

2.2.3. Optimal CE thresholds

The examples discussed above illustrate that the optimal level of surplus appropriation, a^*, by innovators may vary greatly and will generally depend on the particular model in question. Regardless of what view one may take on whether current incentives for innovation are too high or too low, we argue that existing CE thresholds T can be used to achieve any desired level of appropriation for producers, i.e. $T = T(a)$. Optimal levels of appropriation a^*, once determined, can therefore be used to devise optimal CE thresholds $T^* = T(a^*)$. This is most easily seen with the *monetized* equivalent of the threshold rule presented earlier. Consider, for example, the CB adoption rule below:

$$\text{CB} = \frac{p}{E[p(y)|y \leq q(c)]} \leq T \leq 1 \tag{15}$$

This rule simply states that only those technologies will be adopted whose per-person price falls below the average monetized health benefit (i.e. willingness-to-pay) among users by an amount consistent with the CB cutoff, T. In terms of appropriation, a, this threshold rule can generally be rewritten:

$$\text{CB} \leq \frac{a\bar{W} + c}{\bar{W} + c} = T(a) \tag{16}$$

where c is both the demand price and the constant per-person cost of producing the drug and $\bar{W}$ is the average surplus per person consuming the drug, i.e. $E[p(y)|y \leq q(c)] - c$. The main implication of this interpretation is that CE thresholds not only identify the level of innovator appropriation for a given technology but can be set to achieve any desired level of appropriation. For example, if full appropriation is dynamically optimal (i.e. $a^* = 1$), a threshold rule $T(a^*)$ which sets CB = 1 would achieve this outcome by mandating that innovators receive a price equal to the average willingness-to-pay for their product. Moreover, static efficiency would not be comprised since the competitive output is obtained. Therefore, if the average willingness-to-pay for a QALY is US$ 100,000, any threshold rule which does not reimburse technologies costing more than US$ 100,000 per QALY would be dynamically inefficient in this extreme case. Similarly, the threshold may be set at lower levels of appropriation – e.g. at US$ 50,000 per QALY – if too much incentive for innovation already exists.

Intuitively, CE threshold adoption rules therefore function as subsidies which effectively separate the demand price (paid by consumers) from the supply price (paid by the government to manufacturers) and have the potential to maximize both static and dynamic efficiency by ensuring access and incentive for innovation simultaneously. In the simplest case of full appropriation, dynamic efficiency is achieved when the supply price equals the average willingness-to-pay among users—when full appropriation is not optimal, the optimal supply price is adjusted downward to reflect the share of the average willingness-to-pay that is optimally appropriated. In either case, static efficiency is obtained since the demand price faced by consumers is fixed as well at marginal cost c.

In this sense, our interpretation of CE thresholds is related to work by Garber et al. (2006) and Lakdawalla and Sood (2005). The work by Garber et al. (2006) focuses on the effects of insurance on incentives for medical innovation. Their basic point is that because demand subsidization raises monopoly profits, the presence of insurance may provide too much incentive for innovation. In this case, efforts to reduce profits either through direct limits on pricing or advocacy of lower CE thresholds, may enhance efficiency by limiting otherwise excessive incentives for innovation. In fact, insurance may lead to profits that even *exceed* the social surplus generated by a new product, which in most instances implies that dynamic efficiency would be improved by advocating lower CE thresholds.[12] In the context of our model, this scenario would arise if the monopoly price paid by the government or insurer, p, exceeded the average willingness-to-pay $E[p(y)|y \leq q(c)]$ that arises when patients pay only a portion of the supply price, namely c. Therefore, even in the extreme case where full appropriation is dynamically efficient, this may lead to excessive R&D since appropriation would be more than full. The natural solution would therefore be to limit CB thresholds to be no greater than unity and substantially less if strong incentives for innovation already exist.

