Surplus Appropriation from R&D and Health Care Technology Assessment Procedures

[PRELIMINARY AND INCOMPLETE-PLEASE DO NOT QUOTE]

by

Tomas J. Philipson[1]

and

Anupam B. Jena

The University of Chicago

Date: November 7, 2005

Abstract:

Given the rapid growth in health care spending that is often attributed to technological change, many private and public institutions are grappling with how to best assess and adopt new health care technologies. We argue that popular assessment criteria going under the rubric of "cost-effectiveness" often concern maximizing *consumer surplus*, which many times is consistent with maximizing static efficiency after an innovation has been developed. Dynamic efficiency, however, concerns aligning the social costs and benefits of R&D and is therefore determined by how much of the social surplus from the new technology is appropriated as *producer surplus*. We estimate that for the HIV/AIDS therapies that entered the market from the late 1980's onwards, producers appropriated only 5% of the social surplus arising from these new technologies. We show how to translate standard findings of cost-effectiveness to estimates of innovator appropriation for standard studies of over 200 drugs, and find that these studies implicitly support a low degree of appropriation as well. Despite the high annual costs of drugs to patients, the low share of social surplus going to innovators raises concerns about advocating cost-effectiveness criteria that would further reduce appropriation by innovators, and hence further reduce dynamic efficiency by unduly sacrificing future patients' health for current ones.

---

[1] Corresponding author: t-philipson@uchicago.edu. We are thankful for comments from seminar participants at The University of Chicago and the NIH-Director conference *Biomedical Innovation and the Economy* held in Bethesda, June 1, 2005. Philipson is thankful for financial support from The Milken Institute, Santa Monica, CA. Jena received fellowship support from the NIH through the University of Chicago Medical Scientist Training Program. We thank Gary Becker and Casey Mulligan for comments, Lisa Lee and Michael Campsmith (CDC) for providing data, and Jennifer Kates (Kaiser Family Foundation) and Ruigang Song (CDC) for helpful advice.

Section 1: Introduction

Technological change is often argued to be a central force behind the growth in healthcare spending.[2] Given this rapid growth, criteria used by private and public institutions to value the increase in healthcare spending therefore requires a methodology to measure the value of new healthcare technologies brought about by R&D investments. There is a long-standing and vast health economics literature that attempts to assess the value of new technologies by use of so called cost-effectiveness, cost-utility, or cost-benefit analysis, hereafter referred to collectively as CE analysis.[3] This type of CE analysis has been argued to be central in managing new technologies, their adoptions, and their impact on long term healthcare spending.

Although not explicitly stated as such, we argue that CE criteria are implicitly concerned with estimating the observed market level of *consumer surplus* associated with a given technology. In particular, many technology assessments attempt to quantify the health impacts of new technologies for patients or health plans by comparing patient benefits with spending at observed market prices. 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 healthcare spending. Though such estimates may not fully capture unobservable aspects of consumer surplus that would be incorporated in traditional demand estimates, the central theme of standard CE assessment performed in practice seems, nevertheless, aimed at measuring consumer surplus or net benefits. In common CE practice, technologies are deemed more valuable the larger is the patient or health plan benefits above what is spent on them.

However, when new technologies are brought to life from costly R&D, consumer surplus is a very poor guide to inducing optimal (second-best) R&D policy. 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. Put differently, even though measured levels of CE would be higher without the patent system, since patients or health plans would spend less to get the same technology, dynamic efficiency would clearly be lowered. An illustrative case of the dangers of CE criteria may be vaccines, which many times have been estimated to be extremely cost-effective but for that and other reasons lack any appreciable R&D investments.[4] In fact, we argue that many times both dynamic efficiency and patient health is maximized when CE is minimized.

As the ability of innovators to appropriate the surplus of their innovations is central to dynamic efficiency, we investigate the degree to which this takes place for a major breakthrough in medicine—the new drugs to treat HIV/AIDS that came on the market in the late 1980's. HIV/AIDS is an important case to consider in and of itself, partly because it is perhaps the major

[2] See e.g. Newhouse (1992).
[3] The literature is vast, but for examples, see Weinstein and Stason (1977), Johanneson and Weinstein (1993), Gold et al. (1996), Meltzer (1997), Drummond et al. (1997), Garber and Phelps (1997), Garber (2000), Cutler and McClellan (2001), and Cutler (2005).
[4] A major concern here has, of course, been product liability issues. See e.g. Manning (1993).

disease targeted by public sector R&D in the US.[5] For the new HIV drugs that came about during this period, our major finding is that innovators captured only 5% of the social surplus arising from these new technologies. More precisely, consumer and producer surplus from these drugs amounted to roughly $1.33 trillion and $63 billion, respectively. Our main point is that if the new HIV/AIDS therapies are representative of other technologies, the lack of appropriation of social surplus by innovators raises strong concerns about adherence to CE analysis. Despite the high prices of many therapies such as the new HIV drugs, patients and health plans are getting too good a deal in the short run which, of course, hurts them in the long run by insufficient R&D.

This low share of appropriation by innovators can be understood by some simple back-of-the-envelope calculations. For the size of the consumer surplus, consider that there have been about 1.5 million US citizens infected with HIV since the start of the epidemic, some of whom died before drug therapy became available and some of whom lived until or contracted HIV after the second wave of HIV/AIDS drugs entered the market in the mid 1990's. Averaging across all such cohorts, the gain in life-expectancy has been at least 5 years. With a low value of a life-year of $100,000, the added survival has been worth more than $500 thousand per individual and $750 billion in aggregate. This figure, of course, does not include the benefit to those individuals who become infected with HIV in later years but can benefit from drugs introduced to date—doing so, while assuming current incidence rates persist in the future, raises the total consumer value of these drugs above $1 trillion. For the size of the producer surplus, consider that sales of HIV/AIDS drugs have grown from $1 billion to $4 billion annually since the breakthrough drugs came on the market in 1996. From these revenues, one can compute a present value of sales of $74 billion assuming that drugs sell at current levels in the future. We can then subtract the variable costs of production, which are approximated to be 15% of revenue based on estimates of markups stemming from differences in drug prices pre- and post patent expiration.

Moreover, we show that the estimated small share of social surplus captured by innovators turns out to be consistent with alternative, theory-based methods of calculating this share. We argue that, generally, price reductions upon patent expirations may be used to estimate patent-protected markups, and hence the elasticity of demand for patent-protected drugs. Existing estimates suggest that price reductions upon generic entry are on the magnitude of 85% percent, implying a demand elasticity of about 1.17.[6] For a constant elasticity demand curve, it can be shown that this implies an innovator share of potential social surplus of about 10%, which is in the same order of magnitude as our *directly estimated* share of 5%.

Given the low estimated share of social surplus appropriated by producers of HIV/AIDS drugs, this raises the question of whether producers of similarly CE technologies appropriate comparable amounts of social surplus. And if so, can these results be generalized to obtain appropriation estimates that vary with a technology's observed level of cost-effectiveness? Relating our main finding to the traditional CE literature, we demonstrate why and how the CE results of over 200 studies on drugs can be implicitly viewed as identifying the degree of

[5] Public R&D on HIV/AIDS was roughly $2 billion in 2000. Health, in general, is among the three leading industries into which the government allocates its R&D, the other two being defense and aero-space. The National Institutes of Health (NIH) is responsible for allocating the vast majority of the public R&D dollar—in 1999, NIH funding accounted for nearly 81% of public spending on health R&D. Of the $13.9 billion that the NIH spent on research in 1999, nearly $1.8 billion (13%) was spent on HIV/AIDS (Health, United States, 2002).

[6] See e.g. Caves et al (1991). In general, when marginal costs are zero, monopolists will price at unit elasticity. Interestingly, the marginal costs of drug production are thought to be quite low, consistent with our elasticity estimates slightly above one.

innovator appropriation. We discuss the conditions under which there is a negative relationship between cost-effectiveness and innovator appropriation—that is, the often estimated CE of a technology may be used to identify the share of social surplus appropriated by producers of that technology. This is because the CE of a given technology *reveals* information about the cost or demand parameters generating the observed CE. This implies that the existing and vast CE literature can be evaluated in terms of its implications for the limited degree of innovator appropriation. We find that 25% of the interventions considered have estimated appropriations of less than 7%, while 75% have appropriations less than a fourth. If the estimated distribution of producer shares generalizes to the distribution across all health interventions, our empirical finding for producers of HIV/AIDS drugs suggests their appropriation of social surplus is at the twentieth percentile. Importantly, our findings relate to an existing literature on the general inability of innovators to capture the social value of their inventions.[7]

If this lack of appropriation by innovators generalizes to other health care technologies, it naturally raises questions about its causes. It seems natural to suggest that this is ultimately due to a lack of market power of patent holders through weak patents or close substitutes being available. However, we show that in common monopoly models, market power *reduces*, rather than raises, appropriation. In other words, as demand becomes less sensitive to price, the monopolist captures less of the social surplus; the non-appropriated consumer surplus rises faster than profits as the elasticity falls. Even though one's first reaction may be that low appropriation by innovators is due to highly elastic demand induced by weak patents or other forces, our analysis thus shows that for nearly-inelastic demand, appropriation is still small. In fact, a producer share of social surplus of the level estimated for HIV/AIDS, namely 5%, is consistent with monopoly pricing under a constant-elasticity demand curve that is almost as inelastic as it can be, $\varepsilon = 1.06$. The point is that despite the high prices of HIV/AIDS drugs, presumably associated with a low elasticity of demand for these life-saving technologies, producers capture a small share of social surplus.

