# When Is Prevention More Profitable than Cure? The Impact of Time-Varying Consumer Heterogeneity

**Michael Kremer and Christopher M. Snyder**

## Abstract

We argue that in pharmaceutical markets, variation in the arrival time of consumer heterogeneity creates differences between a producer's ability to extract consumer surplus with preventives and treatments, potentially distorting R&D decisions. If consumers vary only in disease risk, revenue from treatments—sold after the disease is contracted, when disease risk is no longer a source of private information—always exceeds revenue from preventives. The revenue ratio can be arbitrarily high for sufficiently skewed distributions of disease risk. Under some circumstances, heterogeneity in harm from a disease, learned after a disease is contracted, can lead revenue from a treatment to exceed revenue from a preventative. Calibrations suggest that skewness in the U.S. distribution of HIV risk would lead firms to earn only half the revenue from a vaccine as from a drug. Empirical tests are consistent with the predictions of the model that vaccines are less likely to be developed for diseases with substantial disease-risk heterogeneity.

JEL Codes: O31, L11, I18, D42

Center for Global Development
www.cgdev.org

Working Paper 334
July 2013

**When Is Prevention More Profitable than Cure? The Impact of Time-Varying Consumer Heterogeneity**

Michael Kremer
Harvard University, Brookings Institution,
Center for Global Development, NBER

Christopher M. Snyder
Dartmouth College, NBER

A previous version of this paper was circulated under the title "Why Is There No AIDS Vaccine?" The authors are grateful to Marcella Alsan, Emmanuelle Auriol, Chris Avery, Bryan Boulier, Ryan Bubb, Jim Dana, Esther Duflo, Glenn Ellison, Amy Finkelstein, Jon Hamilton, Corinne Langinier, Scott Lee, David Malueg, David McAdams, Sendhil Mullainathan, Robert Porter, Michael Schwarz, Andrew Segal, Lars Stole, Heidi Williams, and seminar participants at the American Enterprise Institute, Dartmouth, Harvard, MIT, Northeastern, Northwestern, NYU, Princeton, RAND, Stanford, UCLA, University of Pennsylvania, University of Rochester, University of Toronto, the IAEN Symposium on the Economics of AIDS/HIV in Developing Countries (Barcelona), the IDEI Conference on Markets for Pharmaceuticals and the Health of Developing Nations, the International Industrial Organization Conference (Boston), the NBER Summer Institutes on Health and Aging and on Innovation and the Global Economy, and the Southern Economic Association Conference (Charleston) for helpful comments; to David Blanch flower for sharing his data; and to Lindsey Beckett, Ruben Enikolopov, Cacey Tang, Suzanne Wang, and Dan Wood for excellent research assistance.

CGD is grateful to its funders and board of directors support of this work.

Michael Kremer and Christopher M. Snyder . 2013. "When Is Prevention More Profitable than Cure? The Impact of Time-Varying Consumer Heterogeneity." CGD Working Paper 334. Washington, DC: Center for Global Development.
http://www.cgdev.org/publication/when-prevention-more-profitable

**Center for Global Development**
**1800 Massachusetts Ave., NW**
**Washington, DC 20036**

202.416.4000
(f) 202.416.4050

**www.cgdev.org**

# 1. Introduction

Many public-health and industry experts believe that firms earn more revenue from disease treatments, such as drugs, than preventives, such as vaccines (see, e.g., Rosenberg 1999), and that stronger government support is needed for the development of preventive health technologies. These views are especially common in the case of HIV (human immunodeficiency virus) (see, e.g., Thomas 2002), and indeed governments have established special programs to support HIV vaccine research such as the International AIDS Vaccine Initiative (IAVI).

In this paper, we argue that time-varying consumer heterogeneity can drive a wedge between relative private and social incentives to invest in preventives and treatments. We show that heterogeneity among consumers in disease risk will limit firms' ability to extract consumer surplus from preventives, biasing firms' R&D incentives away from preventives toward treatments compared to a social planner's. Heterogeneity in harm from infection, on the other hand, can lead to the opposite bias, toward preventives. The model also suggests when these biases are likely to be quantitatively important. Firms' bias against investing in preventives will be strongest for diseases with skewed distributions of disease risk (i.e., with high risk concentrated in a small segment of the population). Common diseases cannot exhibit much skewness as a mathematical principle, so the bias toward preventives has the most scope to affect R&D incentives for relatively rare diseases with risk heterogeneity. Calibrations for HIV suggest that the rates of partner change and other risk factors are sufficiently skewed that the bias against HIV preventives may be substantial. Empirically, we find that the relative likelihood of developing treatments compared to preventives is greater for diseases with heterogeneity in disease risk.

To see why heterogeneity in disease risk can lead to a bias against preventives, consider the following numerical example. Suppose a monopoly pharmaceutical manufacturer sells directly to 100 rational, risk-neutral consumers, who will suffer harm quantified as \$100 from contracting the disease. The firm can develop a treatment or a preventive; both are costless to manufacture, are perfectly effective, and have no side effects. Suppose first that the consumers are homogeneous, having the same 19% risk of contracting a disease. If the firm develops a treatment, it can sell to all people who contract the disease at a price (equal to the avoided harm) of \$100. Expected treatment revenue is \$1,900 because an expected 19 consumers contract the disease and buy the treatment. If the firm develops a preventive, it could sell to all 100 consumers at a price, equal to consumers' expected harm of \$19, for total revenue of \$1,900. With homogeneous consumers, the firm's expected revenue is \$1,900, which represents full extraction of consumer surplus in the market, whether it sells a treatment or preventive.

Consider the same example except suppose now that consumers are heterogeneous in disease risk, with 90 having a 10% chance of contracting a disease while 10 have a 100% chance. Because the number of people expected to contract the disease is the same 19 as in the homogenous-consumer case, expected treatment revenue remains the same at \$1,900. On the other hand, the firm's revenue from a preventive falls. The firm can either sell to the 10 high-risk consumers at their expected harm of \$100, or sell to all consumers

at a price of $10, equal to the low-risk consumers' expected harm. Either way, the firm's revenue from a preventive is $1,000, only about half the revenue from a treatment and only about half the social value of the product.

To understand how these results generalize, in Section 2 we provide a simple benchmark model of a monopolist that sells a perfectly safe and effective product, which is costless to produce, directly to rational, risk-neutral consumers. The analysis in Section 3 shows that in this simple benchmark model, if consumers are heterogeneous only in disease risk, then for any disease-risk distribution, monopoly revenue is higher with a treatment than a preventive. The potential social welfare loss from distortions in research incentives is related to the ratio of preventive to treatment revenue: as a percentage of total disease burden, the social welfare loss can be as high as one minus this ratio. The ratio of preventive to treatment revenue equals 1/2 for a uniform distribution of disease risk, is greater than 1/2 for monotonic distributions that are negatively skewed, and is less than 1/2 for monotonic distributions that are positively skewed. Indeed for sufficiently positively skewed distributions, the revenue ratio can be driven to zero; and thus the social cost of distortions in R&D incentives can approach the entire disease burden.

How close to zero the ratio can be driven is limited by the prevalence of the disease in the population. This point is easiest to see in the extreme case in which the disease is ubiquitous: if nearly everyone is expected to contract the disease, there is little scope for the distribution of disease risk to exhibit the dispersion required to generate a substantial gap between preventive and treatment revenue. We compute a tight lower bound on the ratio as a function of disease prevalence and show that this bound is strictly increasing. The implication is that risk heterogeneity can induce little difference in the incentives to develop preventives versus treatments for the most common diseases; diseases must be sufficiently rare for heterogeneity in disease risk to substantially impair firms' relative incentives to develop preventives.

In Section 4 we consider other types of heterogeneity. We first consider the case in which consumers obtain private information not ex ante, as with disease risk, but ex post, after contracting the disease. Polio, for example usually leads to fairly minor, transient symptoms, but sometimes leads to paralysis. If consumers are ex ante homogeneous, but differ in harm ex post, preventive manufacturers will be able to fully extract consumer surplus while treatment manufacturers will not. If there are multiple sources of private information revealed at different times, the correlation among these sources affects relative R&D incentives. For example, if consumers vary ex ante in disease risk and also in a factor such as income that affects both ex ante and ex post willingness to pay and if these factors are independent, firms' bias against preventives will be dampened. We consider a range of different correlation structures among the variety of sources of private information.

In Section 5 we extend the benchmark model to richer, policy-relevant institutional structures such as insurance contracts and government purchasing.[1] If firms can sell future access to their products through

[1]The appendix provides an extension of the model in which we relax the monopoly assumption, allowing for entry by both preventive and treatment manufacturers and also allowing for entry by generics after a period of patent protection. The prospect of an effective treatment can wipe out the market for a preventive, but the currently ill form a captive market for a treatment, providing revenue even under the threat of entry by preventives and generics.

insurance or other contracts, then treatment manufacturers can always imitate preventive manufacturers and so earn at least as much revenue with a treatment as a similarly effective preventive. If third-party purchasers such as HMOs or governments can negotiate with pharmaceutical firms over fixed fees, they can potentially eliminate the deadweight loss associated with pricing above marginal cost. Assuming that bargaining takes place after R&D costs are sunk, under plausible assumptions the pharmaceutical manufacturer and the third-party purchaser will each capture some of the gain associated with eliminating deadweight loss. Thus the biases in R&D decisions we found under direct-to-consumer sales will survive (though typically attenuated) under sales to third-party purchasers. This can be seen as a standard (Klein, Crawford, and Alchian 1978) hold-up problem. However, if bargaining takes place before R&D costs are sunk, it may be possible to match private and social incentives for preventive and treatment R&D. This provides a potential justification for institutional mechanisms that help commit to pricing for preventives in advance, such as those found in the de facto operation of the Advisory Committee on Immunization Policy in the United States or the International Pneumococcus Advance Market Commitment (Kremer and Glennerster 2004; Snyder, Begor, and Berndt 2011).

Having established theoretical bounds on the ratio of preventive to treatment revenue, in Section 6 we calibrate where between these bounds the revenue ratio might fall in practical examples. We focus on the case of HIV. Using U.S. data on the distribution of sexual partners (as well as other risk factors) to infer infection risk, the highly skewed distribution of sexual partners leads to a highly skewed distribution of HIV infection risk, in turn leading calibrated revenue from a preventive to fall short of that from a treatment by a factor of between two and four. The bias persists when the joint distribution of disease risk with income is considered. The results for HIV contrast with additional calibrations for HPV (human papillomavirus), a more common disease with an infection-risk distribution that is consequently less skewed. We find that calibrated revenue for an HPV preventive and is close to that for a treatment, suggesting that firms may have less bias against developing preventives for HPV.

Section 6 also provides a separate set of calibrations based on the joint distribution of income and HIV risk across countries to shed light on how changes in firms' ability to price discriminate internationally could potentially affect R&D incentives for HIV preventives relative to treatments. We find that if firms' existing ability to price discriminate across countries were eliminated, drug revenue could potentially fall below vaccine revenue.

Section 7 empirically tests whether infection-risk heterogeneity affects whether preventives or treatments are developed for different diseases. We construct a unique dataset including proxies for heterogeneity in infection risk (e.g., STIs, disease concentration in certain subpopulations or regions or transmission through specialized vectors) for a cross-section of diseases. We find that disease-risk heterogeneity significantly reduces the probability of vaccine development—by over 25 percentage points—but has no effect on drug development, consistent with the theory from Section 3.

Of course in identifying a new factor, time-varying consumer heterogeneity, which may affect the relative profitability of R&D on preventives and treatments relative to their social value, we do not seek to deny

the potential role of other factors. While acknowledging that other factors—scientific and technological difficulties of developing new products, manufacturing and delivery costs—may differ between preventives and treatments; we note that these other factors will not necessarily create a wedge between relative private and social incentives to invest in preventives as opposed to treatments. In the interest of parsimony, our benchmark model involves rational, risk-neutral consumers who do not face credit constraints. We recognize that risk aversion, credit constraints, and behavioral factors could also affect willingness to pay for preventives rather than treatments. The model could be readily extended to incorporate such factors in future work.

Our work contributes to several literatures. It is well understood that epidemiological externalities may limit the ability of pharmaceutical firms to capture social value from products to prevent or treat infectious disease, and this issue may be more acute for preventives than treatments. Papers that examine firm incentives in the presence of epidemiological externalities include Brito, Sheshinski, and Intrilligator (1991); Boulier (2006); Francis (1997); Geoffard and Philipson (1997); Gersovitz (2003); and Gersovitz and Hammer (2004, 2005). A companion paper (Kremer, Snyder, and Williams 2012) examines the determinants of the magnitude of these effects. However, the analysis in this paper applies to preventives more generally rather than only to infectious diseases, and thus has implications that are analytically distinct than those considered in the literature on epidemiological externalities. For example, our analysis also applies to vaccines that are not subject to epidemiological externalities, such as vaccines against shingles (a recurrence of childhood chickenpox infection in adults), as well as to preventives against non-infectious diseases, such as cholesterol-reducing drugs or heart-disease preventives. Within the class of vaccines against infectious diseases, our analysis suggests that, independent of epidemiological externalities, biases against vaccines will be particularly severe for diseases with skewed distributions of disease risk, such as HIV; but that there may be other diseases with heterogeneity in harm or strong negative correlations between income and infection risk where heterogeneity creates a bias towards vaccines rather than treatments.

Our work is related to the industrial organization literature on monopoly pricing when consumers gradually learn their demands. Lewis and Sappington (1994) and Courty (2003) assume consumers are initially identical, whereas we assume consumers have ex ante private information about their disease risk. Courty and Li (2000) compare optimal ex ante and ex post schemes under general conditions, where ex ante schemes are allowed to involve refunds. Refunds are impossible for preventives because, once the preventive is administered, the benefit is inalienable from the consumer. Clay, Sibley, and Srinagesh (1992) and especially Miravete (1996) are closest to our work. Our application to disease risk calls for a specific mapping from ex ante private values into ex post types, whereas Miravete considers general functional forms for the mapping. The specificity in this one dimension allows us to examine general distributions of ex ante disease risk rather than the particular class of beta distributions examined by Miravete, and to establish bounds on the profit ratio as a function of skewness of the disease-risk distribution and as a function of disease prevalence, all of which are new results in the literature. Our analysis of social welfare in Section 3, calibrations and empirical work, and the appendix analyzing generic competition between preventives and treatments after expiry of intellectual-property rights are new as well.