A CE threshold which ensures efficient utilization of a technology, while providing appropriate incentive for innovation, is also similar to a two-part pricing contract between a government and a manufacturer. For example, in the case of full surplus extraction, the government may pay the manufacturer a lump sum amount equivalent to the total consumer surplus arising from the treatment in exchange for the ability to purchase the treatment at some lower price, say marginal cost c. This is a point stressed by Lakdawalla and Sood (2005), in which the health insurance contract is viewed as a two-part tariff which can allow for statically efficient extraction of consumer surplus by innovators of the utilized treatment. In their model, as in our interpretation of CE thresholds, the extent to which innovators are able to capture surplus from consumers depends on the relative bargaining power of both parties—when governments possess strong bargaining power in negotiations over price, the presence of low CE thresholds may lead innovators to actually appropriate little surplus, which may or may not

[12] An important exception to this is when social surplus is made up not only of consumer surplus but of non-consumer surplus as well, e.g. due to altruism. In this case, profits may efficiently exceed *consumer surplus* (Philipson et al., 2006).

be dynamically efficient depending on the particular model of innovation. The main point, however, is that regardless of what view one takes on the optimal level of appropriation, existing CE thresholds may be optimally designed to ensure both efficient ex-post utilization and dynamically efficient incentives for innovation.

3. Inferring surplus appropriation for a sample of medical technologies

The previous discussion highlighted the importance of surplus appropriation by innovators to dynamic efficiency and the role that existing CE thresholds can play in ensuring both statically efficient access to treatment and dynamically efficient rewards for innovation. As the ability of innovators to appropriate the surplus of their innovations is central to dynamic efficiency, this section illustrates how existing CE estimates can be combined with assumed parameters to exemplify the level of incentives that already exist. The main simplifying assumption is a single supply and demand price faced by both patients and a monopoly producer. This ignores the role of third-party reimbursement (which would lead to separate supply and demand prices) and the ability of the monopolist to charge different prices to different patients, i.e. price discriminate. The purpose of this exercise, therefore, is not to advocate a preferred estimate of existing levels of surplus appropriation, but rather to demonstrate how the large CE literature may be used to inform us on the general magnitude of these levels. To the extent that our calculated incentives are deemed either too low or too high compared to dynamically efficient levels, CE thresholds may be appropriately raised or lowered to improve dynamic efficiency.

3.1. Inferring observed surplus appropriation

In order to illustrate the extent to which innovators are able to capture the net social value of their innovations, we begin by discussing conditions under which the often estimated incremental CE of a given technology may be combined with specific parameter assumptions to infer the share of social surplus appropriated by producers of that technology. Recall from our earlier discussion that the observed CB ratio of a technology can be written as $\text{CB}=p/E[p(y)|y\leq q(p)]$, where $E[p(y)|y\leq q(p)]$ is the average willingness-to-pay among those users whose monetized health benefit exceeds the incremental price paid, p. This, of course, neglects the role of third-party reimbursement which would result in a demand price that is below the supply price. Given a constant marginal cost of production c, the observed surplus appropriation is simply $a_{\text{obs}}=[p-c]/[E[p(y)|y\leq q(p)]-c]$, where the numerator is the manufacturer's profit per person and the denominator is the average social surplus per person available to be captured by the innovator. If $m=p/c$ is the markup of the monopolist's chosen price above costs, it is straightforward to show that for a given price, the observed appropriation may be written as a function of m and the CB ratio as in:

$$a_{\text{obs}}=\frac{m-1}{(m/\text{CB})-1} \tag{17}$$

This expression demonstrates that technologies with low CB ratios implicitly support low levels of observed surplus appropriation. Moreover, when free-entry is possible and firms earn zero profits (price = average cost), surplus appropriation is zero.

The general point, then, is that with information on the degree of market power in an industry, one can use monetized versions of commonly reported CE estimates to infer the degree of appropriation by producers of the relevant technology. For example, consider a technology which costs nothing to produce after it is discovered, commands an incremental price (over some baseline therapy) of US$ 20,000 per person, and yields an average incremental increase in life-expectancy among users of one year. If the average willingness-to-pay for an additional year of life is US$ 100,000, the average ex-post profits and surplus per person would be US$ 20,000 and US$ 100,000, respectively, implying a surplus appropriation of 20%.