The paper may be briefly outlined as follows. Section 2 discusses the effect of supply and demand on observed levels of CE, followed by the difference between dynamic healthcare technology evaluation and static analyses implicit in CE criteria. Section 3 presents estimates of the share of social surplus appropriated by producers of HIV/AIDS drugs. Section 4 discusses how to compare our finding to traditional empirical results in the CE literature and how that literature can be interpreted as also supporting a small share of appropriation by innovators. Section 5 discusses the causes of modest appropriation by innovators by considering the non-trivial relationship between market power and appropriation. Lastly, section 6 concludes.

Section 2: Technology Assessment and Dynamic versus Static Efficiency

In order to discuss how CE analysis relates to static and dynamic efficiency, for a given output level q denote the ex-post social surplus of a new technology by w(q). This social surplus can be divided into a consumer surplus, z(q), and producer surplus (variable profits), $\pi(q)$, as in:

$$w(q) = z(q) + \pi(q) \qquad (1)$$

For example, a commonly analyzed case is when price-discrimination is infeasible, in which case a given output level q induces both profits and consumer surplus according to

[7] See e.g., Mansfield et al. (1977), Mansfield (1985), Levin et al. (1987), Hall (1996), and Nordhaus (2004).

$$\pi(q) = p(q) \cdot q - c(q) \qquad (2)$$

$$z(q) = \int_0^q [p(y) - p(q)]dy = g(q) - p(q)q \qquad (3)$$

where p(q) is the inverse demand function, c(q) is the variable cost function which excludes the fixed cost of R&D, and g(q) is the gross consumer benefit.

2.1 Cost-Effectiveness Criteria

In this standard framework, we argue that typical CE technology evaluation has implicitly centered on consumer surplus, by focusing on how much patients benefit beyond what is spent on the technology after it has been developed. Despite the many forms of such criteria developed to date, their basic goal seems to be to determine whether increased healthcare spending on new technologies is justified by "societal", "health plan", or "patient" benefits in terms of improved health. Absent from the discussion has been the effect of such criteria on the behavior of innovators who make the technologies available in the first place. Although static efficiency is often enhanced with increases in CE, as it implicitly concerns consumer surplus, these criteria are less understood in terms of how they relate to dynamic efficiency when the observed level of CE is the result of rational behavior by market participants.

Common measures of CE ratios relate the (here monetized) patient benefits to observed spending levels. In the traditional framework, this can be expressed by:

$$z_R = \frac{g}{p \cdot q} = 1 + \frac{z}{pq} \qquad (4)$$

This measure (z_R) expresses consumer benefits as a *ratio* to spending, similar to the standard consumer surplus measure (z) that expresses it as a *difference* between the two. Ratios are often estimated through spending per quality- or disability-adjusted life years or through monetized versions of health benefits, in which the value of life is compared to observed spending levels. These attempts, however, are implicitly related to the size of consumer surplus, since they compare consumer benefits to observed spending levels.[8] In particular, static technology assessments in healthcare commonly rely upon the use of "cost-benefit", "cost-utility", or "cost-effectiveness" criteria to determine under what circumstances the value (whose units depend on the measure) of a given technology exceeds what is spent on it. Although it is true that CE analysis concerns the ratio of gross benefit and spending, while consumer surplus concerns their difference, both change in the same direction with unilateral changes in costs and benefits.

Regarding the estimated CE magnitudes, many empirical studies estimate and document z_R ratios above unity for employed technologies (see e.g. references in Introduction). Yet, it would be extremely surprising if correctly measured z_R ratios were found to be below unity, at

[8] 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.

least in a standard market economy. 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, such as e.g. plastic surgery. A new plastic surgery technology would have a z_R ratio above 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. Although this expected and basic outcome has to be qualified by the presence of private or public insurance, it is supported by a large existing and growing empirical health economics literature on the cost-effectiveness of recent innovations.[9]

Little is understood about how CE criteria operate in a market context with traditional supply and demand. Given the obvious dependence of z_R on prices and quantities, it is clear that the observed CE should respond to supply and demand parameters. We can write z_R more generally to illustrate its dependence on cost and demand parameters, θ and η. Specifically,

$$z_R = \frac{g(q(\eta,\theta),\eta)}{p(q(\eta,\theta),\eta)\cdot q(\eta,\theta)} \qquad (5)$$

Note that cost parameters affect gross benefits and spending only indirectly through their effect on equilibrium output and price. Demand parameters, on the other hand, affect gross benefits and spending indirectly, through their effect on equilibrium output and price, but also directly, through their effect on the demand curve itself.

The above expression can be used to determine how a technology's observed CE ratio varies with cost and demand factors, i.e. the elasticity of z_R with respect to θ and η. As with any ratio, the elasticity is simply the elasticity of the numerator minus the elasticity of the denominator. In the Appendix, we derive the following general expressions for the elasticity of z_R with respect to cost and demand parameters, where ε_b^a is the elasticity of *a* with respect to *b*:

$$\begin{aligned}
\varepsilon_\theta^{z_R} &= \left[\frac{1}{z_R}\cdot\varepsilon_\theta^q\right] - \left[\varepsilon_\theta^q\cdot\left(1-\frac{1}{\varepsilon_p^q}\right)\right] \\
\varepsilon_\eta^{z_R} &= \left[\frac{1}{z_R}\cdot\varepsilon_\eta^q + \varepsilon_\eta^g\right] - \left[\varepsilon_\eta^q\cdot(1-\frac{1}{\varepsilon_p^q}) + \varepsilon_\eta^p\right]
\end{aligned} \qquad (6)$$

As this illustrates, changes in cost and demand parameters are not clearly and monotonically related to changes in the observed level of cost-effectiveness, as measured by z_R.

These effects may be illustrated for some commonly analyzed supply and demand schedules. First, consider the case of constant returns to scale and linear demand. If the inverse demand is given by p(q) = a – bq and costs are given by c(q) = cq, one can show that under monopoly pricing, the CE ratio satisfies $z_R = (3a + c)/(2a + 2c)$. An outward shift in demand (increasing a) leads to increases in the observed CE ratio. On the other hand, increases in costs lead to decreases in the observed level cost-effectiveness. However, as opposed to direct measures of CE that would have a unit elasticity with respect to costs, $\varepsilon_c^{z_R} = -2ac/[(a+c)(3a+c)]$ depends on both supply and demand parameters.

[9] See, for example, Cutler (2004).

Second, consider the case in which demand is of the constant elasticity form as in $p(q) = x/q^{1/\varepsilon}$, where $\varepsilon > 0$ is the elasticity of demand with respect to price and x is a scale factor that shifts demand outward. In this case, the CE ratio is $z_R = \varepsilon/(\varepsilon-1)$. This has the striking implication that a technology's observed level of cost-effectiveness can be independent of cost parameters, which occurs when cost affects spending and gross benefits proportionately.[10]

2.3 R&D and Dynamic Efficiency

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. Let technological change be characterized by x(r), an increasing, differentiable, and strictly concave function representing the probability of discovery for a given level of R&D undertaken, r. The optimal level of R&D that maximizes expected payoffs for any hypothetical ex-post prize, k, is denoted r(k) and is defined by:

$$r(k) = \arg\max_{r}\left[x(r)k - r\right] \qquad (7)$$

Our assumptions about x(r) imply that r(k) is an increasing function so that R&D rises with the ex-post reward. In particular, $r(\pi)$ represents the R&D undertaken when those investing in R&D maximize expected profits. If profits drive R&D investments, the expected social surplus is

$$E(z,\pi) = x[r(\pi)]\cdot w - r(\pi) \qquad (8)$$

where $w = z + \pi$ is the social surplus ex-post. This expression directly highlights the well-known implication that dynamic efficiency only 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 (see e.g. Arrow (1961) and Tirole (1988)). In other words, the key factor driving dynamic inefficiency is that profits (π) are less than social surplus (w). 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. Indeed, in this setting, the dynamically efficient R&D investment is r(w), which is obtained when the entire social surplus is appropriated as profits.