Our computation of tight bounds on the ratio of preventive to treatment profit is related to Malueg's (1994) bounds on the ratio of monopoly to competitive welfare as a function of the demand curvature and to Maleug and Snyder's (2006) bound on the ratio of profits from discriminatory to uniform pricing as a function of the number of markets. We also contribute to the literature on the response of innovation in R&D-intensive industries (see Newell, Jaffee, and Stavins 1999; Acemoglu and Linn 2004; Finkelstein 2004).

## 2. Model

We begin with a stylized benchmark model of a monopolist selling directly to consumers, deferring analysis of more complicated models with third-party purchasers to Section 5 and of models with competition among producers to the appendix (Appendix B). The monopoly pharmaceutical manufacturer faces a choice of developing a preventive or treatment.

To simplify the presentation, we will initially consider the case in which preventives and treatments are perfectly effective, have no side effects, and are costless to manufacture and administer. (Proposition 14 in the appendix shows that the key results continue to hold when these assumptions are relaxed.) The firm's only cost is the present discounted value of the fixed cost of developing product $j$, denoted $k_j \in [0,\infty)$, where $j = p$ for the preventive and $j = t$ for the treatment. Let $p_j \in [0,\infty)$ be the present discounted value of the price the firm receives for product $j$. Let $\pi_j$ be producer surplus (equivalently revenue in the case of costless production), $\Pi_j = \pi_j - k_j$ be profit, $CS_j$ be consumer surplus, $WE_j = CS_j + \Pi_j$ be equilibrium social welfare, and $WF_j$ be first-best social welfare (i.e., social welfare when the product's price is set to marginal cost) from product $j$. The difference between these two social-welfare measures is deadweight loss: $DWL_j = WF_j - WE_j$. Using notation that drops the subscript $j$ for products, let $WE$ be equilibrium social welfare given the firm's equilibrium choice of product, $WF$ be first-best social welfare given the first-best choice of product, and $DWL$ be the difference $DWL = WF - WE$.

Consider the case in which the firm sells directly to risk-neutral consumers. Before purchasing any product, consumer $i$ learns his or her disease risk, $x_i \in [0,1]$, i.e., the probability he or she contracts the disease. Assume $x_i$ is a random variable with cumulative distribution function $F(x_i)$. Normalizing the mass of consumers to unity, the mass of consumers with disease risk at least as great as some value $x$ is denoted $\bar{F}(x_i) = 1 - F(x_i)$. The mean disease risk in the population (also the realized disease prevalence in the absence of a preventive) is $x = \int_0^1 x_i \, dF(x_i)$. Assume the firm knows the distribution of $x_i$ in the population but cannot price discriminate across consumers based on $x_i$.[2]

If a consumer contracts a disease and has not had the preventive or does not receive the treatment, he or she experiences harm $h_i \in [0,\infty)$ in present discounted value terms. In this and the next section, we will assume that consumers all would pay the same amount to avoid harm $h$, but in Section 4 we will

[2]Price discrimination can be ruled out if $x_i$ is private information for consumers (for example, related to their sexual behavior or intravenous drug use, conducted in private) or if $x_i$ is public information but discrimination is prevented by the difficulty of controlling resale or other administrative, institutional, or legal barriers.

generalize the analysis to allow consumers to have various sources of heterogeneity in willingness to pay. Let $D = h\int_0^1 x_i\, dF(x_i) = hx$ be the total social burden of the disease, a term we will use to normalize our welfare measures in the subsequent analysis.

We next turn to a preliminary analysis of which product the firm chooses to develop. If the firm develops a preventive, consumers purchase before becoming infected. A consumer with disease risk $p_p/h$ would be indifferent between purchasing the preventive at price $p_p$ and not.[3] The preventive producer thus sells to the mass of consumers $\bar{F}(p_p/h)$ with disease risk $x_i \geq p_p/h$, implying the profit from developing a preventive is

$$\Pi_p = \max_{p_p \in [0,\infty)} [p_p \bar{F}(p_p/h)] - k_p. \tag{1}$$

If the firm develops a treatment, on the other hand, the consumer purchases after becoming infected. The profit from developing a treatment is

$$\Pi_t = hx - k_t. \tag{2}$$

Equation (2) holds because the treatment is optimally sold at a price that extracts the consumer's entire ex post surplus $p_t^* = h$; the treatment is purchased by the mass $x$ of consumers who become infected. The firm develops a preventive if $\Pi_p > \max(\Pi_t, 0)$, a treatment if $\Pi_t > \max(\Pi_p, 0)$, and neither if $\max(\Pi_p, \Pi_t) < 0$.[4]

## 3. Equilibrium with Ex Ante Heterogeneity in Disease Risk

If consumers are homogeneous, then there is no wedge between private and social R&D incentives, and the first best is obtained in equilibrium, as the following proposition states.

**Proposition 1.** *Assume there is no heterogeneity in the distribution of disease risk, i.e., $x_i = x$ for all i. In equilibrium the firm makes the first-best product choice and produces the first-best quantity of the product.*

The proposition follows immediately from the fact that the monopolist can extract 100% of the surplus from homogeneous consumers with either product and thus fully internalizes social welfare.[5]

Heterogeneity in consumers disease risk will drive a wedge between private and social R&D incentives. In the model, the firm cannot perfectly price discriminate based on disease risk and so is no longer able to extract 100% of consumer surplus with a preventive. Producer surplus from a preventive, $\pi_p$, will thus fall below producer surplus from a treatment, $\pi_t$, as Proposition 2, proved in the appendix (Appendix A), states.

**Proposition 2.** *Assume there is nontrivial heterogeneity in the distribution of disease risk; i.e., at least two distinct subintervals of $(0,1]$ have positive measure. Then $\pi_p < \pi_t$.*

[3] Arguments along the lines of Theorem 4 of Harris and Raviv (1981) establish that a simple linear price $p_p$ is optimal among the set of potentially complicated mechanisms that might be used to sell the preventive.

[4] The remaining strategy—the firm develops both products—can be ignored in the analysis because it is weakly dominated given products are perfectly safe, effective, and costless to manufacture. Appendix B allows for the possibility that both products are developed in an extension with general parameter values and potential competition between manufacturers.

[5] The firm may no longer have first best incentives for product development if we depart from the monopoly assumption by allowing patent races, finite patent lives, rent-dissipating competition, etc. Appendix B analyzes these issues further.

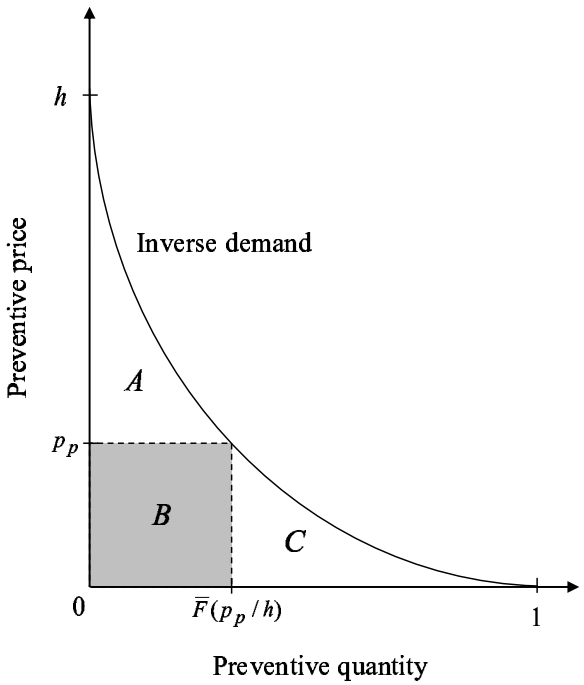


**Figure 1:** Geometric comparison of producer surplus from a preventive and a treatment.

Figure 1 sketches a simple graphical proof of Proposition 2. Producer surplus from a preventive, $\pi_p$, equals the area of the largest rectangle that can be inscribed under inverse demand curve $\bar{F}(p_p/h)$, while $\pi_t$ equals the area under the whole curve. No matter how the rectangle is inscribed, and no matter the shape of the curve, the area of the rectangle will be less than the area under the whole curve, so $\pi_t > \pi_p$.

The result from Proposition 2 that $\pi_p < \pi_t$ has consequences for social welfare because it leaves room for cases in which the firm prefers to develop the treatment even though the preventive is cheaper to develop ($k_p < k_t$) and hence would be developed in the first best. The measure of such cases is what we mean by the firm's "bias" against preventives. The lower is $\pi_p$ relative to $\pi_t$, the greater the firm's bias against preventives. The producer-surplus ratio $\pi_p/\pi_t$ (more precisely, one minus this ratio) provides a convenient index of the bias against preventives because this ratio can be linked to the potential social cost of this bias, as Proposition 3, proved in the appendix, formalizes.

**Proposition 3.** *The difference between first-best social welfare, WF, and equilibrium social welfare, WE, as a percentage of the total disease burden, D, has a tight upper bound given by* $1-\pi_p/\pi_t$. *Formally,*

$$\sup_{(k_p,k_t)\in[0,\infty)^2}\left[\frac{WF-WE}{D}\right] = 1-\frac{\pi_p}{\pi_t}.$$

Proposition 2 states that the firm will be biased against preventives if there is heterogeneity in disease risk, raising the question of how large this bias can possibly be. The next proposition, proved in the appendix, states that in the case in which consumers fall into discrete risk classes, the number of risk classes determines a tight lower bound on the relative producer surplus from a preventive.

**Proposition 4.** *Distributions of consumers into C risk classes can be constructed such that $\pi_p/\pi_t$ can be made arbitrarily close to $1/C$, a lower bound on $\pi_p/\pi_t$.*

The Introduction offered an example of a disease with harm of \$100 and with two risk classes (90 consumers with a 10% chance of contracting the disease and 10 with a 100% chance) in which expected revenue from a treatment was \$1,900, while a preventive producer would earn \$1,000 either by selling at \$100 to 10 high risk customers or \$10 to all 100 customers $\pi_p/\pi_t = 0.53$. The fact that this result was close to 1/2 was no accident: an implication of Proposition 4 is that $\pi_p/\pi_t$ can be driven down as low as, but no lower than, 1/2 in examples with two risk classes. The example can be extended to show how it is possible to keep increasing treatment revenue by adding risk classes while leaving preventive revenue constant. Consider adding a third risk class with 900 individuals with a 1% disease risk. Revenue from a preventive is unchanged at \$1,000 because the firm also earns this much from selling to all 1,000 consumers at the highest price the new consumers are willing to pay (\$1). Expected treatment revenue rises to \$2,800, equal to the \$1,900 earned from the original two risk classes plus \$900 from the nine consumers expected to contract the disease in the added risk class. Adding new risk classes with 10 times the consumers in the previous one having $1/10$ the disease risk leads to a \$900 increase in treatment revenue leaving preventive revenue unchanged. The ratio $\pi_p/\pi_t$ falls from 0.53 to 0.36 to 0.27 as the number of risk classes is increased from 2 to 3 to 4 in this extended example; note these values are close to the $1/C$ bound stated in the proposition.

An immediate consequence of Proposition 4 is that there exist distributions of consumer types such that $\pi_p/\pi_t$ can be made arbitrarily small. This can be seen by taking the limit as $C$ approaches infinity in the proposition.

**Proposition 5.** *There exist distributions of consumers such that $\pi_p/\pi_t$ can be made arbitrarily close to zero.*

When is the bias likely to be large? As the intuition from the two-type example provided in the Introduction suggests, the bias against preventives is especially large when a large segment of the population has a very small probability of contracting the disease and a small segment of the population has a high probability. Translated in more general terms, the bias against preventives should be expected to be largest when the distribution of disease risk is skewed. Proposition 6 provides a formal statement of the relationship between skewness of the disease-risk distribution and the ratio of producer surplus $\pi_p/\pi_t$.

**Proposition 6.** *Let $f(x_i)$ be a differentiable density function associated with consumer types $x_i$. If $f'(x_i) = 0$ (implying $x_i$ is uniformly distributed), then $\pi_p/\pi_t = 1/2$. If $f'(x_i) > 0$ (a sufficient condition for negative skewness), then $\pi_p/\pi_t > 1/2$. If $f'(x_i) < 0$ (a sufficient condition for positive skewness), then $\pi_p/\pi_t < 1/2$.*

The proof is illustrated in Figure 2. The case $f'(x_i) = 0$ is drawn in Panel I of the figure. If $f'(x_i) = 0$, then $x_i$ is uniformly distributed and has no skewness. The associated inverse demand curve $\bar{F}(p_p/h)$ turns out to be linear. Standard results imply that the area of the largest rectangle that can be inscribed under a linear demand curve is half of the area under the curve, so $\pi_p/\pi_t = 1/2$. If $f'(x_i) > 0$ as in Panel II of the

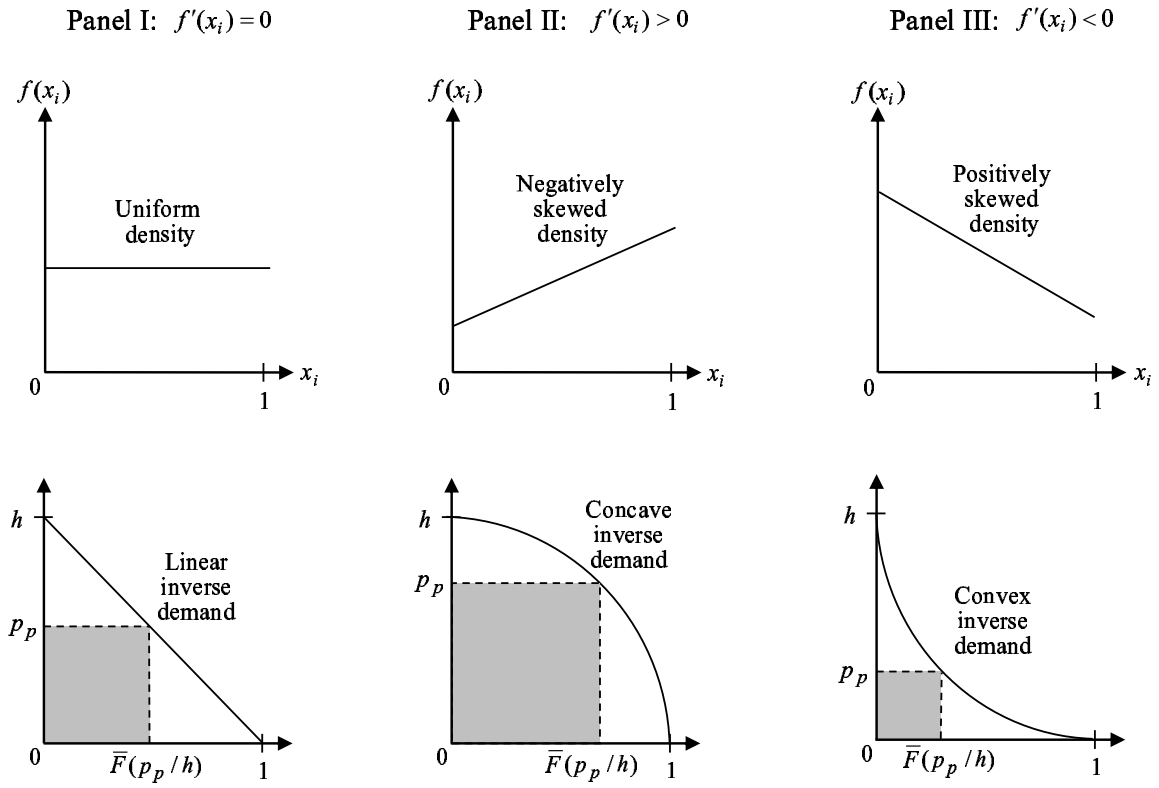


**Figure 2:** Ratio of producer surpluses depends on skewness of density and curvature of inverse demand.

figure, then the distribution of $x_i$ is negatively skewed. The associated inverse demand is then concave. As the figure shows, the area of the largest rectangle that can be inscribed under the inverse demand curve is more than half the area under the inverse demand curve, so $\pi_p/\pi_t > 1/2$. If $f'(x_i) < 0$ as in Panel III of the figure, then the distribution of $x_i$ is positively skewed, and the associated inverse demand is convex. As the figure shows, the area of the largest rectangle that can be inscribed under the inverse demand curve is less than half the area under the curve, so $\pi_p/\pi_t < 1/2$.