3.2. Inferring potential surplus appropriation

With the additional assumption that the distribution of health benefits $F(h)$ and the willingness-to-pay for health $g(h)$ generate a constant elasticity of demand, even less information is needed to infer the level of appropriation from CE estimates. In Appendix A, we show how a technology's CB ratio *alone* identifies its elasticity of demand, which in turns identifies the share of surplus appropriated by the producers of that technology. These assumptions also allow us to distinguish between appropriations of two types of surpluses: observed versus potential. The *observed* surplus discussed above is the surplus which obtains at the market price—the total gross benefit is calculated among those users whose willingness-to-pay for a technology's incremental health benefit exceeds the price. For example, for a monopoly price p_{m}, the appropriation of observed surplus is simply $\Pi(q(p_{\text{m}}))/W(q(p_{\text{m}}))$. Alternatively, the *potential* surplus is that which results if the market quantity is determined competitively and hence relates to the total potential surplus available to an innovator. For a given price $p>c$, the appropriation of potential surplus, a_{pot}, is $\Pi(q(p))/W(q(c))$. The appropriation of potential surplus can therefore be written as:

$$a_{\text{pot}}=\frac{(p-c)q(p)}{[E[p(y)|y\leq q(c)]-c]q(c)} \tag{18}$$

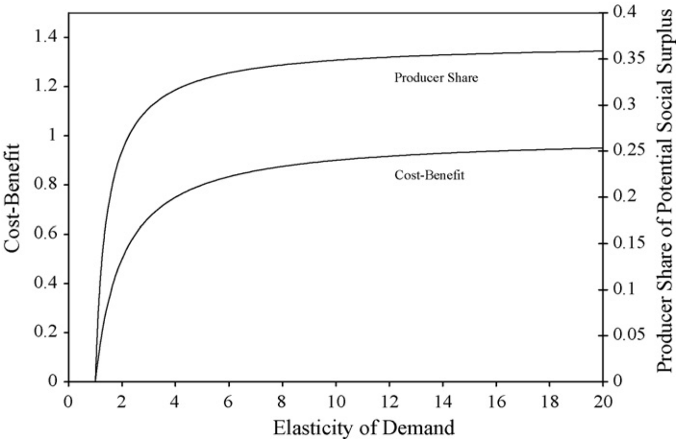

Fig. 1. Elasticity of demand and producer share implied by CB estimates.

where the denominator is the potential surplus available to the innovator and the numerator is the total profit when the quantity supplied is determined by the price p.

Importantly, the size of profits relative to the *potential* social surplus is most relevant to dynamic policy and to the use of CE thresholds to achieve that optimal policy. Since there is a deadweight loss associated with monopoly pricing, the potential surplus from an innovation exceeds the observed surplus. Consequently, calculations of 'surplus' appropriation based on observed surplus will underestimate the level of appropriation by producers of a given technology. Moreover, as the preceding discussion demonstrates, CE thresholds which ensure efficient ex-post utilization of a technology deliver the potential surplus out of which profits to innovators can be allocated based on the dynamically optimal level of appropriation.

To infer the level of potential surplus appropriation from a technology's CB, consider then the common model where the monopolist's variable costs exhibit constant returns, $C(y) = cy$, and there is a constant elasticity inverse demand curve $p(q) = x/q^{1/\varepsilon}$, where $\varepsilon > 0$ is the elasticity of demand with respect to price and x is a scale factor that shifts demand outward. If q_{w} and q_{m} denote the welfare maximizing and monopoly output, respectively, Appendix A shows that the CB ratio under monopoly pricing satisfies:

$$\mathrm{CB} = \frac{p(q_{\mathrm{m}})q_{\mathrm{m}}}{G(q_{\mathrm{m}})} = \frac{\varepsilon - 1}{\varepsilon} = \frac{c}{p(q_{\mathrm{m}})} \tag{19}$$

In other words, a technology's CB is directly related to the familiar percentage markup of price over marginal cost. In addition, the share of *potential* surplus appropriated as profits under optimal monopoly pricing is related to the output expansion due to competition.[13] That is,

$$\frac{\Pi(q_{\mathrm{m}})}{G(q_{\mathrm{w}}) - q_{\mathrm{w}}c} = \frac{q_{\mathrm{m}}}{q_{\mathrm{w}}} = (\mathrm{CB})^{\varepsilon} \tag{20}$$

This interesting result states that, counter-intuitively, the more a monopolist restricts output, as perhaps estimated by patent expirations, the *less* of the surplus it appropriates.[14] Note that as the elasticity approaches unity (below which profits are infinite) from above, the profits, themselves, rise but as a share of social surplus go to zero.[15] This occurs because the non-appropriated consumer surplus rises faster than profits as the elasticity falls. Moreover, as market power declines and elasticity approaches infinity, the share of social surplus appropriated as profits tends to roughly 37%. Finally, there is a direct positive relationship between the CB ratio and innovator appropriation.