More generally, for any technology and preferences, the observed profits associated with a given level of social surplus can be written $\pi(w)$. The main issue, then, is that $\pi(w) < w$. 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 social surplus. Specifically, $\pi(w) = w/2$ in the case of linear demand and $\pi(w) = w\cdot[(\varepsilon-1)/\varepsilon)]^{\varepsilon}$ under constant elasticity of demand.[11,12] In general, then, if the total social surplus associated

[10] In this particular case, it can be shown that elasticities of both spending and gross benefits with respect to cost (c) are equal to $1-\varepsilon$.

[11] The social surplus implicit in these results is the *potential* social surplus available to innovators, i.e. the social surplus that obtains when price is set at its competitive level. This differs from the *observed* social surplus available to the monopolist, which obtains when price and quantity are determined by the monopolist.

[12] 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 and Mechoulan (2003) who discuss R&D under altruism in health care.

with a technology is w, the size of the under-investment in R&D is $r(w)-r(\pi) = r(w)-r(w-z)$, which, since r(.) is an increasing function, rises with the consumer surplus focused on by CE criteria. The fact that dynamic efficiency is driven by the appropriation of social surplus to innovators implies that substituting producer surplus for consumer surplus often raises dynamic welfare. This is analogous to the argument that patents hurt static efficiency but raise dynamic efficiency by engaging in similar substitution.

The important implication of this is that the CE associated with the ex-post market for a technology is not clearly and monotonically related to measures of static or dynamic efficiency. Indeed, in a private market with perfect price discrimination, dynamically efficient R&D occurs because the innovator captures the entire social surplus. Therefore, 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., $z_R = 1$) and that empirical studies citing more cost-effective technologies are, in fact, documenting a dynamic inefficiency! Indeed, as discussed, the underinvestment in R&D from its socially optimal level, r(w) - r(w-z), rises with how "cost-effective" a technology is assessed to be according to traditional CE analysis. In this case, the dynamically efficient minimization of CE is a direct implication of the classic problem of non-appropriation by innovators leading to under-investment in R&D. Importantly, note that minimization of CE in this context still maximizes patient health (as full demand for the health care product obtains) though not consumer surplus.

In a market with extensive public demand, as in many European markets, it is important to note that if CE procedures are used, they may affect the demand function itself, in which case the optimal price, and hence the *observed* CE, will be endogenous to the type of CE analysis used. More precisely, consider a technology that is publicly financed for q patients when the CE ratio is below a certain threshold T. This defines a corresponding price, p(T), at which the CE ratio equals the given threshold, or $g(q)/[p(T)q] = T$. Implicitly, such a public adoption criteria *induces* a demand curve of zero elasticity (in which price is profitably raised) below p(T) and infinite elasticity (in which price is not profitably raised) above p(T). Because the observed CE is endogenous to the adoption criteria (here equaling T regardless of what T is), public adoption criteria based on CE will induce CE to be the highest feasible level. CE adoption criteria will therefore not, as commonly argued, rule in or rule out good versus bad technologies. Instead, they will simply determine the price at which good or bad technologies will be bought as long as markups are positive.

Section 3: Surplus Appropriation for the New HIV/AIDS Drugs

The previous discussion highlighted the importance of surplus appropriation by innovators, and hence low levels of CE, to dynamic efficiency. As the ability of innovators to appropriate the surplus of their innovations is central to dynamic efficiency, we briefly summarize the results of Philipson and Jena (2005), who investigate the degree to which this takes place for the new drugs to treat HIV/AIDS that entered the market from the late 1980s onwards. This analysis will then be used to illustrate how levels of innovator appropriation may be inferred from existing CE estimates in the literature.

3.1 Estimates of Gross Consumer Benefits

The value of life induced by new drug therapies is the value of increased survival for all affected individuals, relative to a benchmark in which no (or worse) therapies exist. In a related work, we develop a methodology to value the increases in survival attributable to the now

standard treatments for HIV/AIDS (Philipson and Jena, 2005). The thought experiment behind the analysis is the following. For a hypothetical individual infected in a given year, we examine how that individual's survival compares to a baseline survival in which no drugs are available. We then attach a monetary value to that increased survival and sum across all infected individuals in that cohort. This process is repeated for each set of cases, cohort by cohort, since the start of the epidemic and aggregated up.

The progression from initial HIV infection to death is represented by a survival function S_t, summarizing the transition from HIV to death for individuals infected with HIV in year t. This survival curve is assumed to be raised by the consumption of new drugs, compared to the counterfactual survival curve S_o experienced by those infected in year zero, taken to be 1979. The gross consumer benefit, g, induced by the new drug consumption is calculated by multiplying the size (or incidence) of cohort t, n_t, by the monetary value of increased survival and summing over all calendar years. Formally, the total discounted gross consumer benefit is written as:

$$g = \sum_{t=1980}^{2000} \beta^{t-1980} n_t \cdot g_t \qquad (9)$$

where $g_t = g(S_o, S_t)$, or the monetary value of increasing survival from the baseline survival S_o to the higher future survival faced by cohort t, S_t. The gain in survival, g_t, is calculated using the infra-marginal valuation formula of Becker et al. (2005). The authors provide the value of survival gains for an infra-marginal change in survival from S_o to S_t under a yearly income y_t as in:

$$V[y_t + e_t, S_o] = V[y_t, S_t] \qquad (10)$$

where V is the indirect *lifetime* utility function and e_t is the yearly compensation required to make the hypothetical individual indifferent between the two survival frontiers.[13] The lifetime value for the gain in survival is calculated by summing the yearly compensation (e_t) over time, discounting by the rate of interest and the new survival probability. That is,

$$g_t = \sum_{d=0}^{\infty} \beta^d \cdot S_o \cdot e_t \qquad (11)$$

To empirically implement this calculation of gross consumer benefits, we apply these formulas to published levels of HIV incidence and estimated changes in survival induced by HIV/AIDS drugs. Figure 1 graphs the HIV incidence figures on which our estimates are based, as well as AIDS incidence and mortality from the CDC's HIV/AIDS Annual Surveillance Reports. Since 1990, the CDC has estimated incidence to be stable at roughly 40,000 individuals infected annually.[14] There have been over 900,000 diagnosed AIDS cases to date, with over 500,000 AIDS related deaths. There are slightly over 400,000 individuals currently living with AIDS and roughly 650,000 individuals living with HIV only.

[13] Note that the use of yearly, as opposed to lifetime, income (y) and compensation (e) in expression 8 follows from the assumption that the discount rate on instantaneous utility equals the market rate of discount, i.e. the interest rate. For more details see Becker, et al. (2005).

[14] This estimate is consistent with data on rates of transmission among hetero- and homo-sexuals and prevalence of both groups in the US (Communication with CDC).