We saw from Proposition 6 that the revenue ratio $\pi_p/\pi_t$ is bounded below if the monotone disease-risk distribution is uniform or negatively skewed. Another lower bound on the revenue ratio can be obtained by focusing on the prevalence of the disease, which in the absence of a preventive equals $x$. Such a bound is empirically useful because prevalence is readily observable. Intuitively, if $x$ is close to 1, most consumers' disease risk must be close to 1, limiting how much heterogeneity there can be in the distribution of disease risk. Lower values of $x$ allow for more heterogeneity in disease risk, but there are limits to this heterogeneity for any given value of $x$.

**Proposition 7.** *Take the prevalence of the disease in the absence of a preventive, x, to be some constant in* $[0,1]$. *A tight lower bound on* $\pi_p/\pi_t$ *is provided by the implicit solution for B in*

$$B[1-\ln(Bx)] = 1. \tag{3}$$

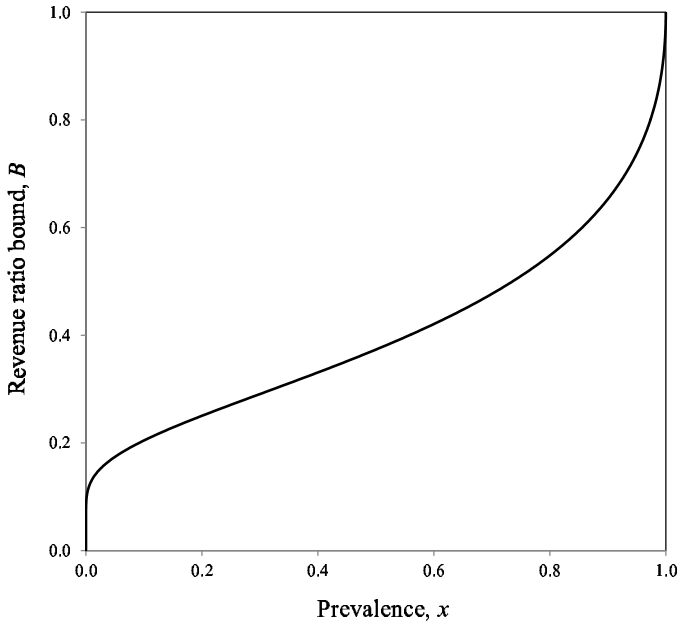


**Figure 3:** Lower bound on ratio of preventive to treatment revenue as function of prevalence.

*B is strictly increasing in x, with* $\lim_{x\to 0} B = 0$ *and* $\lim_{x\to 1} B = 1$.

Figure 3 graphs numerical solutions for $B$ as a function of $x$. The empirical implication of the figure is for the most common diseases, disease-risk heterogeneity cannot be an important factor in a firm's decision to develop a preventive versus a treatment. For example, the figure shows that if the prevalence of the disease is above 0.74, it is mathematically impossible to generate enough disease-risk heterogeneity to drive $\pi_p/\pi_t$ below 1/2. The results from this section that heterogeneity and skewness in disease risk contribute to a bias against preventives are more likely to apply to sufficiently rare diseases.

We conclude the section by drawing out the social-welfare implications of the analysis. The next proposition, proved in the appendix, states that there is socially too little incentive to develop a preventive relative to a treatment.

**Proposition 8.** *If consumers differ only in ex ante disease risk, the firm never develops a preventive in preference to a treatment unless it is socially efficient to do so. There exist cases in which the firm develops a treatment but it would have been socially efficient to develop a preventive.*

Proposition 8 holds whether social efficiency is measured by first-best social welfare ($WF_j$) or equilibrium social welfare ($WE_j$). The main social-welfare implications of Propositions 1 through 6 should also be emphasized. Proposition 5 implies that $1-\pi_p/\pi_t$ can approach one, implying that the potential social cost of the bias against preventives can be as large as the entire disease burden $D$ itself. Proposition 6 implies that the potential social cost of the bias against preventives can be as much as half the disease burden for uniformly distributed disease risk, less for negatively skewed distributions, and more for positively skewed distributions. In sum, the model with consumer heterogeneity in the single dimension of disease risk sug-

gests that R&D decisions may be biased against preventives and that the social loss from these biases can be quite large for positively skewed distributions of disease risk.

# 4. Other Sources of Heterogeneity

The previous section restricted attention to one source of consumer heterogeneity: disease risk, $x_i$. This source gives consumers private information only in the ex ante period; ex post, the act of seeking treatment reveals the consumer's disease status. In this section we examine alternative sources of heterogeneity with different timing structures. Section 4.1 examines heterogeneity in income, $y_i$, a source of private information ex ante that persists ex post because it is not necessarily revealed by disease status. We show that this source of heterogeneity is neutral, not generating a bias toward either preventives or treatments. Section 4.2 examines heterogeneity in disease harm, $h_i$, realized after the disease is contracted, thus an ex post source of private information. We prove that this source of heterogeneity flips the results from the case of heterogeneity in $x_i$ and leads to a bias against treatments.

In addition to looking at the individual sources of heterogeneity, the section provides results for various combinations of them. To allow for multiple sources of heterogeneity, the notation for random variables needs to be generalized. Let $x_i$, $y_i$, and $h_i$ be realizations of random variables $X$ representing disease risk, $Y$ representing income, and $H$ representing disease harm. Let $V \in \{X,Y,H\}$ denote the generic representation of one of these random variables. Associated with $V$, let $v_i$ be a realization, $F_V(v_i)$ the distribution function, $\bar{F}_V(v_i) = 1-F_V(v_i)$ the complementary distribution, $v = \int_0^{\bar{v}} v_i dF_V(v_i)$ the mean, and $[0,\bar{v}]$ the support. Finally, define the value function $R_V = \max_{v_i}[v_i\bar{F}_V(v_i)]$. In words, $R_V$ is the largest area under the "demand curve" $\bar{F}_V(v_i)$ that can be captured by an inscribed rectangle.

## 4.1. Income Heterogeneity

Assume that a consumer's willingness to pay to avoid certain harm is given by his or her income, $y_i$. Assume the consumer learns $y_i$ ex ante, and this private information remains in the ex post period. Assume the firm only knows the distribution of $y_i$ or, if it can observe $y_i$, cannot discriminate on the basis of this information.

Suppose there is no other source of heterogeneity than $y_i$.[6] It is immediate that consumer heterogeneity in $y_i$ alone reduces the revenue the firm can obtain from either product, but does not bias the firm toward either product because the firm faces the same private information ex ante when preventives are sold as ex post when treatments are sold. Producer surplus is the same for both products.

Next consider combined heterogeneity in $x_i$ and $y_i$. Assume the firm cannot discriminate on $x_i$ or $y_i$. The analysis is complicated by the fact that $x_i$ is no longer a pure source of ex ante heterogeneity because it may be correlated with $y_i$, also a source of ex post heterogeneity. The expressions for preventive and treatment profit do not lead to particularly informative comparisons for general distributions between $x_i$ and $y_i$. To build intuition, therefore, our approach will be to analyze three special cases that span the set of

[6] Kessing and Nuscheler (2006) also study monopoly vaccine pricing when income is the sole source of consumer heterogeneity. Their dynamic model generates a feedback effect whereby leaving the poor susceptible increases the willingness to pay of the rich.

possibilities: $x_i$ and $y_i$ are independent; $y_i$ is an increasing deterministic function of $x_i$; $y_i$ is a decreasing deterministic function of $x_i$ (in particular we will take $y_i$ to be inversely proportional to $x_i$). We will use this same three-pronged approach to analyze other combinations of sources of heterogeneity below as well. While the analysis is restricted to just these three special cases here, it is in fact possible to compute and compare preventive and treatment profits given any specific joint distribution of $x_i$ and $y_i$; and in Section 6 we illustrate how this can be calibrated with data on the distributions of HIV risk and income in the United States and the world.

Start then by assuming that $x_i$ and $y_i$ are independent. Define the product $z_i = x_i y_i$, representing the consumer's willingness to pay to avoid harm from the disease from an ex ante perspective, i.e., before knowing whether he or she has contracted it but only knowing his or her risk. Using the standard notation, let $F_Z(z_i)$ be the associated distribution function, $\bar{F}_Z(z_i)$ the complementary distribution, and $z$ the mean. The support of $z_i$, derived from the supports of $x_i$ and $y_i$, is $[0, \bar{y}]$.

First consider the preventive producer's profit-maximization problem. Consumers buy the preventive if $z_i \geq p_p$, implying the demand for the preventive is $\bar{F}_Z(p_p)$. Hence

$$\pi_p = \max_{p_p \in [0,\infty)} [p_p \bar{F}_Z(p_p)] = R_Z. \tag{4}$$

In fact, (4) is the general formula for preventive revenue, which holds whether or not $x_i$ and $y_i$ are independent. Next consider the treatment producer's profit maximization problem. Conditional on contracting the disease, consumer $i$ would be willing to buy the treatment as long as his or her willingness to pay $y_i$ exceeds the price $p_t$. Because $x_i$ is independent of $y_i$, $i$'s probability of contracting the disease is the mean $x$. Hence demand for the treatment is $x\bar{F}_Y(p_t)$, implying

$$\pi_t = \max_{p_t \in [0,\infty)} [x p_t \bar{F}_Y(p_t)] = x R_Y. \tag{5}$$

Revenue expressions (4) and (5) can be ranked. One of the sources of private information integrates out of (5) and becomes the constant $x$; (4) retains both sources of private information and thus reflects lower revenue. We have the following proposition, proved in the appendix.

**Proposition 9.** *Assume there is heterogeneity in the distribution of disease risk among preventive consumers. If $y_i$ is independent of $x_i$, then $\pi_t > \pi_p$.*

The proposition says that adding independently distributed income heterogeneity cannot reverse the bias against preventives found in Proposition 2 when consumers were heterogeneous in disease risk alone. Although adding independently distributed income heterogeneity cannot reverse the bias against preventives, it will reduce the bias as the next proposition, proved in the appendix, shows.

**Proposition 10.** *Adding income heterogeneity that is distributed independently from the heterogeneity in disease risk causes $\pi_p/\pi_t$ to rise at least weakly (strictly for continuous distributions).*

Next, consider the extreme case of positive correlation, letting $y_i$ be a deterministic function of $x_i$ that is increasing. Ex ante, the two sources of private information compound each other; ex post one of them disappears. Because there is less private information ex post, treatments generate more revenue than preventives as the next proposition, proved in the appendix, states.

**Proposition 11.** *Assume there is heterogeneity in the distribution of disease risk among preventive consumers. If $y_i$ is an increasing function of $x_i$, then $\pi_t > \pi_p$.*

Thus far we have not uncovered a case in which the firm is biased against treatments. Such a case can arise when income and disease risk are negatively correlated. This is easiest to see in the extreme case in which $x_i$ and $y_i$ are inversely proportional: $x_i y_i = z_i = z$ for all $i$. In this case the demand for preventives would be homogeneous across consumers, allowing a preventive monopolist to extract all social welfare—the entire disease burden $D$. A treatment monopolist, on the other hand, cannot fully extract $D$ if there is nontrivial heterogeneity in $y_i$. This leads all the results from Section 3 to flip. Preventives now deliver the first best. As in Proposition 2, the firm is guaranteed to have a bias, only now against treatments. This bias can be quantified and bounded as in Propositions 3–5, can be shown to depend on the skewness in the distribution of $y_i$ as in Proposition 6, and can be connected to inefficient product development as in Proposition 8.[7]

If the firm is able to discriminate on the basis of one of the combined sources of heterogeneity, then the analysis is essentially identical to the case in which there is no heterogeneity in that variable. For example, suppose consumers vary in both $x_i$ and $y_i$ but the firm is able to perfectly price discriminate on the basis of $y_i$. (Firms can accomplish a limited form of this sort of discrimination in an international context by charging different prices across countries differing in their income levels.) The qualitative analysis from Section 3 carries over to this case. Preventive and treatment revenue can be calculated for a given $y_i$ using the conditional distribution $F_{X|Y}(x_i|y_i)$ and then integrated over $y_i$.