Under these assumptions, a given estimated CB ratio implies a specific constant elasticity of demand, which in turn implies the degree to which a firm appropriates its potential social surplus. More generally, the above relationship between CB and surplus appropriation can be used to infer the share of potential surplus appropriated by those producers whose technologies are examined in existing CE studies. Fig. 1, below, graphs the relationship between surplus appropriation, cost–benefit, and market power (interpreted as a reduction in the elasticity of demand). As market power decreases, the producer's share of potential social surplus approaches slightly more than a third, while CB approaches 1. As described earlier, CB is bounded from above by unity since individuals only purchase goods for which the incremental benefits exceed the incremental costs.

[13] It is straightforward to show that the share of observed surplus appropriated by producers is $(\varepsilon - 1)/(2\varepsilon - 1)$, which is greater than the potential surplus appropriated.

[14] This result may not be unique to this particular demand structure. For a linear demand curve, it is well known that monopoly output is half the competitive output and that a monopolist always appropriates half the surplus, so that the surplus condition above holds.

[15] It may even be that demand and cost parameters do not affect the share of surplus appropriated by the producer. This is the case when demand is linear (as often estimated) and there are constant returns to scale in production, in which case the share appropriated by producers is always two thirds.

Table 1
Calculated producer share of potential social surplus for several cost-effective technologies

Intervention	Spending per QALY (US$)	CB		Producer share of potential social surplus	
		US$ 50,000	US$ 100,000	US$ 50,000	US$ 100,000
Captopril therapy	4,000	0.08	0.04	0.06	0.03
Ticlopidine therapy	48,000	0.96	0.48	0.36	0.24
Mesalamine therapy	6,000	0.12	0.06	0.09	0.04

Note: CE and producer share of surplus are presented for two, separate values of an additional quality adjusted life year. Description of interventions: (1) Captopril therapy in patients with myocardial infarction, (2) Ticlopidine therapy in patients with high risk stroke, and (3) Mesalamine to maintain remission in Crohn's disease. For a more detailed description, see Neumann et al. (2000).

3.3. *Calculations of observed and potential surplus appropriation*

The above figure illustrates how one can potentially use estimates of cost-effectiveness from the large health economic literature to infer the share of surplus appropriated by producers of the relevant technology. As shown, under an assumption of monopoly pricing and constant elasticity demand, the observed CB ratio identifies both the elasticity of demand and the potential surplus appropriation. We exemplify this general identification strategy using CE estimates from the literature. Neumann et al. (2000), for example, review the cost-effectiveness of pharmaceuticals using the established "cost-utility" method which focuses on costs per QALY gained and therefore concern both the prolongation and quality of life. The authors note that while no accepted standards exist for how much benefit a technology must confer to be deemed a "good value," the range between US$ 50,000 and US$ 100,000 per QALY has typically been a benchmark. In the context of our framework, this value (or range) is the gross benefit to consumers of a technology which leads to an additional quality adjusted year of life.

Table 1 presents the spending required to obtain an additional QALY for several patent-protected interventions (under patent at the time of the original study) reviewed by the authors. For example, an intervention with an incremental price of US$ 1000 that leads to an increase in 0.2 QALYs requires the same incremental spending per QALY as an intervention with an incremental price of US$ 5000 that leads to an additional QALY. While the magnitude of gross benefit differs across the two interventions, the gross benefit per QALY is the same (namely in the range described above). Thus, assuming the gross benefit arising from an additional quality adjusted year of life is between US$ 50,000 and US$ 100,000, we can compute monetized versions of these CE estimates, as well as the implied shares of potential social surplus appropriated by producers (given by Expressions (18) and (19)).

Table 1 demonstrates that those technologies deemed to be extremely cost effective may also result in low surplus appropriation by producers. For example, the highly cost effective Captopril therapy results in roughly 3–6% of potential social surplus going to producers.