Figure 1: Estimates of HIV Incidence, AIDS Incidence, & Deaths from AIDS

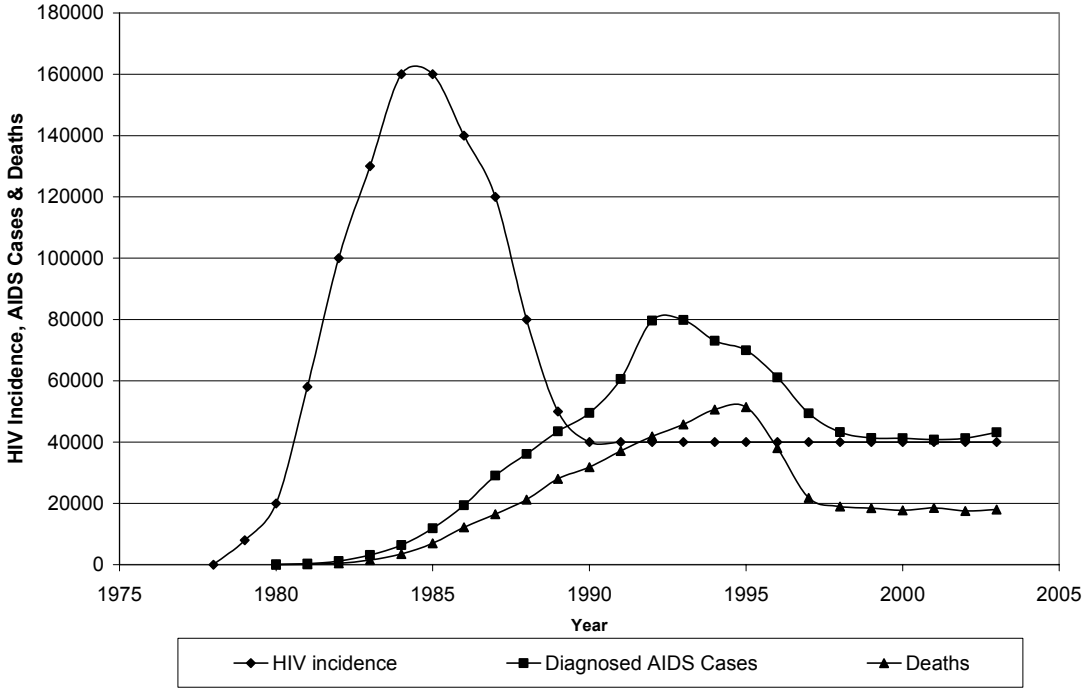
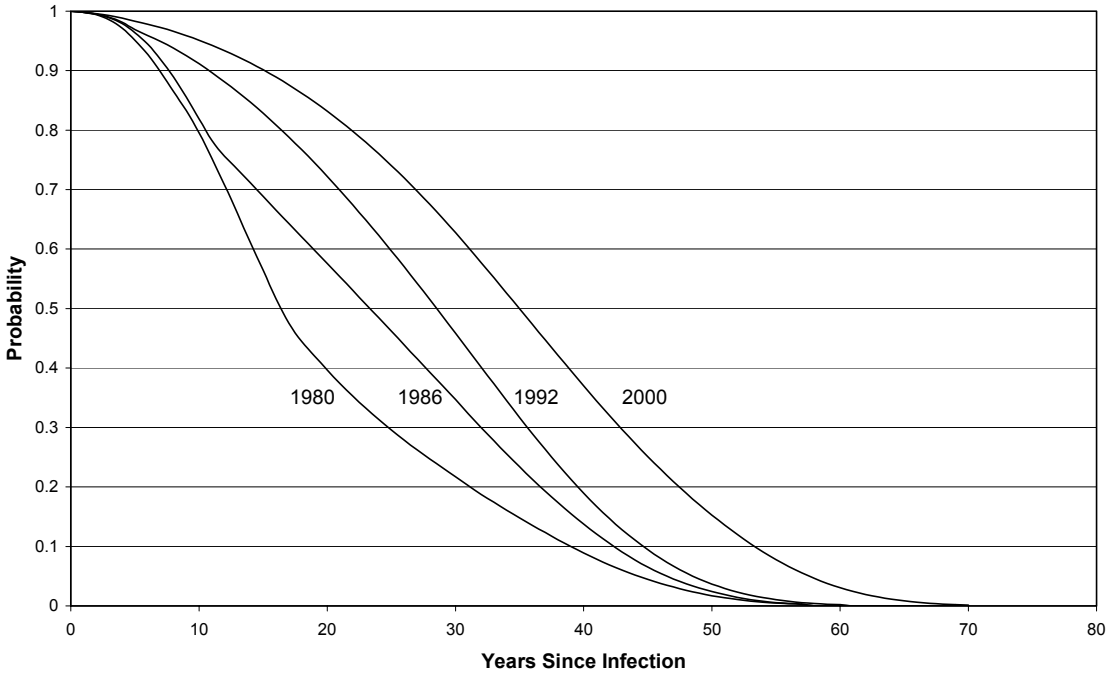


Figure 1 displays the large declines in both AIDS deaths and diagnoses following the introduction of the breakthrough drugs (Protease Inhibitors and Non-Nucleoside Reverse Transcriptase Inhibitors) in 1995 and 1996.[15] The dual decline is consistent with HIV/AIDS drugs slowing the progression from HIV to AIDS and from AIDS to death.

Using various data sources and methods, we then estimate the improvements in HIV survival for each year from 1979 to 2000 (see Philipson and Jena (2005) for a detailed discussion). Figure 2 presents these changes in survival for various years of infection.

[15] The apparent increase in AIDS diagnoses around 1993 is, in part, due to a change in the definition of AIDS. Prior to 1993, a diagnosis of AIDS was based on the clinical finding of an opportunistic infection. After 1993, the diagnosis was expanded to include individuals with CD4 counts below 200 per cubic millimeter.

Figure 2: Survival from HIV by Year of Infection



The improvements in survival depicted in the above graph are due to increases in both the time to onset of AIDS (after being infected with HIV) and the period of time alive after a diagnosis of AIDS. The life-expectancy of the average individual infected with HIV increased roughly 15 years since the start of the epidemic, from 19 to 34 years.[16]

The value of improved survival for a given cohort depicted in Figure 2 is computed by multiplying that cohort's incidence of HIV by the value of increased survival experienced by a single individual in that cohort. The income used in the calculations is GDP per capita (in year 2000 dollars). All figures are discounted back to 1980.

Table 1: Value of Gains in Survival for HIV Infected Individuals

		Value of Survival Gains ($)	
Year of HIV Infection	HIV Incidence	Individual ($)	Aggregate ($ Billion)
1980	20,000	17,655	0.35
1981	58,000	39,361	2.28
1982	100,000	60,256	6.03

[16] These figures are consistent with those in the literature (see e.g. Lichtenberg (2005) and Philipson and Jena (2005)). To further examine the robustness of these estimates, we *predict* the number of individuals alive with HIV/AIDS in 2003, based on the annual reported incidence of HIV and our estimated survival curves. We then compare this to the *reported* number of individuals living with HIV/AIDS in 2003. The predicted and reported figures differ by only 12,000 people (out of nearly 1 million alive with HIV/AIDS).

1983	130,000	84,941	11.04
1984	160,000	116,156	18.59
1985	160,000	146,874	23.50
1986	140,000	178,968	25.06
1987	120,000	214,389	25.73
1988	80,000	250,284	20.02
1989	50,000	287,924	14.40
1990	40,000	322,311	12.89
1991	40,000	339,957	13.60
1992	40,000	383,328	15.33
1993	40,000	432,908	17.32
1994	40,000	567,422	22.70
1995	40,000	613,839	24.55
1996	40,000	696,951	27.88
1997	40,000	718,603	28.74
1998	40,000	730,179	29.21
1999	40,000	738,839	29.55
2000	40,000	740,515	29.62
Total Discounted Value			398

All figures are discounted to 1980 and are in year 2000 dollars.

The above results suggest that the aggregate value of improved survival experienced by all individuals infected with HIV to date has been nearly $400 billion. This, of course, ignores the value of increasing survival for all individuals who have not contracted HIV yet. To add this component, we forecast the value to future cohorts of HIV infected individuals by assuming that all cohorts experience the same aggregate gain in survival g_t as the last cohort, 2000. Hence, we assume

$$g_{2000+t} = g_{2000} \forall t > 0 \qquad (12)$$

Assuming that the future incidence of HIV is equivalent to the last period, we calculate the discounted sum of future gains for individuals infected with HIV in the future. We then add this amount to the value to date shown above, namely $398 billion. This leads to an aggregate value of increased survival for all past and future cohorts of nearly $1.4 trillion.

3.2 Producer Surplus

The overall producer surplus obtained from R&D is determined by the present value of producer surplus to firms producing HIV/AIDS drugs:

$$\pi = \sum_{t=1980}^{\infty} \beta^{t-1980} \cdot \pi_t \qquad (13)$$

An upper bound of producer surplus is the aggregate sales for HIV/AIDS drugs. An alternative lower estimate accounts for variable costs of production. We apply existing estimates of markups for brand-name drugs (as estimated from patent expirations) to estimate these costs.

Figure 3 presents estimates of national spending on HIV/AIDS drugs broken down by public and private payers. The estimates for total spending are from IMS Health and are reported

in Lichtenberg (2005). Public spending is approximated by the sum of Medicaid and ADAP expenditures. The Medicaid estimates include both federal and state contributions and are calculated from the Medicaid State Drug Utilization Data using National Drug Codes (NDC) for all antiretrovirals introduced since 1987.[17]