## 4.2. Harm Heterogeneity

In this subsection, we analyze of consumer heterogeneity in harm $h_i$ from the disease. We will model this as an ex post source of private information, revealed to the consumer after he or she contracts the disease. Conceptually, we are taking $h_i$ to be a fairly narrow measure of harm, mainly representing the severity of the physical damage caused by the disease. Any aspect of harm that the consumer could predict ex ante (e.g., lost income from a given period of sick leave) is assumed to be embodied in $y_i$. Assume the consumer learns $h_i$ upon contracting the disease but the firm only ever knows the distribution of $h_i$ in the population (or cannot discriminate based on $h_i$ if it observes $h_i$).

Suppose $h_i$ is sole source of consumer heterogeneity. It is immediate that switching the source of private information from $x_i$ ex ante to $h_i$ ex post flips the results from Section 3, just as the results were flipped in the case of extreme negative correlation between disease risk and income ($x_i y_i = z$ for all $i$).

[7] There is no analogue to Proposition 7 because $y_i$ does not have a natural upper bound as does the probability $x_i$.

Next, consider combining heterogeneity in $h_i$ with other sources of heterogeneity. Begin by assuming consumers are heterogeneous in $x_i$ and $h_i$ but have the same income, $y$. Assume first that $x_i$ and $h_i$ are independent. Revenue from a preventive and treatment can be written respectively as

$$\pi_p = \max_{p_p} p_p \bar{F}_X(p_p/h) \tag{6}$$

$$\pi_t = \max_{p_t} x p_t \bar{F}_H(p_t) \tag{7}$$

Applying a straightforward change of variables leads to the following proposition, proved in the appendix.

**Proposition 12.** *Suppose $x_i$ and $h_i$ are distributed independently and that consumers have the same income $y$. The firm earns more revenue from a preventive if and only if $R_X/x > R_H/h$ and from a treatment if and only if $R_H/h > R_X/x$.*

To understand the proposition, recall the definition of $x$ as the mean of the distribution of disease risk: $x = \int_0^{\bar{x}} x_i dF_X(x_i) = \int_0^{\bar{x}} \bar{F}_X(x_i) dx_i$, where the last equality follows from integrating by parts. Because $x$ is the area under "demand curve" $\bar{F}_X$, it represents the potential rent that can be extracted from the market ex ante. Recall the definition of $R_X$ as the area of the largest rectangle that can be inscribed under $\bar{F}_X$, or in other words the greatest rent that can be extracted from the market ex ante with a linear price. Hence $R_X/x$ is the surplus extraction ratio for the ex ante period. Similarly, $R_H/h$ is the surplus extraction ratio for the ex post period. The proposition says that when the sources of heterogeneity are independent, the surplus extraction ratios for the ex ante and ex post periods can be computed in isolation. Whichever of the ratios is greater, the product sold in the associated period generates more revenue. The proposition implies that the bias could go either way in theory.

Next, consider the case in which $h_i$ is an increasing deterministic function of $x_i$. The case turns out to be similar to the one in which consumers are heterogeneous in $x_i$ and $y_i$ and $y_i$ is an increasing function of $x_i$. Arguments similar to the proof of Proposition 9 can be used to show that $\pi_t > \pi_p$. Next, consider the case in which $h_i$ is inversely proportional to $x_i$: $x_i h_i = z$ for all $i$ for some $z$. It is immediate that consumers are homogeneous from an ex ante perspective, and so the whole disease burden $D$ can be extracted with a preventive. Heterogeneity in $h_i$ remains in the ex post period, so treatments will not be able to extract all of $D$. Thus the firm's bias is toward preventives. The results are the same as with heterogeneity in $h_i$ alone.

Moving to the remaining combination of sources of heterogeneity to be analyzed, suppose consumers are heterogeneous in $h_i$ and $y_i$ but have the same disease risk, $x$. As before we will consider three cases: independence, extreme positive correlation, and extreme negative correlation. If $h_i$ and $y_i$ are independent, we have results analogous to Proposition 9 and 10, but with the inequalities flipped because the variable combined with $y_i$ involves ex post rather than ex ante heterogeneity. Thus we have that adding independently distributed heterogeneity in $y_i$ cannot reverse the firm's bias against treatments found with heterogeneity in $h_i$ alone but will reduce the bias.

The remaining cases in which $h_i$ is a deterministic function of $y_i$—whether increasing or inversely proportional—can be analyzed together. Indeed, if $h_i$ is any deterministic function of $y_i$, the result will

**Table 1:** Summary of results for alternative sources of heterogeneity

| | |
|---|---|
| Firm's bias toward treatment | Ambiguous bias |
| *Heterogeneity in $x_i$ alone | Independent variation in $x_i$ and $h_i$ |
| *Independent variation in $x_i$ and $y_i$ | |
| *Perfect positive correlation between $x_i$ and $y_i$ | No bias |
| Perfect positive correlation between $x_i$ and $h_i$ | *Heterogeneity in $y_i$ alone |
| | †Perfect positive correlation between $y_i$ and $h_i$ |
| Firm's bias toward preventive | †Perfect negative correlation between $y_i$ and $h_i$ |
| †Heterogeneity in $h_i$ alone | |
| †Independent variation in $y_i$ and $h_i$ | |
| *Perfect negative correlation between $x_i$ and $y_i$ | |
| Perfect negative correlation between $x_i$ and $h_i$ | |

Notes: The firm is said to be biased toward treatment if $\pi_t > \pi_p$, toward preventive if $\pi_p > \pi_t$, and exhibits no bias if $\pi_p = \pi_t$. Perfect positive correlation refers to the case in which the second variable is a deterministic, increasing function of the first variable. Perfect negative correlation refers to the case in which variables are inversely proportional. *Indicates entries that do not involve heterogeneity in $h_i$, possibly empirically relevant for the HIV example discussed in the text. †Indicates entries that involve heterogeneity in $h_i$ but not $x_i$, possibly empirically relevant for the polio example discussed in the text.

be the same. Because the consumer knows $y_i$ ex ante and ex post, and knowledge of $y_i$ gives knowledge of $h_i$, consumers have the same private information ex ante and ex post. Thus just as the firm has no bias toward either produce with heterogeneity in $y_i$ alone, it will have no bias in this case either, as the next proposition, proved in the appendix, states.

**Proposition 13.** *Assume that consumers have the same disease risk x, that they are heterogeneous in income $y_i$, and that harm $h_i$ is a deterministic function of $y_i$. Then $\pi_p = \pi_t$.*

The results in Proposition 13 do not mirror the analogous results for combined heterogeneity in $x_i$ and $y_i$ when $y_i$ a deterministic function of $x_i$. Whereas there is no bias toward either product under the conditions of Proposition 13, when $y_i$ is a deterministic function of $x_i$ there can be bias; moreover, the direction of bias depends on the slope of the function. The difference can be explained with reference to entropy, a measure of uncertainty from information theory. With combined heterogeneity in $x_i$ and $y_i$, the elimination of private information in $x_i$ upon realization of disease status has a real effect on entropy facing firms. The direction of the effect depends on the slope of the function linking $y_i$ to $x_i$. On the other hand, when $h_i$ is a deterministic function of $y_i$, the realization of $h_i$ ex post does not change entropy because this information was already completely embodied in $y_i$.

### 4.3. Summary

Consideration of various sources of heterogeneity in various combinations with various correlation structures led to a rich set of results. For reference, the results are summarized in Table 1, organized by the conditions generating a bias toward one product or the other.

Further progress can be made in digesting these results by determining which conditions are empirically

relevant for various actual diseases. For example, until the development of antiretrovirals, heterogeneity in $h_i$ may have been less important for HIV than heterogeneity in infection risk. While the time of death varied, HIV virtually always led to AIDS and ultimately death. Thus for HIV those cases that remain after putting aside heterogeneity in $h_i$—the starred entries in Table 1—would be of most empirical relevance.[8]

Focusing on just the starred entries, one can work through a decision tree to further narrow down the relevant theoretical results. One could next ask whether heterogeneity in disease risk is likely to be important. If not, then the theory suggests that bias toward preventives or treatments will not be much cause for concern. If there is substantial heterogeneity in disease risk, (as with HIV) then the next question is the skewness in risk, which will affect the potential for bias. Finally, one can ask about the correlation of disease risk and income in order to estimate the sign and magnitude of bias in R&D incentives between preventives and treatments. Typically, firms will be biased toward treatments and against preventives unless there is substantial negative correlation between risk and income. Indeed, this correlation is negative in the case of HIV, so it is difficult to sign bias between treatments and preventives a priori. In our calibrations for the case of HIV in Section 6, we will first try to measure the importance of heterogeneity in disease risk. Then we will try to determine the direction and importance of the correlation of disease risk and income.

For other diseases, harm could be the most important source of heterogeneity. Polio is one possible example. Before the development of the Salk vaccine in the 1950s, polio epidemics affected a wide swath of the US population (Howard 2005); perhaps the most famous victim was President Franklin Roosevelt, whose legs were paralyzed by polio. Only around 5% of polio infections result in any symptoms. Of the infections resulting in symptoms, most result in a mild, flu-like illness. Only around 10% of the symptomatic infections (0.5% of total infections) result in severe nerve damage such as that suffered by Roosevelt (Mueller, Wimmer, and Cello 2005).[9] The entries in Table 1 that may be empirically relevant for polio, those involving heterogeneity in $h_i$ but not $x_i$, are marked with a dagger. The implied theoretical result for this case is that if there is any bias at all it should be toward preventives, not treatments. Assuming that polio epidemics were widespread, and not strongly correlated with income, the theory would suggest that firms would have stronger R&D incentives for a polio vaccine than for a polio treatment. In fact, a preventive was developed for polio (the Salk vaccine, followed by the Sabin vaccine), but as yet no good pharmaceutical treatments

[8]The case without substantial heterogeneity in $h_i$ may have more empirical relevance than is apparent at first glance. Recall that $h_i$ is a fairly narrow definition of harm, embodying only those elements of harm severity that the consumer cannot predict until contracting the disease. In some cases, harm varies with patient age, weight, or other patient characteristics that patients know ex ante. For other diseases exhibiting substantial harm heterogeneity, patients must be treated before the presentation of severe symptoms to avoid the harm from these symptoms. For example, syphilis eventually leads to blindness in about 15% of untreated cases; however, blindness cannot be reversed by antibiotic treatments for syphilis (Euerle and Chandrasekar 2012). This sort of heterogeneity would not be a source of private information for consumers in either the market for preventives or treatments and thus would not generate a bias toward either product. Further, producers may be better able to discriminate if the heterogeneity is in ex post harm rather than ex ante risk. The producer could offer different versions of the drug, targeting serious cases with a high-priced version with either a high dosage or in a presentation that is suited to be administered in hospitals. In practice, the price differentials can be huge: Lau et al. (2011) found that subject hospital paid from 35 to 240 times more for the intravenous than the pill form for the drugs studied.

[9]Whether this heterogeneity in eventual harm corresponds to heterogeneity in $h_i$ in the sense relevant for the model would depend on whether people have private information on heterogeneity in harm at the time of taking the treatment, and of course we cannot know that since the treatment does not exist.

exist for the disease (Howard 2005). Of course, these outcomes could have been driven by the underlying technological possibility set rather than differences in commercial incentives.

# 5. Alternative Purchasing Arrangements

In our benchmark model, consumers purchase pharmaceuticals directly from the manufacturer. In this section, we extend the model to consider alternative purchasing arrangements. While we still consider highly stylized environments, we argue some of the cases are useful for understanding policy.

Subsection 5.1 considers the case in which the manufacturer can offer an insurance plan for its product. Our main finding is that such ex ante contracts for future product access benefit treatment more than preventive manufacturers because treatment manufacturers gain the additional option of selling either ex ante or ex post, whichever is more profitable.

Subsection 5.2 considers the case in which a third party (HMO, insurer, or small-country government) bargains with the firm over the bulk purchase of the product on behalf of a group of consumers. We first consider the case in which the group of consumers is small enough that the purchaser does not consider the impact on R&D incentives. By negotiating a nonlinear tariff with a pharmaceutical firm, the buyer can reduce the deadweight loss associated with prices above marginal cost for marginal units. In Australia, for example, consumer purchases of most pharmaceuticals are subsidized by the government through the Pharmaceutical Benefits Scheme; the program allows manufacturers the option to negotiate with government officials over a nonlinear tariff via a Deed of Agreement (Australian Government Department of Health and Ageing 2009). We will model the third party in these situations as seeking to maximize consumer surplus and will take the set of consumers covered by the intermediary as exogenous to the structure of the contract with the pharmaceutical firm (thus abstracting from adverse-selection issues). These assumptions fit the Australian policy well; they also fit the case of employer-sponsored insurance plans in which coverage for the disease in question is a small part of the benefits package (abstracting from any agency problems between the ultimate consumer and third-party intermediaries).

Subsection 5.3 analyzes the complementary case of ex ante bargaining by a purchaser that represents enough consumers that it seeks to influence R&D incentives. We will see that a third-party purchaser that covers a large proportion of consumers would like to commit to bargaining before the firm sinks R&D costs because this avoids a hold-up problem associated with ex post bargaining. We argue that the Advisory committee on Immunization Practices in the U.S. and the Pneumococcus Advance Market Commitment internationally many help play this role.

## 5.1. Insurance Contracts

Abstracting for the moment from the role of third-party intermediaries, we begin with a straightforward extension of the model to allow the manufacturer to sell insurance for its products to the consumer rather than selling the product directly. This alternative contractual form has no bearing on preventive sales because

preventives must be sold ex ante so function exactly like insurance in our model. Insurance can only make a difference for treatment sales.

Suppose that a treatment manufacturer can offer a contract for future access to its product to consumers ex ante, before their infection status is realized. We call this contract "insurance" here, but it can be any future contract. The possibility of offering insurance offers the treatment manufacturer the option of imitating the strategy of a preventive manufacturer. Hence the treatment manufacturer is assured of earning at least as much from its product as from a similarly effective preventive. This result holds even in cases identified in Section 4 in which, in the absence of insurance contracts, treatments generated less producer surplus than preventives, cases including private information about ex post harm and perfect negative correlation between ex ante disease risk and income. The results of the previous sections could be reinterpreted as indicating when the manufacturer would choose to sell its treatment—ex ante versus ex post—and when this choice would result in social distortions.[10]

For reference, we conclude the subsection by cataloging the results involving insurance allowing for general forms of heterogeneity. If ex ante consumer valuations are homogeneous, there will be no distortion of R&D incentives: preventives can be used to appropriate the full social benefit even absent insurance; treatments can achieve the same result via insurance contracts. If consumer valuations are heterogeneous ex ante, R&D incentives for preventives will typically be suboptimal as we have discussed at length; insurance has no bearing on these incentives. Incentives will also be suboptimal for treatments if ex post consumer valuations are heterogeneous. Insurance can improve these incentives if selling the treatment ex ante via insurance contracts is more profitable than direct sales of the treatment ex post.