While Table 1 presents calculations of the producer share of social surplus for only three interventions, cost-effectiveness estimates from a larger sample of interventions could be used to infer the distribution of producer shares. We illustrate this using data from over 200 published cost-utility analyses contained in the Harvard Cost-Effectiveness Analysis (CEA) Registry. Importantly, the studies included in the CEA Registry are not random and therefore cannot be expected to yield a generalized distribution of innovator appropriation.[16] Nonetheless, they do provide an initial benchmark to *illustrate* the levels of appropriation that may already exist. Including analyses from 1976 to 2001, the Registry reports the spending per QALY of various interventions compared to benchmark comparator groups. This spending per QALY can in turn be used to calculate the share of potential social surplus appropriated by the producer of that technology, as in Table 1 above.[17] This can be compared to calculations of the producer's actual appropriation, identified by the technology's CE and average mark-up as in expression (16) above. As a simplifying assumption, we apply existing estimates of markups for brand-name drugs (as estimated from patent expirations) to approximate variable costs as 15% of sales (see e.g. Caves et al., 1991).[18]

Fig. 2 plots the distribution of observed and potential producer shares for the interventions considered. Because the studies included in the Registry are not a random sample of all technologies, however, we can only infer the distribution of producer shares conditional on inclusion into the Registry.

Since the constant elasticity of demand assumption predicts a producer appropriation of social surplus of no more than 37%, all interventions considered in Fig. 2 have estimated producer shares less than this amount. The median intervention requires a spending per QALY of roughly US$ 19,000, which corresponds to a producer share of potential (actual) social surplus of nearly 13% (17%).[19] Approximately 25% of the interventions considered have estimated potential appropriations of less

[16] The Registry is not limited to only pharmaceutical interventions. More detailed information can be found at: http://www.hsph.harvard.edu/cearegistry/.

[17] For these calculations, we assume the gross benefit of an additional QALY to be US$ 100,000. Consequently, we limit our attention to those interventions with published costs of less than US$ 100,000 per QALY gained.

[18] Caves et al. (1991) estimate that with 20 generic competitors, the ratio of prices between generic- and brand-drugs is roughly 17%. We use the price of generic drugs as an upper bound of the marginal costs of production.

[19] If the gross benefit of an additional QALY is assumed to be US$ 50,000 (rather than US$ 100,000), the median intervention has an implied producer share of social surplus closer to 20%.

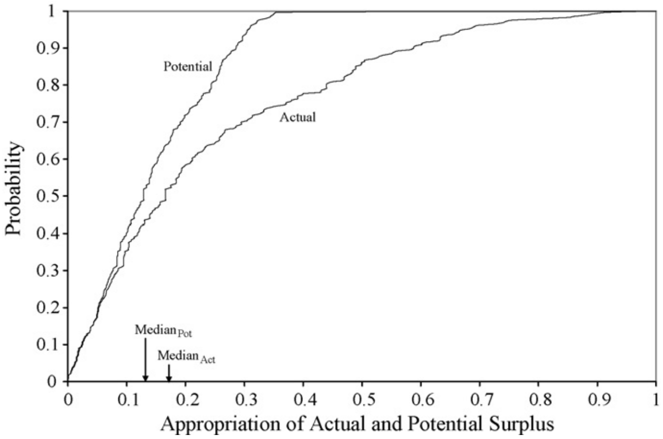

Fig. 2. Cumulative distribution of actual and potential social surplus appropriation.

than 7%, while 75% have appropriations less than a fourth. Moreover, 75% of the interventions have an actual appropriation of less than 40%. These calculations can be compared to directly estimated levels of appropriation for producers of HIV/AIDS drugs presented in Philipson and Jena (2005). This analysis values the gains in HIV/AIDS survival in the past twenty years and calculates appropriation using data on sales of HIV/AIDS drugs during that same period. The main finding is that firms appropriated roughly a twentieth of the social surplus generated by these technologies, placing HIV/AIDS therapies at the 20th percentile of appropriation in the CEA Registry.

Our illustrative calculations of appropriation can be compared to alternative, theory-based methods of calculating this share. Specifically, given the previously described relationship between the elasticity of demand and the share of potential social surplus appropriated by innovators, one can use information on price reductions after patent expiration to estimate patent-protected markups (Caves et al., 1991). These markups identify the elasticity of demand for patent-protected drugs, which for simplicity is assumed to be constant, and thus the share of surplus allocated to the producer. In particular, the larger is the price reduction upon patent-expiration, the lower is the elasticity and the *smaller* is the share of surplus allocated to the producer. Existing estimates suggest that price reductions are on the magnitude of 85%, implying a demand-elasticity around 1.17. This elasticity implies a producer share of social surplus of 10%, which is well within the range calculated above. This is true even though prices for patented drugs are often high, presumably due to the inelastic nature of demand.[20]