Figure 3: National Spending on HIV/AIDS Drugs

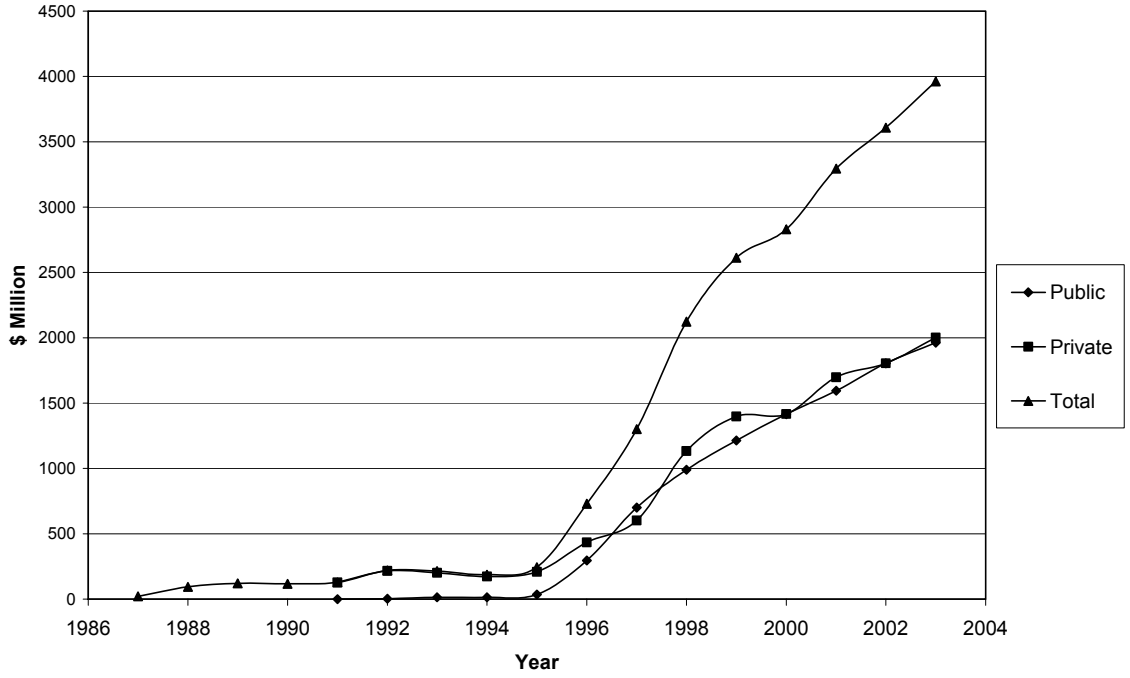


Figure 3 demonstrates the large increase in spending on HIV/AIDS drugs in the past ten years. Since 1995, spending has increased from $250 million to almost $4 billion, largely due to increased spending on the new drugs represented by protease inhibitors and nucleoside reverse transcriptase inhibitors. Figure 3 also depicts the large share of total spending on HIV/AIDS drugs that comes from public sources, nearly 50% from 1997 onwards. In order to estimate the aggregate lifetime profits from HIV/AIDS drugs, we assume that future sales are equal to last-period sales, in this case national sales of HIV/AIDS drugs in 2000. We use estimates from the literature on the prices of generic drugs relative to their branded counterparts to assume variable costs to be no more than 15% of sales (Caves, et al., 1991). With sales in the future being equivalent to year 2000's patent-protected sales, we estimate lifetime sales to be roughly $74 billion. This suggests a lifetime variable cost of production of $11.1 billion (74 x 0.15) and lifetime profits of $62.9 billion.

Using the above figures, we can decompose the total lifetime value of HIV/AIDS drugs into consumer surplus, producer surplus (profits), and production costs. Recall that we estimated the total value, g, to be nearly $1.4 trillion, discounted to 1980 in year 2000 dollars. This amounts to a lifetime consumer surplus, z, of roughly $1.33 trillion ($1.4 trillion – $74 billion). With a social surplus (total value net of production costs), w, of $1.38 trillion, almost 95% is captured in the form of consumer surplus.

[17] http://www.cms.hhs.gov/medicaid/drugs/drug5.asp

Section 4: Extending the Analysis to Traditional Cost-Effectiveness Studies

Given the low estimated share of social surplus appropriated by producers of HIV/AIDS drugs, this raises the question of whether producers of similarly CE technologies appropriate comparable amounts of social surplus. And if so, can these results be generalized to obtain appropriation estimates that vary with a technology's observed level of cost-effectiveness? We begin this section by discussing conditions under which the often estimated CE of a given technology may, in fact, be used to infer the share of social surplus appropriated by producers of that technology.

Recall from our earlier discussion that the ratio of gross benefit to spending, z_R, can be written as g(q)/[p(q)q]. Similarly, the degree of observed surplus appropriation can be written as $\pi(q)/w(q)$, where w(q) is the *observed* social surplus associated with a level of output q, and $\pi(q)$ is the level of profit induced by that quantity. If m(q)=p(q)/[c(q)/q] is the markup above average costs, it is straightforward to show that for a given level of output, appropriation may be written as a function of the CE level as in[18]:

$$\frac{\pi(q)}{w(q)} = \frac{m(q)-1}{m(q)\cdot z_R - 1} \quad (14)$$

This expression demonstrates that highly cost-effective technologies (those with high z_R) 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 commonly reported CE estimates to infer the degree of appropriation by producers of the relevant technology.

For the case of HIV/AIDS, calculating the appropriation ratio based on our estimates is straightforward. First, recall that we estimated gross benefits to consumers to be nearly $1.4 trillion with spending levels of $74 billion. This implies a CE or z_R ratio of roughly 18. Estimates of the average markup can, in turn, be obtained from information on price reductions after patent expiration, which suggests that average costs are as low as 15% of patented prices. Put together, the average markup and estimated CE of HIV/AIDS drugs imply a producer appropriation of observed social surplus of only 5%, identical to our directly estimated level of appropriation.

With more restrictive cost and demand assumptions, even less information is needed to infer the level of appropriation from CE estimates. Under constant returns to scale and constant elasticity demand, it can be shown that a technology's CE *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 appropriation of two types of surpluses, observed versus potential. The *observed* surplus (presented earlier) is the surplus which obtains at the market quantity. For example, for a monopoly quantity q_m, the appropriation of observed surplus is simply $\pi(q_m)/w(q_m)$. Alternatively, the *potential* surplus is that which would result if the market quantity were determined competitively ($q = q_c$) and hence relates to the total potential surplus available to an innovator. Importantly, the size of profits relative to the *potential* social surplus is most relevant to dynamic policy. For a monopoly quantity q_m, the appropriation of potential surplus is $\pi(q_m)/w(q_c)$. Since there is a deadweight loss associated with monopoly pricing, the potential surplus from an innovation exceeds the observed surplus.

[18] To see this, note that $\pi(q) = p(q)q - c(q)$ and $w(q) = g(q) - c(q)$. Substituting $z_R = g(q)/[p(q)q]$ into the expression for w(q) and simplifying the appropriation share (π/w) yields the above result.

Consequently, estimates of 'surplus' appropriation based on observed surplus will underestimate the deficiency in appropriation by producers of a given technology.

More precisely, consider the common model where variable costs exhibit constant returns, $c(y) = cy$, and there is a constant elasticity 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_c and q_m denote the competitive and monopoly output, respectively, the Appendix shows that the ratio of gross benefit to spending (i.e., z_R) under monopoly pricing satisfies:

$$z_R = \frac{g(q_m)}{p(q_m)q_m} = \frac{\varepsilon}{\varepsilon - 1} = \frac{p(q_m)}{c} \qquad (15)$$

In other words, a technology's CE, as described by the ratio of gross benefit to spending, 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 equals the output expansion due to competition.[19] That is,

$$\frac{\pi(q_m)}{g(q_c) - q_c \cdot c} = \frac{q_m}{q_c} = \left(\frac{1}{z_R}\right)^{\varepsilon} \qquad (16)$$

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 is appropriated.[20] 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.[21] 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%.[22] Finally, there is a direct negative relationship between cost-effectiveness and innovator appropriation.

Under these assumptions, a given estimated CE or z_R ratio implies a specific elasticity of demand, which in turn implies the degree to which a firm appropriates social surplus. In the case of HIV/AIDS, for which z_R is roughly 18, the implied elasticity of demand is around 1.06, which (according to equation 16) implies a producer share of social surplus of 5%.

More generally, the above relationship between CE and surplus appropriation can be used to infer the share of surplus appropriated by those producers whose technologies are examined in existing CE studies. Figure 4, below, graphs the relationship between surplus appropriation, cost-effectiveness, and market power (interpreted as a reduction in the elasticity of demand). As market power decreases, the producer's share of social surplus approaches slightly more than a

[19] 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.