### 5.2. Ex Post Bargaining with Third Parties

Next we turn to bulk purchases on behalf of a group of consumers by a purchaser that does not seek to, or cannot, influence R&D incentives, for example an HMO, insurer, or small-country government. We model this as a Nash bargain between the bulk purchaser and firm over sale of an existing product, with threat points if bargaining breaks down given by the equilibrium with direct-to-consumer sales studied above. We will see that distortions to R&D incentives found previously, even if they are attenuated, persist.

Formally, suppose the firm and buyer engage in Nash bargaining over the sale of product $j$ after the firm has decided which product to develop and has sunk its investment $k_j$ in R&D. Assume that the buyer's objective is to maximize consumer surplus. Consider a general form of Nash bargaining in which $\phi \in (0,1)$ indexes the firm's share of the gains from bargaining. Then the firm's ex ante equilibrium surplus from Nash bargaining is

$$N_j = \Pi_j + \phi[WF_j - \Pi_j - CS_j], \tag{8}$$

[10]For the option to sell ex ante to be guaranteed not to reduce producer surplus requires the assumption that the treatment manufacturer can commit to a price path. It if cannot commit, it may face the problem of a durable-good monopolist conjectured by Coase (1970), whereby anticipated sales to low-value consumers ex post constrains how high prices can be ex ante. An interesting set of asymmetries arises between treatments and preventives in the context of a durable-good monopoly, but we omit reporting them here for space considerations.

i.e., the firm's threat point from selling on the private market ($\Pi_j$) plus its share of the ex post gains from moving to the first-best "pie" ($WF_j$) over and above the sum of parties' threat-point surpluses ($\Pi_j$ for the firm and $CS_j$ for the buyer). (Note that (8) is an ex ante surplus, reflecting the accounting convention of netting out the sunk cost $k_j$ from all the firm's payoff terms.) Substituting $DWL_j = WF_j - WE_j = WF_j - (\Pi_j + CS_j)$ into (8) yields $N_j = \Pi_j + \phi DWL_j$. Thus the firm's objective function with ex post buyer procurement is its objective function with private procurement, $\Pi_j$, plus a term reflecting its share of the deadweight loss eliminated in ex post bargaining. The presence of this second term may mitigate the firm's bias against the product that extracts fewer surpluses on the private market but need not eliminate the bias. The results should generalize to any bargaining structure in which the firm and bulk buyers each capture some share of the deadweight loss under private purchases.

The fact that procurement by a bulk purchaser need not eliminate bias in the firm's incentives is an instance of the familiar hold-up problem (Klein, Crawford, and Alchian 1978). The firm decides which product to develop before negotiating with the government. Recognizing that it does not appropriate all the surplus in bargaining, the firm may distort its decision in order to appropriate more surplus. The literature on the hold-up problem focuses on distortions at the intensive margin of how much to invest; in our setting, the hold-up problem also leads to a distortion at the extensive margin of which product to develop.[11]

### 5.3. Ex Ante Bargaining with Third Parties

Now suppose that the firm and the third-party bulk purchaser bargain in advance of the development of products instead of after. Suppose further that the bulk purchaser represents most consumers, so the revenue it provides forms most of the firm's producer surplus. It is straightforward to see that if the parties engage in Nash bargaining (or another efficient bargaining process), they will now reach an efficient outcome. This outcome can be implemented with a two-part tariff, with the bulk purchaser using a fixed fee to transfer surplus to the manufacturer and then being able to purchase as many units as desired at marginal cost.

In the U.S. the de facto policies of the U.S. Advisory Committee on Immunization Practices (ACIP) provides something close to ex ante price setting. The ACIP analyzes the cost effectiveness of new vaccines, recommending that a vaccine be added to the immunization schedule if its price falls below a threshold which would make it cost effective. While the ACIP's recommendations are not legally binding, they are almost always followed in practice. Firms respond by pricing at this threshold. This policy effectively commits the government to a price-setting procedure that ends up tying the price to the value generated by the vaccine. (See Barder, Kremer, and Levine 2004, chapter 2, for a discussion of these examples.)

Another example of ex ante bargaining is provided by advance market commitment programs for vaccines of the type described by Kremer and Glennerster (2004). A pilot program was implemented for pneumococcal vaccine by a group of donors including the Gates Foundation, countries, and other sponsors. This group committed to help finance purchase of pneumococcal vaccine still in development, covering strains of

[11] Stole and Zwiebel (1996), among others, identify a different extensive-margin distortion resulting from the hold-up problem, in their case a distortion in the firm's technology choice.

the disease common in developing countries at a price targeted to be between unit production cost and the vaccine's social value (see Snyder, Begor, and Berndt 2011 for description and analysis).

In summary, individual bargaining before infection status is realized can avoid some losses associated with ex post heterogeneity in valuation (Subsection 5.1). Bargaining by an agent who maximizes welfare of an exogenously defined group can avoid static losses (Subsection 5.2). If barganing takes place not just before infection status is realized, but before R&D investments are sunk, any R&D distortions caused by heterogeneity among consumers can be avoided (this subsection). In considering various purchasing arrangements, all three subsections maintained the assumption of a monopolist pharmaceutical firm. Appendix B considers the effect of competition among manufacturers who obtained temporary monopoly when they develop new products, showing that the prospect of competition from future generic treatments may further reduce R&D incentives for preventives.

## 6. Calibrations for Sexually Transmitted Infections

The remainder of the paper turns from theory to measurement. We begin with calibrations illustrating how to apply the model to assess whether firms may have biased R&D incentives for a particular disease for which we have information on relevant sources of consumer heterogeneity. We focus on the case of HIV because it is an important disease, we have reasonable proxies for the joint distribution of HIV disease risk and income, and as argued in Section 4.3 HIV may exhibit less harm heterogeneity than some other diseases, allowing us to focus on just disease risk and income heterogeneity for which we have better data.

Section 3 showed that heterogeneity in disease risk alone could lead firms to favor treatments over preventives, while Section 4.1 showed that negative correlation between income and disease risk could potentially lead firms to favor preventives over treatments. Thus a bias towards either preventives or treatments is possible a priori, with the direction and size of the bias depending on the joint distribution of disease risk and income. The focus of this section will be on finding empirical measures of this joint distribution, first for the U.S. market and then for the international market.

In Subsection 6.1 we use individual-level data for the U.S. market to calibrate revenue for an HIV preventive and treatment. Revenue from the preventive is generally much lower than from the treatment, only one quarter to one half as much, providing a potential contributing factor for the continued delay in developing HIV vaccines relative to drugs. These results contrast with additional calibrations for HPV, a much more common disease than HIV and with an infection-risk distribution that is consequently less skewed. The calibrated revenue for an HPV preventive is close to that for a treatment, suggesting that firms may have less bias against developing preventives for HPV.

In Subsection 6.2 we move from U.S. to cross-country data on the joint distribution of HIV risk and income. These calibrations allow us to explore the effect of international price discrimination in the pharmaceutical market on relative incentives to invest in preventives. The calibrations suggest that restricting the scope for international price discrimination could substantially reduce revenue from HIV drugs, possibly

**Table 2:** Preventive/treatment producer surplus ratio in calibrations for the U.S. market

| Survey: | GSS | GSS | NHANES | GSS | GSS |
|---|---|---|---|---|---|
| Income heterogeneity: | No | No | No | Yes | Yes |
| Income elasticity: | — | — | — | 1.0 | 0.4 |
| Ages in sample: | All | 35–40 | All | All | All |
| | (1) | (2) | (3) | (4) | (5) |
| HIV calibrations | | | | | |
| HIV1: Linear model | 0.253 | 0.260 | 0.227 | 0.496 | 0.356 |
| HIV2: Kaplan model, $\beta = 0.06\%$ | 0.251 | 0.265 | 0.246 | 0.504 | 0.363 |
| HIV3: Kaplan model, $\beta$ varies by demographics | 0.375 | 0.402 | 0.371 | 0.571 | 0.461 |
| HPV calibrations | | | | | |
| HPV1: Kaplan model with $\beta = 13.5\%$ | 0.482 | 0.517 | 0.547 | 0.830 | 0.707 |
| Observations | 17,255 | 2,478 | 2,457 | 15,827 | 15,827 |

below that from vaccines.

## 6.1. U.S. Market

The U.S. pharmaceutical market is by far the world's largest and is widely seen as the driver of firms' R&D decisions. Several surveys report information on risk factors for HIV and other sexually transmitted infections such as numbers of sexual partners. We will try several different approaches to mapping the relationship between observed characteristics and disease risk and employ data from two different surveys.

Our first calibrations use nationally representative data on the lifetime number of sexual partners broken down by the individual's gender and sexual orientation and the partners' genders from the 1989–2004 General Social Survey (GSS) to calibrate the model of Section 3.[12] The distribution of lifetime sexual partners is highly positively skewed: the median is 3 but the mean is 10.7. This skewness induces skewness in the distribution of disease risk in our calibrations, which in turn leads to a large gap between the producer surplus from a preventive and treatment.

Column (1) of Table 2 contains the results from calibrations that use GSS data and that account for disease-risk heterogeneity but not income heterogeneity. The calibration labeled HIV1 involves a simple linear mapping from lifetime sexual partners to infection risk with a constant probability of transmission per partner. Figure 4 graphs the resulting inverse demand curve for this calibration. The positively skewed distribution of disease risk produces a highly convex inverse demand curve. Recall $\pi_p$ is given by the area of

[12] We use the cleaned version of the GSS data used in Blanchflower and Oswald (2004) among other studies. Income is based on the family income variable interpolated as the median of the bands or, for top-coded observations, 1.25 times the top code. Other top-code factors produced essentially the same results. Income is converted into 2004 dollars using the Consumer Price Index. We label "lifetime sexual partners" the response to the survey question asking the number of sexual partners since age 18.

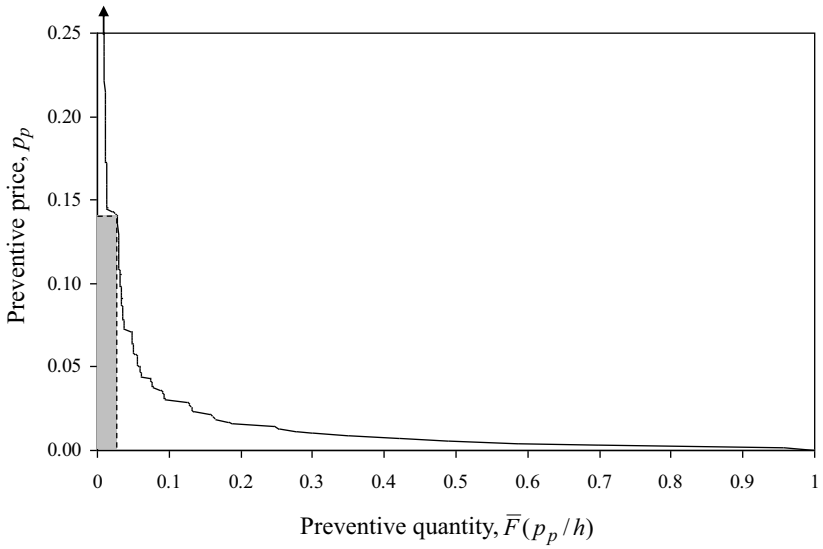


**Figure 4:** Inverse demand curve for calibration in which probability of infection assumed linear in lifetime number of sexual partners. (To aid visualization, the vertical axis has been truncated from $p_p = 1$ to $p_p = 0.25$.)

the largest rectangle that can be inscribed under the curve (the shaded rectangle in the figure) and $\pi_t$ by the area under the curve. It is apparent that $\pi_p$ is much less than $\pi_t$; to be precise, $\pi_p/\pi_t = 0.253$. As shown in the figure, the firm's optimal strategy in this calibration turns out to be to sell the preventive at a high price to a small segment of high-risk individuals.

In the row of calibrations labeled HIV2, we replace the simple linear model with a model due to Kaplan (1990), in which a person with $n$ sexual partners has probability $1-(1-\beta)^n$ of ever contracting the disease, where $\beta$ is the probability of contracting the disease from any given partner. We take $\beta = 0.06\%$, equal to an estimate of the current HIV prevalence rate in the United States, which according to UNAIDS (2004) is 0.6%, times the average per-partner transmission rate, which following Rockstroh *et al.* (1995) we take to be 10%. The estimated figure for $\pi_p/\pi_t$, 0.252, is quite similar to that from the linear model.[13]

In the row of calibrations labeled HIV3, we allow the $\beta$ in the Kaplan model to vary by sexual orienta-

[13]Results are insensitive to varying $\beta$ by one third in either direction.

tion,[14] race,[15] and intravenous (IV) drug use.[16] These are important sources of disease-risk heterogeneity in the population: our estimates suggest that HIV is over 60 times more prevalent among homosexual than heterosexual males, eight times more prevalent among blacks than whites, and over 30 times more prevalent among IV-drug users than others. Although one might expect these additional potential source of heterogeneity to reduce the relative profitability of preventives, in fact $\pi_p/\pi_t$ increases from 0.252 to 0.316 in column (1). The firm ends up concentrating its sales of the preventive among even higher-risk individuals compared to the previous calibration. Although sales fall, the price can be increased enough that the overall profitability for the preventive rises.

Columns (2) and (3) provide robustness checks. Column (2) repeats the calibrations from column (1) for a single age cohort, 35 to 40 year olds. At the cost of a smaller sample size, the calibrations address the potential concern that number of sexual partners may have different meanings for people in different age cohorts because older cohorts have had a longer time to accumulate partners and also lived in environments with different sexual norms. The producer-surplus ratio $\pi_p/\pi_t$ increases slightly across calibrations from column (1) to (2), for example from 0.253 to 0.260 for the linear model. Column (3) repeats the calibrations from column (1) using a different data source for infection risk: the 2003–2004 National Health Examination Survey (Centers for Disease Control 2005), or NHANES. The resulting producer surplus ratios are close to their analogues in column (1).