4. Concluding remarks and future research

As CE analysis has begun to be more heavily considered for public technology adoption, less emphasis has been placed on the effect of such criteria on the behavior of innovators who make the technologies available in the first place. We analyzed the relationship between CE analysis and economic efficiency and found that greater CE of a technology may be inconsistent with static efficiency, dynamic efficiency, as well as improved patient health. Intuitively, CE thresholds closely resemble any number of mechanisms that fix price and therefore have similar predictions for both static and dynamic economic efficiency. These mechanisms include explicit price controls, rate of return regulations that limit the overall rate of return a company's portfolio of drugs may earn from a given public purchaser, as well as global spending budgets that earmark fixed levels of spending for a given number of patients who demand medical treatments. The point of our analysis is not that the impact of "price controls" on static and dynamic efficiency are not well-recognized – they certainly are – but that widely advocated CE policies can be thought of in much the same way as existing price control policies and thus have the same implications for economic efficiency.

As the ability of innovators to appropriate the surplus of their innovations is central to dynamic efficiency, we illustrated how existing CE estimates can be used to infer this level of appropriation and found it to be potentially low. That being said, the major point of our analysis is not that existing incentives for innovation are too high or too low, but that existing and future CE-policies for technology adoption do have implications for dynamic welfare which should be considered in assessing the merits of such policies. And because CE thresholds closely relate to other mechanisms to control price, our illustrative results on appropriation may be useful in determining the merit of such policies as well.

[20] Somewhat counterintuitive, levels of appropriation may be low despite the seemingly high prices that characterize many medical technologies. Although it is natural to suggest that this is due to lack of profits and market power, even with free pricing and nearly-inelastic demand, the share of social surplus allocated to producers may be small. This is most easily illustrated by the constant elasticity case, in which a producer share of social surplus of 5% is consistent with monopoly pricing under a demand curve that is almost as inelastic as it can be, $\varepsilon = 1.08$. In fact, higher prices induced by lower elasticities of demand may lead to *less* surplus captured by inventors. Although profits, of course, rise as the elasticity of demand falls, the *share* of social surplus appropriated by innovators may fall if the non-appropriated consumer surplus rises faster.

Several issues may be important in generalizing our conclusions and are therefore suitable for future research. The first concerns the interpretation of CE analysis in a non-monopoly context; the field of "industrial organization of technology adoption under CEA" needs to be better understood. Another concern is the effect of altruism, which seemingly motivates much of public financing, on optimal technology adoption and the efficient form of surplus appropriation.[21] A third concerns the effect of ex-post inefficiencies such as moral hazard. Fourth, the impact of the joint demand of physicians and patients on observed levels of CE must be examined further. Fifth, the effect of improved treatment on disease prevalence, whether through increased life-expectancies among infected individuals or increased risky behavior (due to lower costs of infection induced by treatment) among non-infected individuals, must be considered (see e.g. Philipson, 2000). Sixth, the role of public funding, comprising almost half of US medical R&D spending, on the optimal degree of appropriation is not well understood. While much basic research in the US is financed by tax-payers (mainly through the NIH), little analysis exists on the implications of that for optimal appropriation.[22]

Lastly, in conducting cost-effectiveness assessments, better methods should be developed to measure actual production costs rather than marked-up prices which proxy for such costs. For example, generic prices of close substitutes may be used to measure production costs rather than using marked-up prices as seems to be universally done in practice.[23] Despite this important issue, however, even if one could measure costs perfectly and did not need to estimate unobserved costs by observed prices, our analysis implies that traditional cost-effectiveness assessments would be concerned with the wrong measure, namely, total ex-post surplus. The reason is that the division of the surplus, not the total surplus, is what matters for dynamic R&D policy. Static efficiency is raised by total surplus being larger ex-post, but innovation incentives may be harmed if such an expansion of surplus reduces profits.

The overall conclusion we hope to emphasize is that much more research needs to be done on the implications for innovation of using the implicit price controls of CE-threshold policies for technology adoption. This is particularly true for empirical analysis where the static effects of lowering price, and hence lowering CE values, are better understood than the dynamic consequences for future technologies. The choice should not be seen as one between cheap or expensive technologies once marketed – as CE adoption suggests – but one between an initially expensive technology and no technology, the latter which would entail higher real prices for producing a healthy life.