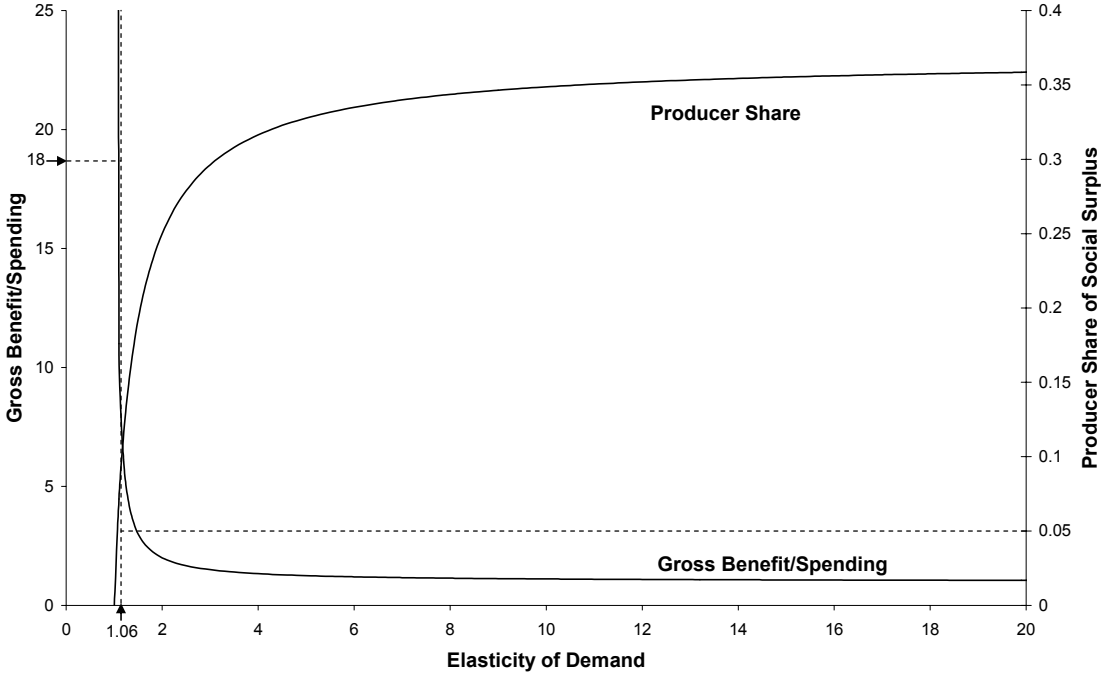
[20] This result may not be unique to this particular demand structure. For a linear demand curve, it is well known that monopoly output is half the competitive output and that a monopolist always appropriates half the surplus, so that the surplus condition above holds.

[21] 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.

[22] Note that while $(\varepsilon-1)/\varepsilon$ approaches unity as elasticity becomes infinite, $[(\varepsilon-1)/\varepsilon]^{\varepsilon}$ does not do the same.

third, while z_R approaches 1. As described earlier, z_R is bounded from below by unity since individuals only purchase goods for which the benefits exceed the costs.

Figure 4: Elasticity of Demand and Producer Shares Implied by CE Estimates



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. For example, consider the technologies to treat HIV/AIDS. With an estimated ratio of gross benefit to spending of roughly 18, this implies an elasticity of demand of 1.06 and a producer share of social surplus of a twentieth.

We exemplify this general identification strategy using estimates of cost-effectiveness from the literature. Neumann et al. (2000) review the cost-effectiveness of more than 200 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 $50,000 and $100,000 per QALY has been a benchmark for the US. 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 4 presents the spending required to obtain an additional QALY for several interventions reviewed by the authors. For example, an intervention with a price of $1,000 that leads to an increase in 0.2 QALYs requires the same spending per QALY as an intervention with a price of $5,000 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 $50,000 and $100,000, we can compute estimates of the ratio of gross benefit to spending per QALY for these interventions, as well as the implied shares of social surplus appropriated by producers.

Table 4: Estimated Producer Share of Social Surplus for Several Cost-Effective Technologies

Intervention	Spending per QALY ($)	z_R (Gross Benefit/Spending)		Producer Share of Surplus	
		$50,000	$100,000	$50,000	$100,000
Captopril Therapy	4,000	12.5	25	0.06	0.03
Hormone Replacement Therapy	12,000	4.2	8.4	0.15	0.09
INH Prophylaxis	18,000	2.8	5.6	0.20	0.12
Hip Fracture Prevention	34,000	1.5	3.0	0.30	0.19
Chemotherapy for Breast Cancer	58,000	.9	1.8	---	0.26

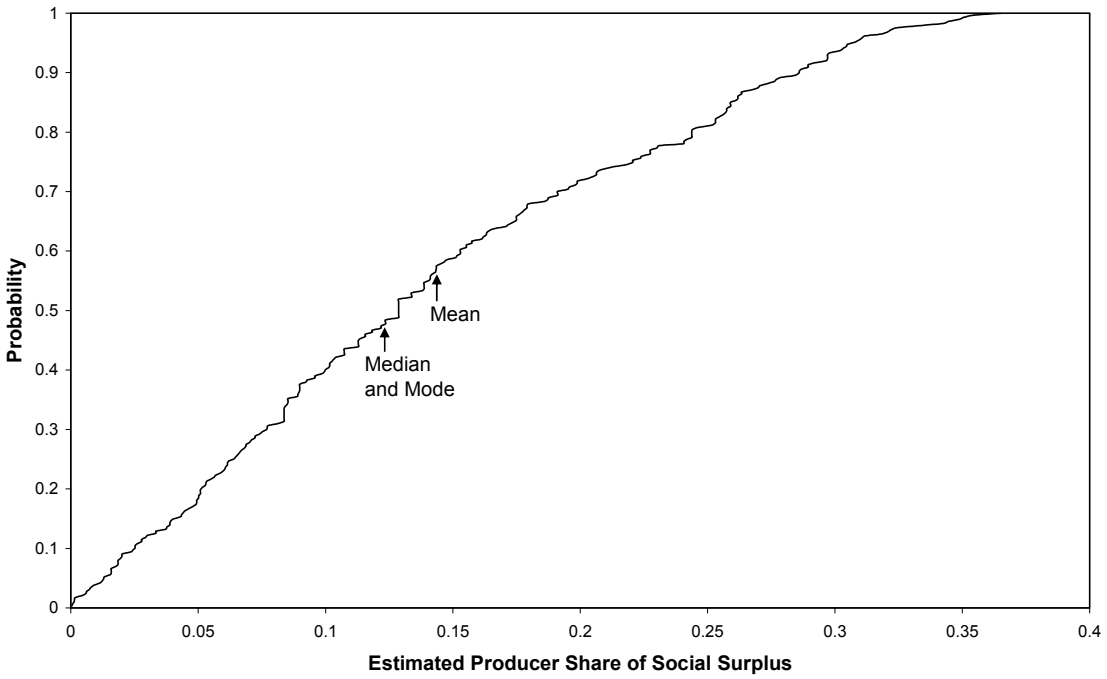
Notes: CE and producer share of surplus are presented for two, separate values of an additional quality adjusted life year. The final intervention has a gross benefit less than cost when gross benefit per QALY equals $50,000.
Description of Interventions—1) Captopril therapy in patients with myocardial infarction, 2) Hormone replacement therapy (HRT), 3) Isoniazid (INH) prophylaxis for tuberculosis, 4) Treatment to prevent hip fracture in patients with osteoporosis, and 5) Chemotherapy for breast cancer. For a more detailed description, see Neumann et al. (2000).

Table 4 demonstrates that, as illustrated by the case of HIV/AIDS, 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 only 3% - 6% of social surplus going to producers.

While Table 4 presents estimates of the producer share of social surplus for only five interventions, cost-effectiveness estimates from a large, *random* sample of interventions could be used to estimate the distribution of producer shares. We use data from over 200 published cost-utility analyses contained in the Harvard Cost-Effectiveness Analysis (CEA) Registry to estimate this distribution.[23] 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 estimate the share of social surplus appropriated by the producer of that technology, as in Table 4 above.[24] Because the studies included in the Registry may not be a random sample of all technologies, however, we can only estimate the distribution of producer shares conditional on inclusion into the Registry. Figure 5 plots the distribution of estimated producer shares for the interventions considered.

[23] The Registry is not limited to only pharmaceutical interventions. More detailed information can be found at: http://www.hsph.harvard.edu/cearegistry/
[24] For these calculations, we assume the gross benefit of an additional QALY to be $100,000. Consequently, we limit our attention to those interventions with published costs of less than $100,000 per QALY gained.

Figure 5: Cumulative Distribution of Estimated Producer Shares



Since the constant elasticity of demand assumption predicts a producer appropriation of social surplus of no more than 37%, all interventions considered in Figure 5 have estimated producer shares less than this amount. The median intervention requires a spending per QALY of roughly $19,000, which corresponds to a producer share of social surplus of nearly 13%.[25] Moreover, 25% of the interventions considered have estimated appropriations of less than 7%, while 75% have appropriations less than a fourth. If the estimated distribution of producer shares generalizes to the distribution across all health interventions (i.e., not only those included in the Registry), our empirical finding for producers of HIV/AIDS drugs suggests their appropriation of social surplus is at the twentieth percentile.

Section 5: Causes of Low Degree of Surplus Appropriation: Market Power?