Column (4) repeats the calibrations from column (1) allowing for heterogeneity in income along with infection risk. The assumption from Section 4.1 are maintained: willingness to pay to avoid harm from the disease is proportional to income ($y_i$) and price discrimination based on $y_i$ is impossible. An individual's demand for a preventive equals his or her disease risk $x_i$ multiplied by $y_i$. Producer surplus from a preventive is calculated as the rectangle of maximum area under this inverse demand curve. The demand curve for a drug is constructed by ordering consumers by $y_i$ and then stepping off the expected drug quantity $x_i$ each

[14]For the male partners of males, we scale $\beta$ up in two stages. We first multiply by 36.8, the estimated prevalence of HIV among homosexual males relative to the general population, computed by taking the percentage of people living with HIV in 2004 who contracted the disease from male-to-male contact—199,085 out of 462,792 cases in the 35 reporting states according to the Centers for Disease Control (2006a)—and dividing by the percentage of homosexual males in the population, estimated to be 1.2% in our GSS data. We further scale $\beta$ by a factor of three to reflect the estimate from Royce *et al.* (1997) that HIV is three times more likely to be passed between males than from males to females. For the rest of the sample, we scale $\beta$ by 0.58, equal to the prevalence of HIV among the population that is not homosexual male relative to the prevalence in the general population (including homosexual males). Given the small number of bisexual males in the GSS sample, 0.2%, the results do not depend on how the transmission rates for their male and female partners are treated (we allow for differential rates) and indeed are similar if bisexual males are omitted from the calculations.

[15]We take the $\beta$ parameters which have been adjusted to reflect variation in infection risk by sexual orientation as described in the previous footnote, and further scale them by 2.55 for African American, 0.324 for whites, and 1.00 for Hispanics, estimated from statistics from the Centers for Disease Control (2006a). Implicit in this scaling is the assumption that an individual matches with partners of the same race.

[16]The GSS does not report IV drug use, so we resort to other data sources. A study of HIV prevalence among IV drug users in U.S. drug treatment centers (Centers for Disease Control 2006b) found that HIV prevalence averaged 18% but varied across cities, ranging from 1% in a Los Angeles to 36% in New York City. Coupled with an estimate of the total number of HIV cases due to IV drug use from Centers for Disease Control (2006a), we can back out the total number of IV drug users in different infection-risk categories and append simulated observations to the GSS data to represent the population of IV drug users. Since we do not have information on income for IV drug users, for the calibration in column (4) we take their income to be the threshold for U.S. Medicaid eligibility (75% of the $9,827 poverty line as of 2004).

consumer would buy at this reservation price. Comparing the results to column (1), we see that accounting for heterogeneity in income about doubles the revenue ratio. Though the bias against preventives is reduced, the calibrations in column (4) still suggest that the producer surplus from treatments is nearly twice that from preventives.

Column (5) is similar to (4) but takes 0.4 as an empirical estimate of income elasticity of the willingness to pay to avoid harm rather than the 1.0 implicitly used in column (4).[17] Not surprisingly given that the income elasticity in column (5) is about midway between the implicit value of zero in column (1) and one in column (4), the revenue ratios in column (5) are about midway between those in columns (1) and (4).[18]

As a counterpoint to the calibrations for HIV, Table 2 adds a set of calibrations for a much more common disease, HPV. These calibrations, labeled HPV1, are directly comparable to the HIV2 calibrations—both are Kaplan models with fixed values of $\beta$—but $\beta$ is increased from 0.06% to 13.5%.[19] The ratio of preventive-to-treatment producer surplus in the HPV1 calibrations is about double that for HIV2 across all five columns. Indeed, the HPV calibrations including income heterogeneity come quite close to 1. With a disease as prevalent as HPV, the disease risk cannot be very positively skewed, putting a bound on the discrepancy between preventive and treatment revenue, as shown in Figure 3.

## 6.2. International Market

Firms currently have considerable ability to price discriminate across countries, but there is an active policy debate on whether this ability should be curtailed—for example, in the contexts of parallel trade for pharmaceuticals within the European Union (Danzon 1998) or re-importation of Canadian pharmaceuticals in the United States (Pecorino 2002). The calibration in this section suggests that the abolition of international price discrimination would substantially reduce the profitability of drugs. The calibration also illustrates the possibility raised in Section 4.1 that the bias against preventives can be reversed if disease risk $x_i$ and willingness to avoid harm (as proxied by income $y_i$) are sufficiently negatively correlated and drug access cannot be sold before disease status is realized.

[17] Getzen (2000) surveys empirical studies of the income elasticity of health expenditures. For purposes of the table, we are interested in the U.S. income elasticity of out-of-pocket expenditures. This is provided by the handful of studies using U.S. micro data from an historical period when most of the population was uninsured. The 0.4 figure, estimated by Anderson, Collette, and Feldman (1960) using 1953 data, is in the middle of the $[0.2, 0.7]$ range from these studies. Micro studies using data from the modern era with more insured consumers find income elasticities near zero. Using such an income elasticity would generate the same results in column (1).

[18] It is also possible to calibrate the impact of government purchases. Consider a Medicaid program that has the firm and government engage in Nash bargaining over the supply of product $j$ to all consumers below a certain income threshold (say 75% of the U.S. poverty line, the threshold for Supplemental Security Income eligibility) and that the firm sells to the rest of the consumers as usual on the private market. The addition of this Medicaid program increases the reported revenue ratio for some calibrations and decreases the ratio for others, but overall the changes are modest. For example, in the HIV3 calibration, the addition of the Medicaid program increases ratio in column (4) from 0.571 to 0.611 and decreases the ratio in column (5) from 0.461 to 0.438.

[19] This value of $\beta$ is computed as the HPV prevalence rate times its transmission rate. Dunne *et al.* (2007) estimated the prevalence among U.S. women of the HPV strains classified as posing a high cervical-cancer risk as 15.2%. Dunne et al. estimated the prevalence of the four strains included in the Gardasil HPV vaccine as 3.4%, but the vaccine also offers cross-protection against other high-risk strains (Ault 2007). Data from Hernandez *et al.* (2008) data imply an HPV transmission rate of 88.8%: of the 18 couples in which one partner had an HPV strain that the other did not at the beginning of their study, 16 ended up transmitting a strain to the other.

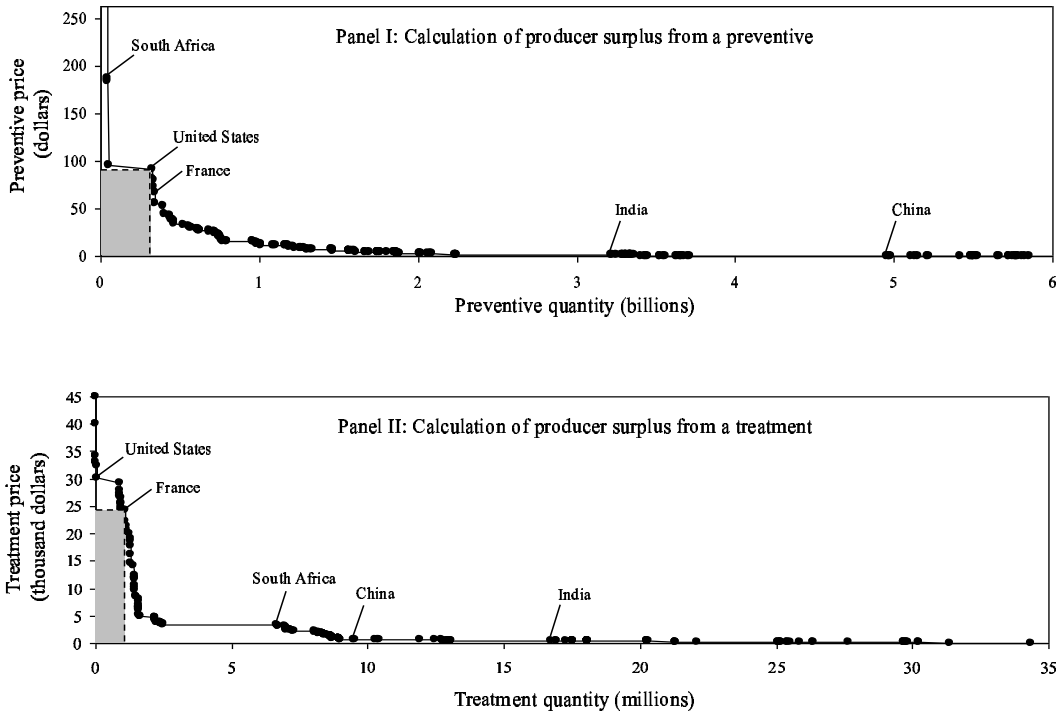


**Figure 5:** Comparison of producer surplus from an HIV preventive to that from a treatment in international example with income heterogeneity and no price discrimination. (Axes scaled so that a unit of area represents the same producer surplus in both panels.)

We consider the market as consisting of the entire world population and treat all individuals within any given country as homogeneous, with the same income and chance of infection; the analysis could be extended to allow for distributions of $x_i$ and $y_i$ within each country. We use country-level data on per-capita GNP, population, and HIV prevalence to approximate our two sources of consumer heterogeneity.[20] We approximate $x_i$ by the fraction of people within a given country that are HIV-positive and $y_i$ by per-capita GNP. The correlation between $x_i$ and $y_i$ across countries is significantly negative at $-0.13$, raising the possibility that $\pi_p > \pi_t$.

Figure 5 shows the inverse demand curve for an HIV preventive in the upper panel and for a treatment in the lower panel. The demand curves are derived as explained in the previous subsection. The firm maximizes preventive profit by charging the price that just induces consumers in the United States to buy and strictly induces consumers in Switzerland, Swaziland, Namibia, the Bahamas, South Africa, and Botswana to purchase the preventive. The profit-maximizing drug price just induces consumers in France to buy and strictly induces consumers in 16 other countries to buy. The axes on the two panels of Figure 5 have been scaled so that a unit of area in both represents the same revenue. The rectangle for the preventive is slightly larger: $\pi_p/\pi_t = 1.13$.[21]

[20]Population data are 1998 data from World Bank (2000); per-capita GNP data are 1998 data calculated with the World Bank Atlas method in 2000 U.S. dollars from World Bank (2000); HIV data are the estimated number of HIV-positive 0-to-49 year olds at the end of 1999 by country from UNAIDS (2000).

[21]As we did in moving from column (4) to (5) in Table 2, we can replace the implicitly assumed value of 1.0 for the income

The analysis suggests that impeding international price discrimination would diminish revenue from an HIV treatment more than from a preventive, and in the extreme could reduce treatment revenue below preventive revenue if treatment access cannot be sold before infection status is realized. Nonetheless, even in the unlikely case of a policy that abolished international price discrimination entirely, there would be an important sense in which the bias against preventives would persist. Although producer surplus from a preventive is 1.13 times that from a treatment in our calibration, at equilibrium prices, social surplus from a preventive is 1.31 times larger than from a treatment, and nearly five times as many lives would be saved from a preventive as from a treatment. This is because it is privately optimal for the firm to target a treatment only to high income countries. The deadweight loss from monopoly pricing is much larger with treatments than preventives. Hence, the firm might develop a treatment even if a preventive would yield greater social surplus and save many more lives.

## 7. Empirical Tests

The calibration in the previous section suggests that in the case of HIV (and other rare STIs), disease-risk heterogeneity may substantially reduce R&D incentives for preventives (even in the presence of substantial negative correlation between income and disease risk). In contrast, the model suggests no reason why disease-risk heterogeneity would affect incentives for development of treatments (controlling for disease prevalence and the joint distribution of disease risk with income and harm).

In this section we present a first-pass empirical test of whether disease-risk heterogeneity affects the probability that vaccines (the preventives we study) and drugs (the treatment we study) have been developed over the last century for a sample of about 100 infectious-disease-causing microorganisms. Since quantitative information on the distribution of disease risk is not systematically available for a cross-section of diseases, we develop several proxies for heterogeneity and positive skewness in disease risk and combine these proxies into a single indicator. Among others, the proxies include sexual transmission and concentration of risk in an identifiable subpopulation or subregion. To the extent that these proxies are imperfect measures of the shape of the disease-risk distribution, the power of our tests will be reduced.

This indicator characterizing the disease-risk distribution is used as a right-hand-side variable in a model of product (vaccine or drug) development. We use a linear probability model to study a 0–1 measure of whether a product has been developed for a disease. The presumption underlying the model is that lucrative products are more likely to be developed. We control for the type of organism causing disease (virus vs. bacterium) because it is believed to be relatively easier technologically to develop vaccines rather than drugs for viral disease and data on organism type is readily available.[22]

elasticity of health expenditures with an empirical estimate. The elasticity of 0.4 used above in the U.S. calibrations (see footnote 17) will not be appropriate in the present international context because estimates of the elasticity across countries are generally much greater than within country. The range of estimates from the handful of cross-country studies surveyed by Getzen (2000) is $[1.2, 1.4]$. Using the 1.3 midpoint of this range, estimated by Newhouse (1977), leads to only a slight change in the value of $\pi_p/\pi_t$ in the calibration, from 1.13 to 1.18.

[22] We limited attention to bacterial and viral diseases because all variation in the availability of products for other types of

**Table 3:** Descriptive statistics

| Variable | Obs. | Mean | Std. dev. | Min. | Max. |
|---|---|---|---|---|---|
| Vaccine development indicator | 91 | 0.29 | 0.45 | 0 | 1 |
| Drug development indicator | 91 | 0.69 | 0.46 | 0 | 1 |
| Infection-risk heterogeneity | 91 | 0.46 | 0.50 | 0 | 1 |
| Childhood onset | 91 | 0.15 | 0.36 | 0 | 1 |
| Viral indicator | 91 | 0.43 | 0.50 | 0 | 1 |
| Prevalence (max. over period) | 51 | 0.52 | 1.11 | 0 | 4.74 |

Of course, many other factors are important determinants of product development, including ease of the science involved, other cost factors, government subsidies, and as discussed in Section 4.2 particular forms of harm heterogeneity. Lacking data on these factors, we will include them in the error term. We see no particular reason to expect these factors to be systematically correlated with our indicator of risk heterogeneity, but of course future research could seek to control for these factors.