Appendix A

Assume a constant elasticity demand function and constant returns to scale as in:

$$p(q) = \frac{x}{q^{1/\varepsilon}}; \qquad C(q) = cq$$

where $\varepsilon > 0$ is the elasticity of demand with respect to price, and x is a demand shifter. This results in an optimal quantity and price of

$$q_{\mathrm{m}} = \left[\frac{c\varepsilon}{x(\varepsilon-1)}\right]^{-\varepsilon}; \qquad p_{\mathrm{m}} = \frac{c\varepsilon}{\varepsilon-1}.$$

The gross consumer benefit $G(q_{\mathrm{m}})$ can be expressed by the following formula:

$$G = \int_0^{q_{\mathrm{m}}} p(q)\,\mathrm{d}q = \frac{x\varepsilon}{\varepsilon-1}(q_{\mathrm{m}})^{\varepsilon-1/\varepsilon}.$$

Similarly, the maximized profit can be written as:

$$\Pi = p(q_{\mathrm{m}})q_{\mathrm{m}} - cq_{\mathrm{m}} = \frac{cq_{\mathrm{m}}}{\varepsilon-1}$$

We can now determine the share of profits in potential social surplus, i.e. the social surplus that obtains in perfect competition with $p = c$. Specifically,

$$\frac{\Pi(q_{\mathrm{m}})}{G(q_{\mathrm{w}}) - q_{\mathrm{w}}c} = \frac{(cq_{\mathrm{m}})/(\varepsilon-1)}{(x\varepsilon/\varepsilon-1)(q_{\mathrm{w}})^{\varepsilon-1/\varepsilon} - q_{\mathrm{w}}c} = \frac{(cq_{\mathrm{m}})/(\varepsilon-1)}{q_{\mathrm{w}}((x\varepsilon/\varepsilon-1)(q_{\mathrm{w}})^{-1/\varepsilon} - c)}$$

$$= \frac{(cq_{\mathrm{m}})/(\varepsilon-1)}{q_{\mathrm{w}}((x\varepsilon/\varepsilon-1)((x/c)^{\varepsilon})^{-1/\varepsilon} - c)} = \frac{(cq_{\mathrm{m}})/(\varepsilon-1)}{q_{\mathrm{w}}((c\varepsilon/\varepsilon-1) - c)} = \frac{(cq_{\mathrm{m}})/(\varepsilon-1)}{q_{\mathrm{w}}(c/\varepsilon-1)} = \frac{q_{\mathrm{m}}}{q_{\mathrm{w}}}$$

[21] Philipson et al. (2006) discuss optimal technology assessment in the presence of altruism that motivates public health care delivery, in general, and R&D into third-world diseases, in particular.

[22] The discrimination between public and private funding may be mitigated by private expenditures towards the licensing of publicly funded discoveries.

[23] We thank Richard Frank for suggesting this measurement strategy.

That is, the share of profits in potential social surplus is equal to the ratio of the monopolist output to the competitive output. In terms of the exogenous parameters, this simplifies to:

$$\frac{\Pi(q_{\mathrm{m}})}{G(q_{\mathrm{w}})-q_{\mathrm{w}}c}=\left(\frac{\varepsilon-1}{\varepsilon}\right)^{\varepsilon}$$

Using the above expressions, it is straightforward to derive the CB ratio as well:

$$\mathrm{CB}=\frac{p(q_{\mathrm{m}})q_{\mathrm{m}}}{G(q_{\mathrm{m}})}=\frac{\varepsilon-1}{\varepsilon}$$

References

Arrow, K., 1961. Economic welfare and the allocation of research for invention. In: Nelson, E.R. (Ed.), The Rate and Direction of Inventive Activity: Economic and Social Factors. Princeton University Press.

Bethan, G., Harris, A., Mitchell, A., 2001. Cost-effectiveness analysis and the consistency of decision making: evidence from pharmaceutical reimbursement in Australia (1991–1996). Pharmacoeconomics 11, 1103–1109.

Caves, R., Whinston, M., Hurwitz, M., 1991. Patent Expiration, Entry, and Competition in the U.S. Pharmaceutical Industry. Brookings Paper on Microeconomic Activity, Microeconomics, pp. 1–66.

Cutler, D.M., 2004. Your Money or Your Life: Strong Medicine for America's Health Care System. Oxford University Press, New York, NY.

Cutler, D.M., McClellan, M., 2001. Is technological change in medicine worth it? Health Affairs 20, 11–29.