What forces contribute to the low degree of appropriation by producers as exemplified by the HIV/AIDS drugs and as revealed by the discussed CE estimates? It seems natural to suggest that this is ultimately due to a lack of profits and market power. However, the derivation in the previous section implies that market power *reduces* appropriation, as opposed to raising it, in common models. Even though *profits*, of course, rise as the elasticity of demand falls, many times the *share* of social surplus appropriated by the monopolist *falls*. In other words, as demand becomes less sensitive to price, the monopolist captures less of the social surplus. This occurs because the non-appropriated consumer surplus rises faster than profits as the elasticity falls.

The relationship between the elasticity of demand and the degree of surplus appropriation by firms greatly affects the interpretation of our estimated share of surplus appropriated by innovators, whether estimated directly, as done already, or inferred from other variables, as done

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

shortly. The low degree of surplus appropriation by innovators may potentially be interpreted as prices being held down by: 1) the threat of public regulation if pharmaceutical companies raise prices, or 2) patents that are weakly enforced or too narrowly defined to allow patent-protected monopolies to raise price appropriately. But, our above example demonstrates that even with free pricing and nearly-inelastic demand, the producer share of social surplus may still be very small. In fact, a producer share of social surplus of only 5% is consistent with monopoly pricing under a constant-elasticity demand curve that is almost as inelastic as it can be, $\varepsilon = 1.06$, as an elasticity below unity, of course, leads to infinite profits.

Moreover, given its relationship to the elasticity of demand, the share of social surplus appropriated by innovators can be inferred from demand estimates. This can be compared to our direct estimates for HIV/AIDS. In particular, one can use information on price reductions after patent expiration to estimate patent-protected markups (Caves et al. (1991).[26] These markups identify the elasticity of demand for the patent-protected drugs and thus the share of surplus allocated to the producer. In particular, the derivation above implies that 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% percent, implying a demand-elasticity around 1.17. This elasticity implies a producer share of social surplus of about 10%, which is highly related to our major finding that the share of social surplus appropriated by R&D investors in the area of HIV/AIDS research is, in fact, quite low. This is true even though prices for these drugs are high, presumably due to the inelastic nature of demand.

Section 6: Concluding Remarks and Future Research

We argued that popular technology assessment criteria in healthcare going under the rubric of "cost-effectiveness" are not well understood in terms of how they operate in a market context with traditional supply and demand. We further argued that CE criteria are often implicitly concerned with maximizing the observed level of consumer surplus, which is many times consistent with maximizing static efficiency after an innovation has been developed. Dynamic efficiency, however, aligns the social costs and benefits of R&D and is therefore determined by the how much of the social surplus from a new technology is appropriated by innovators. For the case of HIV/AIDS, our earlier estimates suggested that producers appropriated only 5% of the social surplus arising from new drug therapies. Given the low degree of appropriation by producers of the highly cost-effective HIV/AIDS therapies, we showed how other CE estimates in the literature could be related to the standard framework—our main finding was that these CE estimates implicitly support a low degree of surplus appropriation by producers, comparable to our directly measured estimates for HIV/AIDS. Despite the high annual costs of these drugs to patients, the low share of social surplus going to innovators raises concerns about advocating cost-effectiveness criteria that would further reduce appropriation share, and hence further reduce dynamic efficiency.

In addressing why producer surplus is so small, and why the CE of therapies may be inefficiently high, one may be tempted to argue that there is a lack of market power of those holding patents on these new technologies. However, even without substitutes and very broad patents, as when demand is highly inelastic, we argued that the share of the social surplus

[26] In "Patent Expiration, Entry, and Competition in the U.S. Pharmaceutical Industry," R. Caves, M. Whinston, and M. Hurwitz 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.

allocated to the producer may be very small. This point was illustrated using the constant elasticity of demand case, in which an elasticity of 1.06 was shown to be consistent with a producer share of social surplus of 5%. In general, we showed that higher prices (such as those of HIV/AIDS drugs) induced by lower elasticities of demand often lead to less surplus captured by inventors.

In addition, the small estimated share of social surplus appropriated by investors sheds important light on the recent growth of alternative funding mechanisms to stimulate HIV/AIDS research, e.g. through advance purchasing contracts of governments or private foundations.[27] Given that there is a social surplus above a trillion dollars that is not appropriated by R&D investors, a few billion dollars added to stimulate innovation, as these public or private contracts seem to provide, seems to pale in comparison to interventions that would better allow innovators to capture the value of their innovations. Moreover, since both spending and markups are higher in the US than in the rest of the world (drug sales in the US account for more than half of worldwide spending) and price controls dominate foreign markets, estimates of appropriation based on US markets alone *over-estimate* worldwide appropriation.

It is important to stress that arguments about the difference between static and dynamic efficiency are a different matter than whether prices used for calculating spending in CE analysis reflect costs of production in general, and average costs of production (reflecting R&D costs), in particular. Under traditional CE analysis, even if one could measure costs perfectly, and did not need to approximate unobserved costs by observed prices, one would be concerned with the wrong measure, total ex-post surplus. This is because the *division* of the surplus is what matters for dynamic R&D policy, as opposed to only the *total* surplus which is relevant for static policy. In particular, this holds true whether the costs represented are marginal or average costs, the latter potentially including fixed costs such as R&D. In both cases, the division of social surplus is ignored but is what drives optimal R&D policy.

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" 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.[28] 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 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.[29]

---

[27] See e.g. The Center for Global Development (2005) and the efforts undertaken by The Gates Foundation (2005).

[28] Philipson, Mechoulan, and Jena (2005) discuss optimal technology assessment in the presence of altruism that motivates public healthcare delivery, in general, and R&D into third-world diseases, in particular.

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

Last, but not least, future research should consider the implications of CE analysis in environments in which there is excessive R&D due, for example, to so called patent-races or other forces (see Philipson et al. (2005) for an analysis of this issue in a more general setting). 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. If such patent races lead R&D to be over-provided, our conclusions emphasizing under-provision of R&D may be altered. However, there appears to be an almost universal policy towards subsidizing (as opposed to taxing) R&D, such that most nations have decided that the forces operating towards over-provision are dominated by those operating towards under-provision. In light of this, although incentives favoring over-provision may change the quantitative conclusions of our analysis, the qualitative conclusion that CE criteria limit already under-provided R&D seems generally applicable to most research areas and countries.

Our analysis and evidence, if they generalize to other technologies, suggest that interventions aimed at raising innovator appropriation may be desirable. This would lead to lower observed levels of cost-effectiveness for these technologies as traditionally measured from market prices, but would raise efficiency by not unduly sacrificing the health and well being of future patient populations for the benefit of current ones.

References

Arrow, Kenneth, "Economic Welfare and the Allocation of Research for Invention," in The Rate and Direction of Inventive Activity: Economic and Social Factors," e. R. Nelson, Princeton University Press (1961).

Becker, Gary S., Philipson, Tomas, Soares, Rodrigo, "The Quantity and Quality of Life and the Evolution of World Inequality," American Economic Review, 95 (2005): 277-291.

Caves, R., Whinston, M., Hurwitz, M., "Patent Expiration, Entry, and Competition in the U.S. Pharmaceutical Industry," Brookings Paper on Microeconomic Activity, Microeconomics (1991): 1-66.

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

Cutler, David M., and Richardson, Elizabeth, "The Value of Health: 1970-1990," American Economic Review, 88 (1998): 97-100.

Cutler, David M., and McClellan, Mark, "Is Technological Change in Medicine Worth It?," Health Affairs, 20 (2001): 11-29.

Duggan, Mark, and Evans, William, "The Impact of HIV Antiviral Treatments: Evidence for California's Medicaid Population," Working Paper.

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

Health, United States; National Center for Health Statistics with Chartbook on Trends in the Health of Americans, Hyattsville, MD (2002).

Garber, Alan M. and Phelps, Charles E., "Economic Foundations of Cost-Effectiveness Analysis," Journal of Health Economics 16 (1997): 1-32.

Garber, Alan M., "Advances in Cost-Effectiveness Analysis of Health Interventions," NBER Working Paper 7198 (1999).

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

Griliches, Zvi. R&D and Productivity: The Econometric Evidence, University of Chicago Press (1998).

Hall, Bronwyn H., "The Private and Social Returns to Research and Development," Technology, R&D, and the Economy, ed. B. Smith and C. Barfield, Brookings Institution/American Enterprise Institute, 1996.

Johannesson, M., and Weinstein, M.C., "On the Decision Rules of Cost-Effectiveness Analysis," Journal of Health Economics 12 (1993): 459-467.

Kates, J., and Wilson, A. "Medicaid & HIV/AIDS," Pub. 7172, Henry J. Kaiser Family Foundation (2004).

Lakdawalla, Darius, and Philipson, Tomas J.," Intellectual Property and Non-Price Discrimination," University of Chicago, Department of Economics Working Paper, 2005.