The dataset was constructed by a team of research assistants including a senior medical student. A list of disease-causing organisms was taken from Harpavat and Nissim (2001), a widely-used teaching reference that covers the most clinically important organisms. This source provided summary information on type of organism (bacterium, virus, parasite, fungus), available treatments, whether children or adults are disproportionately affected, sexual and insect transmission, etc.[23] This information was only available for a subset of 51 "notifiable" diseases as defined by the Centers for Disease Control (CDC). We collapsed the resulting time series on prevalence for each disease into a single number by taking the maximum prevalence over the time series. [24]

Two issues arise in using the restricted sample of CDC notifiable diseases. First, the restricted sample is considerably smaller than the full sample. Second, it is not a randomly selected sample. Notifiable diseases are significantly more likely to have had some product (vaccine or drug) developed for them than others, presumably because they are associated with some factor that makes them a greater public-health concern

organisms (parasitic, fungal) would be captured by organism fixed effects.

[23]This basic source was supplemented by the microbiology reference Mandell, Bennett, and Dolin (2009). Dates of product development were compiled from Mandell, Bennett, and Dolin (2009), the dates of vaccine development supplemented by public-health websites (Centers for Disease Control 2009, National Network for Immunization Information 2009, Immunization Action Coalition 2009, U.S. Food and Drug Administration 2009) and the dates of drug development by medical histories (Corey, Kürti, and Czakó 2007; Greenwood 2008). Historical data on disease prevalence was taken from the *Morbidity and Mortality Weekly Report* (various dates, spanning 1944–2007).

[24]We use the historical maximum to address the problem that a product's introduction may reduce the disease's prevalence, inducing a correlation between the prevalence variable and the regression error. The maximum captures prevalence in the absence of a drug or vaccine. The results are similar using alternative prevalence measures such as the mean over the period rather than the maximum.

**Table 4:** Impact of infection-risk heterogeneity on product development

| | Full sample (coefficients) | | | Restricted sample (coefficients) | | |
|---|---|---|---|---|---|---|
| Variable | Vaccine developed (1) | Drug developed (2) | Difference (3) = (1) – (2) | Vaccine developed (4) | Drug developed (5) | Difference (6) = (4) – (5) |
| Infection-risk heterogeneity | –0.265*** (0.090) | –0.003 (0.098) | –0.262* (0.145) | –0.400*** (0.136) | –0.044 (0.089) | –0.355** (0.143) |
| Childhood onset | | | | 0.408*** (0.130) | –0.242* (0.122) | 0.650*** (0.130) |
| Viral | | | | 0.204* (0.121) | –0.693*** (0.116) | 0.897*** (0.143) |
| Prevalence (max. over period) | | | | –0.022 (0.025) | 0.011 (0.023) | –0.033 (0.027) |
| Constant | 0.408*** (0.071) | 0.694*** (0.067) | –0.286*** (0.101) | 0.491*** (0.123) | 1.037*** (0.043) | –0.546*** (0.124) |
| $R^2$ | 0.09 | 0.00 | | 0.39 | 0.67 | |
| Observations ($n$) | 91 | 91 | | 51 | 51 | |

Notes: Ordinary least squares regressions in which dependent variable is an indicator for development of product. Bacterial is omitted organism category in the restricted-sample regressions. White (1984) heteroskedasticity-robust standard errors reported in parentheses. Significantly different from 0 in a two-tailed test at the *10% level, **5% level, ***1% level.

(greater prevalence, harm, or transmissibility). Our strategy will be to focus on the results from the full sample but also report results for the restricted sample for robustness.

Table 3 provides descriptive statistics for the dataset. The indicator for infection-risk heterogeneity deserves special comment because it is the regressor of central interest. This indicator is set to 1 if a discrete high-risk group could readily be defined from a review of the disease's epidemiology and transmission patterns. Specifically, the indicator is set to 1 if the disease satisfies at least one of the following conditions:

- sexually transmitted;
- transmitted by animal contact;
- chiefly affects a small population of either hospitalized patients, immuno-compromised individuals, intravenous-drug users, or soldiers;
- organism has restricted ecological habitat (e.g., tropics for malaria).

Table 4 reports the results from a linear probability model, which regresses an indicator for product (vaccine or drug) availability on infection-risk heterogeneity using ordinary least squares. Results from alternative specifications (probit, logit) are quite similar. Consider the results from the full sample in columns (1)–(3) in which infection-risk heterogeneity is the only covariate. The –0.265 coefficient in the first row of column (1) indicates that vaccines are 26.5 percentage points less likely to have been developed for diseases

with infection-risk heterogeneity, significant at the 1% level. The analogous coefficient in column (2) indicates that there is no statistically significant effect of infection-risk heterogeneity on drug development. The difference between the vaccine and drug coefficients in column (3) indicates that infection-risk heterogeneity reduces vaccine development 26.2 percentage points more than it does drug development, a difference significant at the 10% level.

The difference between the constant terms in column (3) indicates that vaccines are less common than drugs, the average disease being 28.6 percentage points less likely to have a vaccine than a drug, significant at the 1% level. This result may capture a host of factors besides heterogeneity in infection risk that may make vaccines harder to market than drugs, such as tendencies for people to invest less on prevention or the greater epidemiological externalities from vaccines.

One concern with results is that our infection-risk heterogeneity may be proxying for more than just the shape of the risk distribution; it may be proxying for low overall disease burden, as diseases that are transmitted through specialized vectors or concentrated in subpopulations may have an overall low prevalence. Virtually any theory would suggest that firms would have less of an incentive to develop products for low-burden diseases, and so a significantly negative coefficient on our proxy may not be a dispositive test of the particular theory in Section 3. This concern is partially addressed in the specification involving the full sample by focusing not on the negative coefficient in the vaccine regression in isolation but on a comparison of the vaccine to the drug regression. If infection-risk heterogeneity were proxying for low overall disease burden, one would expect to find a negative effect on drug development as well, but the coefficient on infection-risk heterogeneity in column (2) is close to 0. The result in column (3), which can be viewed as a difference-in-differences, indicates that our proxy is having a statistically significantly different effect on vaccine than on drug development.

The concern is further addressed by the specification involving the restricted sample, reported in columns (4)–(6), adding an explicit prevalence measure as well as other controls. The sample is restricted to the subset of 51 notifiable diseases for which we have prevalence data. The results are if anything a bit stronger than in the full sample, with infection-risk heterogeneity decreasing the probability of vaccine development by a statistically significant 40.0 percentage points, but having essentially no effect on drug development, resulting in a differential effect on vaccines vs. drugs reported in column (6) of 35.5 percentage points, now significant at the 5% level.

The additional controls in column (4)–(6) regressions are of some independent interest. Vaccines are significantly more likely to be developed for diseases that disproportionately affect children and drugs significantly less likely. This is consistent with the lower cost of delivery of vaccines that can be integrated into childhood immunization programs. Viral diseases show the same pattern consistent with the widespread view among scientists, that the technology of vaccine production is particularly suitable for viruses. The prevalence measure does not show up as important in any regression. One explanation is that the restricted sample, including as it does only diseases listed as notifiable by the Centers for Disease Control, already selects for diseases with high aggregate health burden, so that within this group, and prevalence and harm

may be negatively correlated across diseases.[25]

We also tested the hypothesis by running a Cox proportional hazards model on the date (from 1945 to present) of product development. In the specification using the full sample, infection-risk heterogeneity cuts the hazard of vaccine development by more than two-thirds but does not reduce the hazard of drug development. The reduction in vaccine hazard is significantly larger than the reduction in the drug hazard at the 5% level. In the specification restricting the sample to CDC notifiable diseases with information on prevalence and other variables (using the full panel of prevalence data for that variable), infection-risk heterogeneity leads to a similar reduction in the hazard of vaccine development as in the specification involving the full sample. There is now also some evidence it reduces the hazard of drug development (at the 10% level). The point estimate implies that infection-risk heterogeneity reduces the vaccine hazard by twice as much as the drug hazard. Due to large standard errors in the specification using the restricted sample, the reduction in the vaccine hazard is not significantly different from the reduction in drug hazard. Overall, the results from both the linear-probability and hazard models are consistent with the idea that infection-risk heterogeneity reduces incentives to develop vaccines.

## 8. Conclusion

In this paper, we argued that time-varying sources of private information for consumers, combined with differences in the timing of when preventives and treatments are administered, may affect firms' ability to extract consumer surplus under direct sales to consumers. Thus the wedge between private and social R&D incentives will be different for preventives than for treatments.

We focus on a benchmark model in which a monopolist sells directly to consumers, but also consider extensions to other environments. If consumers vary only in their disease risk, a monopolist can extract less revenue from preventives—which are sold when consumers still have private information about their disease risk—than from treatments which are sold after consumers' disease status is realized, at which point there is no heterogeneity among those with positive valuation. We showed that the firm's bias toward treatments is likely to be largest for diseases with a right-skewed risk distribution, i.e., diseases with much of the risk concentrated in a small segment of the population. Thus biases against vaccines are more likely for sexually transmitted infections than infections with airborne transmission, for example.

We then broadened the analysis to encompass other sources of consumer heterogeneity with different timing structures. If consumers are initially homogeneous but learn about how severely they are affected by the disease only after contracting it ex post, then the benchmark results are reversed. Treatment manufacturers will not be able to fully extract consumer surplus, but preventive manufacturers will.

Allowing firms to sell insurance contracts for their products creates a potentially valuable option for a

[25] Consistent with this explanation, we ran a regression similar to the specification in columns (4)–(6) but using the full sample of 91 observations and replacing the prevalence variable with an indicator for CDC-notifiable diseases. This indicator was quite often large, positive and statistically significant. We prefer the specification reported in columns because it involves a more homogeneous set of diseases and because the omitted CDC-notifiability indicator may be endogenous, in particular if the CDC is more likely to require notification for diseases that are part of immunization programs.

treatment manufacturer, which can choose to sell treatment insurance ex ante (before the disease is contracted) or continue just selling the treatment ex post. The option is worthless for a preventive manufacturer, whose product already functions like insurance because it is administered ex ante.

A rich set of additional results came from analyzing various combinations of sources of consumer heterogeneity. For example, allowing consumers to vary in income in addition to disease risk may reverse the benchmark result that treatment exceeds preventive revenue found when consumers were heterogeneous just in disease risk. This reversal only obtains under certain conditions: the correlation between income and disease risk must be sufficiently negative; the firm cannot be able to price discriminate on the basis of income or to offer insurance contracts for treatments sold in advance of consumers learning their disease status.

Bargaining by bulk purchasers, such as HMOs, insurers, or governments, can address static monopoly pricing distortions, but if this bargaining occurs after product development, bulk purchases reduce but do not eliminate gaps between private and social incentives for product development. The manufacturer cares about the outcome on the private market because this is its threat point in negotiations with the bulk purchaser. In contrast ex ante price setting, as in the de facto operation of the Advisory Committee on Immunization, practices in the U.S., or under Advance Market Committments internationally, could provide optimal R&D incentives.

As discussed in Appendix B, an extension incorporating competition between a preventive and a treatment as well as later generic entrants suggests an additional bias against preventives. Future entry of generic treatments constrains the pricing of preventives, but treatment pricing is unaffected by competition from preventives.

A calibration using estimates of the joint distribution of income and HIV risk in the United States suggests that an unconstrained monopolist would find it optimal to sell even a costless vaccine with no side effects at a high price to a small fraction of the population, and would earn only about half the revenue obtained by a treatment manufacturer. In contrast, for HPV, vaccine revenue would almost equal drug revenue. The difference is that HIV is rare enough that the skewness in number of sexual partners generates skewness in HIV infection risk while HPV is so prevalent that it is mathematically impossible for HPV infection risk to exhibit much skewness. Although many other factors are involved, this may be a contributing factor (i) for why a preventative was developed for HPV more quickly than for HIV; (ii) for why when a preventative was developed for HIV, it was based on an existing treatment, and (iii) for why there was an eight year lag between the time the manufacturer completed the FDA approval process for Truvada as treatment and as a preventative.[26] Truvada is expected to continue selling at a high price and to be used as a preventive only by a small segment of people with extreme HIV risk (Grady 2012); its use as a preventive was thus not regarded by the Gilead CEO as a "significant commercial opportunity" (Fey Cortez and Bennett 2011).

Calibrations for HIV revenue in the international market suggest that eliminating price discrimination

[26] An HPV vaccine, Gardasil, was approved by the U.S. Food and Drug Administration (FDA) in 2006 but an HIV preventive was not approved until 2012. Based on the trials conducted by the manufacturer, the FDA approved Truvada first as an HIV treatment in 2004 but did not approve its use in a daily regimen to protect healthy individuals from HIV infection until 2012.

across countries would substantially reduce incentives to develop HIV drugs but would have much less effect on incentives for vaccine development.

As an empirical test of the model, using a novel dataset on infectious diseases, we regressed indicators for whether drugs or vaccines have been developed on an indicator for heterogeneity in disease risk, which we constructed from underlying proxies, along with other controls. In line with the basic theory, we found vaccines are significantly less likely to have been developed for diseases with heterogeneity in infection risk, such as STIs, but we found no similar effect for drugs.

One important topic for future work is extending the benchmark model to examine the ways in which these effects play out under the range of realistic institutional features of health-care markets, such as employer-sponsored health care plans, government subsidies, rules allowing patients to consume pharmaceuticals only on doctors' advice/prescription, technological advances improving the flow of information or the accuracy of testing, etc. Related to these latter issues, our model raises the paradoxical possibility that improvements in consumer medical information—whether from more physician attention, advertising and public-health campaigns, freer access to medical information over the internet, or advances in testing technologies—may increase heterogeneity in perceived disease risk and thus reduce firms' incentives to develop preventives, potentially reducing welfare. For example, consider the case of new genetic screens for breast cancer. While the lifetime risk of breast cancer is about 12% in the U.S. population of women, the risk rises to 60% among those for whom genetic testing reveals a harmful mutation in the BRCA1 or BRCA2 genes (National Cancer Institute 2009), and correspindingly falls among those found not to have this mutation. It is well understood that genetic testing can exacerbate adverse selection in insurance markets (see, e.g., Oster et al. 2010, which documents increased take-up of long-term care insurance for individuals who test positive for the Huntington-disease gene). Our model suggests another channel for genetic testing to reduce welfare, possibly reducing firms' incentives to invest preventives by increasing perceived heterogeneity in diesease risk.