Drummond, M., et al., 1999. Current trends in the use of pharmacoeconomics and outcomes research in Europe. Value in Health 2, 323–332.

Drummond, M., et al., 1993. Economic evaluation of pharmaceuticals: a European perspective. Pharmacoeconomics 4, 173–186.

Drummond, M.F., O'Brien, B., Stoddart, G.L., Torrance, G.W., 1997. Methods for the Economic Evaluation of Health Care Programmes. Oxford University Press.

Garber, A.M., 1999. Advances in Cost-Effectiveness Analysis of Health Interventions. NBER Working Paper 7198.

Garber, A.M., Phelps, C.E., 1997. Economic foundations of cost-effectiveness analysis. Journal of Health Economics 16, 1–32.

Garber, A.M., Jones, C.I., Romer, P.M., 2006. Insurance and incentives for medical innovation. BE Press Forum for Health Economics and Policy, Forum: Biomedical Research and the Economy, Article 4.

Gold, M.R., Siegel, J.E., Russell, L.B., Weinstein, M.C., 1996. Cost-Effectiveness in Health and Medicine. Oxford University Press.

Hall, B.H., 1996. In: Smith, B., Barfield, C. (Eds.), The Private and Social Returns to Research and Development, Technology, R&D and the Economy. Brookings Institution/American Enterprise Institute.

Johannesson, M., Weinstein, M.C., 1993. On the decision rules of cost-effectiveness analysis. Journal of Health Economics 12, 459–467.

Lakdawalla, D., Sood, N., 2005. Insurance and Innovation in Health Care Markets. NBER Working Paper.

Levin, R.C., Klevorick, A.K., Nelson, R.R., Winter, S.G., Gilbert, R., Griliches, Z., 1987. Appropriating the returns from industrial research and development. Brookings Papers on Economic Activity 3, 783–831.

Manning, R.L., 1994. Changing rules in the Tort Law and the market for childhood vaccines. Journal of Law and Economics 37, 247–275.

Mansfield, E., 1985. How rapidly does new technology leak out? The Journal of Industrial Economics 34, 217–223.

Mansfield, E., Rapoport, J., Romeo, A., Wagner, S., Beardsley, G., 1977. Social and private rates of return from industrial innovation. The Quarterly Journal of Economics 91, 221–240.

Meltzer, D., 1997. Accounting for future costs in medical cost-effectiveness analysis. Journal of Health Economics 16, 33–64.

Neumann, P.J., Sandberg, E.A., Bell, C.M., Stone, P.W., Chapman, R.H., 2000. Are pharmaceuticals cost-effective? A review of the evidence. Health Affairs 19, 92–109.

Newhouse, J., 1992. Medical care costs: how much welfare loss? Journal of Economic Perspectives 6 (3), 3–21.

Nordhaus, W.D., 2004. Schumpeterian Profits in the American Economy: Theory and Measurement. NBER Working Paper.

Phelps, C.E., Parente, S.T., 1990. Priority setting in medical technology and medical practice assessment. Medical Care 28, 03–723.

Philipson, T., 2000. In: Newhouse, J., Culyer, A. (Eds.), Economic Epidemiology and Infectious Disease, Chapter in Handbook of Health Economics. Elsevier B.V., North Holland.

Philipson, T., Jena, A.B., 2005. Who Benefits from Medical Technologies? Estimates of Consumer and Producer Surpluses for HIV/AIDS Drugs. Forum for Health Economics and Policy, Forum: Biomedical Research and the Economy, Article 3. Berkeley Electronic Press (Also NBER Working Paper #11810).

Philipson, T., Jena, A.B., Mechoulan, S., 2006. Intellectual Property & External Consumption Effects: Generalizations from Pharmaceutical Markets. NBER Working Paper #11930.

Raftery, J., 2001. NICE: faster access to modern treatments? Analysis of guidance on new health technologies. British Medical Journal 323, 1300–1303.

Tirole, J., 1988. The Theory of Industrial Organization. The MIT Press.

Weinstein, M.C., Manning Jr., W.G., 1997. Theoretical issues in cost-effectiveness analysis. Journal of Health Economics 16, 121–128.

Weinstein, M.C., Stason, W.B., 1977. Foundations of cost-effectiveness analysis for health and medical practices. New England Journal of Medicine 296, 716–721.