Lichtenberg, Frank R., "The Impact of Increased Utilization of HIV Drugs on Longevity and Medical Expenditure: An Assessment Based on Aggregate U.S. Time-Series Data," Working Paper , Columbia Business School, (2005).

Levin, Richard C., Klevorick, Alvin K., Nelson, Richard R., Winter, Sidney G., Gilbert, Richard, and Griliches, Zvi, "Appropriating the Returns from Industrial Research and Development," Brookings Papers on Economic Activity, Vol. 1987 (3), pp. 783-831.

Manning, Richard L., "Changing Rules in the Tort Law and the Market for Childhood Vaccines," Journal of Law and Economics, Vol. 37 (1), April, 1994, pp. 247-275.

Mansfield, Edwin, "How Rapidly Does New Technology Leak Out?," The Journal of Industrial Economics, Vol. 34(2), December, 1985, pp. 217-223.

Mansfield, Edwin, Rapoport, John, Romeo, Anthony, Wagner, Samuel, and Beardsley, George, "Social and Private Rates of Return from Industrial Innovation," The Quarterly Journal of Economics, Vol. 91(2), May, 1977, pp. 221-240.

Meltzer, David, "Accounting for Future Costs in Medical Cost-Effectiveness Analysis," Journal of Health Economics 16 (1997): 33-64.

Nordhaus, William D., "Schumpeterian Profits in the American Economy: Theory and Measurement," NBER Working Paper (2004).

Phelps, Charles E., and Parente, Stephen T., "Priority Setting in Medical Technology and Medical Practice Assessment," Medical Care 28 (1990): 703-723.

Philipson, T.J., "Economic Epidemiology and Infectious Disease," Chapter in Handbook of Health Economics. Edited by J. Newhouse and A. Culyer, 2000, Elsevier B.V., North Holland.

Philipson, T., and Mechoulan, S., "Intellectual Property & External Consumption Effects: Generalizations from Pharmaceutical Markets," NBER Working Paper #9598, 2003.

Philipson, T., and Jena, A.B., "Who Benefits from Medical Technologies? Estimates of Consumer and Producer Surpluses for HIV/AIDS Drugs," Forthcoming, Forum for Health Economics & Policy, 2005.

Pharmaceutical Research and Manufacturers of America, PhRMA Annual Membership Survey, 2003.

Neumann, Peter J., Sandberg, Eileen A., Bell, Chaim M., Stone, Patricia W., and Chapman, Richard H., "Are Pharmaceuticals Cost-Effective? A Review of the Evidence," Health Affairs, 19 (2000): 92-109.

Tirole, Jean. The Theory of Industrial Organization, The MIT Press (1988).

Weinstein, Milton C., and Manning, Willard G., Jr. "Theoretical Issues in Cost-Effectiveness Analysis" Journal of Health Economics 16 (1997): 121-128.

Weinstein, M.C., and Stason, W.B., “Foundations of Cost-Effectiveness Analysis for Health and Medical Practices,“ New England Journal of Medicine 296 (1977): 716-721.

Appendix

Share of Social Surplus Appropriated to Consumers and Producers Under Constant Elasticity Demand

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

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

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

$$q_m = \left[\frac{c \cdot \varepsilon}{x \cdot (\varepsilon - 1)}\right]^{-\varepsilon} \qquad p_m = \frac{c \cdot \varepsilon}{\varepsilon - 1}.$$

The gross consumer benefit $g(q_m)$ can be expressed by the following formula:

$$g = \int_0^{q_m} p(q)dq = \frac{x \cdot \varepsilon}{\varepsilon - 1} \cdot (q_m)^{\frac{\varepsilon - 1}{\varepsilon}}.$$

Similarly, the maximized profit can be written as:

$$\pi = p(q_m) \cdot q_m - c \cdot q_m = \frac{c \cdot q_m}{\varepsilon - 1}.$$

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,

$$\frac{\pi(q_m)}{g(q_c) - q_c \cdot c} = \frac{(cq_m)/(\varepsilon - 1)}{\frac{x\varepsilon}{\varepsilon - 1}(q_c)^{\frac{\varepsilon - 1}{\varepsilon}} - q_c \cdot c} = \frac{(cq_m)/(\varepsilon - 1)}{q_c\left(\frac{x\varepsilon}{\varepsilon - 1}(q_c)^{\frac{-1}{\varepsilon}} - c\right)} = \frac{(cq_m)/(\varepsilon - 1)}{q_c\left(\frac{x\varepsilon}{\varepsilon - 1}\left(\left(\frac{x}{c}\right)^{\varepsilon}\right)^{\frac{-1}{\varepsilon}} - c\right)}$$

$$= \frac{(cq_m)/(\varepsilon - 1)}{q_c\left(\frac{c\varepsilon}{\varepsilon - 1} - c\right)} = \frac{(cq_m)/(\varepsilon - 1)}{q_c\left(\frac{c}{\varepsilon - 1}\right)} = \frac{q_m}{q_c}$$

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:

$$\frac{\pi(q_m)}{g(q_c) - q_c \cdot c} = \left(\frac{\varepsilon - 1}{\varepsilon}\right)^{\varepsilon}$$

Using the above expressions, it is straightforward to derive the ratio of gross benefit to spending, z_R, as well:

$$z_R = \frac{g(q_m)}{p(q_m)q_m} = \frac{\varepsilon}{\varepsilon - 1}$$

Elasticity of Cost-Effectiveness with Respect to Supply and Demand Parameters

The ratio of gross benefit to spending (z_R) can be written generally as:

$$z_R = \frac{g(q(\eta,\theta),\eta)}{p(q(\eta,\theta),\eta)\cdot q(\eta,\theta)}$$

where η represents a vector of demand parameters and θ a vector of cost parameters. The elasticity of z_R with respect to demand and supply parameters is simply the elasticity of the numerator minus the elasticity of the denominator.

Noting that $g(q(\eta,\theta),\eta) = \int_0^{q(\eta,\theta)} p(y,\eta)dy$, we obtain:

$$\varepsilon_\theta^{z_R} \equiv \frac{d\ln z_R}{d\ln\theta} = \left[\frac{\theta}{g}\cdot\frac{dg}{d\theta}\right] - \left[\frac{\theta}{pq}\frac{d(pq)}{d\theta}\right]$$

$$= \left[\frac{\theta}{g}\cdot p\cdot\frac{dq}{\theta}\right] - \left[\frac{\theta}{pq}\cdot\left(q\frac{dp}{dq}\frac{dq}{d\theta} + p\frac{dq}{d\theta}\right)\right]$$

$$= \left[\frac{1}{z_R}\varepsilon_\theta^q\right] - \left[\varepsilon_\theta^q\cdot\left(1 - \frac{1}{\varepsilon_p^q}\right)\right]$$

where ε_p^q is the elasticity of demand with respect to price, and ε_θ^q is the elasticity of demand with respect to the cost parameter. Both are evaluated at (η,θ).

Similarly, consider the elasticity of z_R with respect to demand parameters:

$$\varepsilon_\eta^{z_R} \equiv \frac{d\ln z_R}{d\ln\eta} = \left[\frac{\theta_D}{g}\cdot\frac{dg}{d\eta}\right] - \left[\frac{\eta}{pq}\frac{d(pq)}{d\eta}\right]$$

$$= \left[\frac{\eta}{g}\cdot\left(p\cdot\frac{dq}{d\eta} + \int_0^{q(\eta,\theta)}\frac{dp(y,\eta)}{d\eta}dy\right)\right] - \left[\frac{\eta}{pq}\cdot p\cdot\frac{dq}{d\eta} + \frac{\eta}{pq}\cdot q\cdot\left(\frac{dp}{dq}\frac{dq}{d\eta} + \frac{dp}{d\eta}\right)\right]$$

$$= \left[\frac{pq}{g}\cdot\varepsilon_\eta^q + \int_0^{q(\eta,\theta)}\frac{p(y,\eta)}{g}\varepsilon_\eta^p(y,\eta)dy\right] - \left[\varepsilon_\eta^q\cdot(1 - \frac{1}{\varepsilon_p^q}) + \varepsilon_\eta^p\right]$$

$$= \left[\frac{1}{z_R}\cdot\varepsilon_\eta^q + \varepsilon_\eta^g\right] - \left[\varepsilon_\eta^q\cdot(1 - \frac{1}{\varepsilon_p^q}) + \varepsilon_\eta^p\right]$$

Where ε_{η}^{g} is the elasticity of the gross benefit with respect to the demand parameter, *holding output constant*, ε_{η}^{q} is the elasticity of optimal demand with respect to the demand parameter, and ε_{η}^{p} is the elasticity of price (or the WTP) with respect to the demand parameter evaluated at (η,θ).