This paper has suggested one factor (disease-risk heterogeneity) that reduces incentives to develop HIV vaccines. A companion paper (Kremer, Snyder, and Williams, 2012) builds an integrated economic and epidemiological model to investigate a different factor: by reducing disease transmission, vaccines have a positive externality not present with existing treatments. The companion paper finds that the externality is particular large for rare diseases. Because firms do not appropriate the externality, it ends up reducing their profits and R&D incentives. Holding constant the total burden of disease, firms will find developing preventives for common but less serious diseases like the flu more profitable than for rarer but more deadly diseases. Since HIV is rare in the high-income countries that account for the bulk of pharmaceutical revenue, the model suggests that firms will be able to capture a greater fraction of the social value of HIV drugs than of HIV vaccines. Hence both the present and companion paper suggest incentives for R&D on HIV vaccines may be suboptimal.

The market distortions against vaccine development we discuss could potentially be corrected through subsidies to HIV vaccine R&D beyond those for pharmaceutical R&D in general (as under the International

AIDS vaccine initiatives), or through Advance Market Commitments to purchase vaccines if they are developed (Kremer and Glennerster 2004). To the extent that policymakers are uncertain about the scientific feasibility of producing an HIV vaccine, and potential vaccine developers have private information about the probability of success, Advance Market Commitments may be particularly attractive.

# Appendix A: Proofs of Propositions

**Proof of Proposition 2:** Substituting $\pi_p = \Pi_p + k_p$ and

$$\bar{F}(p_p/h) = \int_{p_p/h}^{1} dF(x_i)$$

into equation (1) and making the change of variables $\hat{x} = p_p h$ yields $\pi_p = h\int_{\hat{x}^*}^{1} \hat{x}^* \, dF(x_i)$, where

$$\hat{x}^* = \underset{\hat{x}\in[0,1]}{\operatorname{argmax}} \left[ h \int_{\hat{x}}^{1} \hat{x} dF(x_i) \right]. \tag{A1}$$

Substituting $\pi_t = \Pi_t + k_t$ and $x = \int_0^1 x_i dF(x_i)$ into equation (2) yields $\pi_t = h\int_0^1 x_i dF(x_i)$. Thus,

$$\begin{aligned} &\pi_t - \pi_p \\ &= h\int_0^1 x_i dF(x_i) - h\int_{\hat{x}^*}^{1} \hat{x}^* \, dF(x_i) && \text{(A2)} \\ &= h\int_0^{\hat{x}^*} x_i dF(x_i) + h\int_{\hat{x}^*}^{1} (x_i - \hat{x}^*) \, dF(x_i). && \text{(A3)} \end{aligned}$$

Both terms in (A3) are nonnegative. There cannot be a measure one of consumers at $\hat{x}^*$ by maintained assumption. Thus, there must be a positive measure on either a subset of $(0, \hat{x}^*)$, in which case the first term in (A3) is positive, or on a subset of $(\hat{x}^*, 1]$, in which case the last term in (A3) is positive. In either case, $\pi_t - \pi_p > 0$. *Q.E.D.*

**Proof of Proposition 3:** We have

$$\begin{aligned} &\sup\left(\frac{WF - WE}{D}\right) \\ &= \max_{j,\ell\in\{p,t\}} \left\{ \sup\left[ \left(\frac{WF_\ell - WE_j}{D}\right) \right.\right. \\ &\qquad \left.\left. \times \mathbf{1}(\Pi_j = \max(\Pi_p, \Pi_t)) \right] \right\} && \text{(A4)} \\ &= \max\left\{ \sup\left[ \left(\frac{WF_p - WE_p}{D}\right) \mathbf{1}(\Pi_p \geq \Pi_t) \right], \right. \\ &\qquad \left. \sup\left[ \left(\frac{WF_p - WE_t}{D}\right) \mathbf{1}(\Pi_t \geq \Pi_p) \right] \right\}, && \text{(A5)} \end{aligned}$$

where $\mathbf{1}(\cdot)$ is the indicator function and where the suprema are all taken over parameters $(k_p, k_t) \in [0,\infty)^2$. Equation (A4) holds by definition of *WF* and *WE*. To see (A5), note that if a treatment is developed in the first best, then $WE_t = D - k_t = WF_t = WF \geq WE_p$. Thus if $\ell = t$, then $j = t$ as well. But then $WF_t - WE_t = 0$, implying that the term in braces in (A4) equals zero for $\ell = t$. We will see below that the term in braces in (A4) is non-negative for $\ell = p$, so we can restrict attention to maximizing the term in braces in (A4) over $\ell = p$, which leaves the two possible terms in braces in (A5). Manipulating the first braced term from (A5):

$$\begin{aligned} &\sup\left[ \left(\frac{WF_p - WE_p}{D}\right) \mathbf{1}(\Pi_p \geq \Pi_t) \right] \\ &\leq \sup\left(\frac{WF_p - WE_p}{D}\right) && \text{(A6)} \\ &= \sup\left[ \frac{(D - k_p) - (\pi_p + CS_p - k_p)}{D} \right] && \text{(A7)} \\ &= 1 - \frac{\pi_p}{\pi_t} - \frac{CS_p}{\pi_t}. && \text{(A8)} \end{aligned}$$

Condition (A6) follows from $\mathbf{1}(\Pi_p - \Pi_t) \leq 1$, (A7) from the definitions of $WF_p$ and $WE_p$, and (A8) from simple algebra. Manipulating the second braced term from equation (A5):

$$\begin{aligned} &\sup\left[ \left(\frac{WF_p - WE_t}{D}\right) \mathbf{1}(\Pi_t \geq \Pi_p) \right] \\ &= \sup\left[ \left(\frac{k_t - k_p}{D}\right) \mathbf{1}(\pi_t - k_t \geq \pi_p - k_p) \right] && \text{(A9)} \\ &= \frac{\pi_t - \pi_p}{D} && \text{(A10)} \\ &= 1 - \frac{\pi_p}{\pi_t}. && \text{(A11)} \end{aligned}$$

Equation (A9) holds by substituting the definitions of $WF_p$, $WE_t$, $\Pi_t$, and $\Pi_p$ and simplifying. Equation (A10) holds by noting that the greatest value of $k_t - k_p$ subject to the constraint $\pi_t - \pi_p \geq k_t - k_p$ equals $\pi_t - \pi_p$. Equation (A11) follows from dividing numerator and denominator through by $\pi_t$ and noting $D/\pi_t = 1$ since the firm can extract 100% of social welfare with a treatment so that $\pi_t = D$. Since $CS_p \geq 0$, (A11) at least weakly exceeds (A8). Equation (A11) is non-negative by Proposition 2. Hence (A5) equals (A11). *Q.E.D.*

**Proof of Proposition 4:** A distribution of consumers into $C$ risk classes involves $2C$ parameters $\{m_c\}_{c=1}^C$ and $\{x_c\}_{c=1}^C$ satisfying the following feasibility conditions:

$$\begin{aligned} m_c &\in (0,1) \text{ for all } c = 1, \ldots, C, && \text{(A12)} \\ \sum_{c=1}^{C} m_c &= 1, && \text{(A13)} \\ 0 \leq x_1 &\leq \cdots \leq x_C \leq 1. && \text{(A14)} \end{aligned}$$

We will choose these $2C$ parameters so that $\pi_p/\pi_t$ is very close to $1/C$. We will do this by having the risk-class masses $\{m_c\}_{c=1}^C$ decline geometrically and arranging the risk-class probabilities $\{x_c\}_{c=1}^C$ so that the firm is indifferent between serving all consumers with a low price for the preventive than serving a smaller group with higher prices.

Let $\theta \in (0, 1/2)$. Define risk-class masses

$$m_c = \begin{cases} \theta^{c-1} & \text{if } c > 1 \\ 1 - \sum_{c=1}^{C-1} \theta^c & \text{if } c = 1. \end{cases} \tag{A15}$$

It can be shown that this geometrically declining sequence respects constraints (A12) and (A13). We define the risk-class probabilities recursively as follows: set $x_C = 1$, and set

$$h x_c \sum_{i=c}^{C} m_i = h x_{c+1} \sum_{i=c+1}^{C} m_i. \tag{A16}$$

for $c = 1, \ldots, C-1$. The left-hand side of (A16) is the profit from charging a price $hx_c$ and selling the preventive to risk classes $c$ and higher. The right-hand side is the profit from charging a price $hx_{c+1}$ and selling to risk classes $c+1$ and higher. It is easy

to see that the risk-class probabilities respect constraint (A14). From equation (2), we have $\pi_t = \sum_{c=1}^{C} hm_c x_c$. By construction implicit in (A16), we have $\pi_p = hx_1$; that is, it is weakly most profitable to charge $hx_1$ for the preventive and sell to all consumers. Thus,

$$\frac{\pi_t}{\pi_p} = \frac{\sum_{c=1}^{C} hm_c x_c}{hx_1} \tag{A17}$$

$$= m_1 + \sum_{c=2}^{C} \frac{m_c}{m_c + \cdots + m_C} \tag{A18}$$

$$= 1 - \sum_{c=1}^{C-1} \theta^c + \sum_{c=2}^{C} \frac{\theta^{c-1}}{\theta^{c-1} + \cdots + \theta^{C-1}}. \tag{A19}$$

Equation (A17) follows from previous arguments. Equation (A18) holds since it is equally profitable to sell the preventive to all consumers at price $hx_1$ or to consumers in risk classes $c$ and above at price $hx_c$, so that $hx_1 = hx_c(m_c + \cdots + m_C)$, implying $x_c = x_1/(m_c + \cdots + m_C)$. Equation (A19) holds by substituting for $\{m_c\}_{c=1}^{C}$ from equation (A15). Taking limits, $\lim_{\theta \to 0}(\pi_t/\pi_p) = 1 - 0 + \sum_{c=2}^{C} 1 = C$, or, equivalently, $\lim_{\theta \to 0}(\pi_p/\pi_t) = 1/C$. This shows that for any $\epsilon > 0$, and for the definitions of the parameters in (A15) and (A16), we can find $\theta > 0$ such that $\pi_p/\pi_t < 1/C + \epsilon$. To prove $\pi_p/\pi_t \geq 1/C$ for all distributions of consumers into $C$ risk classes, note

$$\begin{aligned} C\pi_p &= C \max_{c \in \{1,\ldots,C\}} \left[ hx_c \left( 1 - \sum_{i=1}^{c-1} m_i \right) \right] \\ &\geq C \max_{c \in \{1,\ldots,C\}} \{hx_c m_c\} \\ &\geq \sum_{c=1}^{C} hx_c m_c \\ &= \pi_t. \end{aligned}$$

Hence $\pi_p/\pi_t \geq 1/C$. *Q.E.D.*

**Proof of Proposition 7:** Let $B$ be the value of the following minimization problem, labeled MIN1:

$$\min_{\bar{F}} \left\{ \frac{\pi_p}{\pi_t} \right\} \tag{A20}$$

subject to

$$x \geq m, \tag{A21}$$

where $m$ is some constant in $[0,1]$ and where the minimization is taken over the set of all functions $\bar{F}$ satisfying the following three conditions:

$$\bar{F}(0) = 1, \tag{A22}$$

$$\bar{F}(x_i) \in [0,1] \text{ for all } x_i \in [0,1], \tag{A23}$$

$$\bar{F}(x_i) \text{ is non-increasing.} \tag{A24}$$

$B$ provides a tight lower bound on $\pi_p/\pi_t$ for a disease with a prevalence rate of at least $m \in [0,1]$.

We next establish several facts that will allow us to transform MIN1 into an equivalent minimization problem. First, integrating by parts shows

$$x = \int_0^1 x_i \, dF(x_i) = \int_0^1 \bar{F}(x_i) \, dx_i. \tag{A25}$$

Second, we can show constraint (A21) binds. To do so, note that as the constraint is relaxed, the solution to MIN1 approaches 0 by Proposition 5. But $\pi_p/\pi_t$ approaches 0 for finite $\pi_t$ only if $\pi_p$ approaches 0. Furthermore, $\pi_p$ approaches 0 if and only if $x$ approaches 0, violating constraint (A21). Third, having established (A21) binds, we have $\pi_t = hx = hm$. Fourth, $\pi_p = h \max_{x \in [0,1]} \{x\bar{F}(x)\}$. Substituting these four facts into MIN1 gives the equivalent problem, labeled MIN2:

$$\frac{1}{m} \min_{\bar{F}} \left\{ \max_{x \in [0,1]} x\bar{F}(x) \right\} \tag{A26}$$

subject to

$$\int_0^1 \bar{F}(x) \, dx \geq m, \tag{A27}$$

where the minimization is again taken over the set of all functions $\bar{F}$ satisfying (A22)–(A24).

We proceed to solve MIN2. Let $\bar{F}^*(x)$ be any solution to MIN2, and let $x^* = \text{argmax}_{x \in [0,1]} \{x\bar{F}^*(x)\}$. Because $x^*$ is a maximizer, $x\bar{F}^*(x) \leq x^*\bar{F}^*(x^*)$ for all $x \in [0,1]$. Because $\bar{F}^*(x)$ is a solution to MIN2 and thus MIN1, it must generate a value of $B$ in objective function (A26), which upon rearranging implies $x^*\bar{F}^*(x^*) = Bm$. Combining these equalities with condition (A23) implies, for all $x \in [0,1]$,

$$\bar{F}^*(x) \leq \min\{1, Bm/x\}. \tag{A28}$$

Consider the function $\bar{F}^{**}(x)$ given by the right-hand side of (A28), i.e., $\bar{F}^{**}(x) = \min\{1, Bm/x\}$. It can be verified that $\bar{F}^{**}$ yields $B$ as the value of the objective function (A26), that it respects constraint (A27), and that it satisfies conditions (A22)–(A24). Hence $\bar{F}^{**}$ must also be a solution to MIN2.

We argued that the constraint (A21) binds, implying that the equivalent constraint (A27) must also bind. Substituting $\bar{F}^{**}$ into (A27) treated as an equality yields

$$\int_0^1 \min\{1, Bm/x\} \, dx = m, \tag{A29}$$

which after integrating yields

$$Bm[1 - \ln(Bm)] = m. \tag{A30}$$

Canceling terms and substituting $x = m$ from binding constraint (A21) gives the expression for $B$ in (3). *Q.E.D.*

**Proof of Proposition 8:** For a treatment, $\Pi_t = WE_t = WF_t$. Since the firm extracts all social surplus with a treatment, the firm always develops a treatment if it is socially efficient (by either social-welfare measure $WE_t$ or $WF_t$) to do so.

For a case in which $WE_p > WE_t$ but $\Pi_t > \Pi_p$, suppose $x_i$ is uniformly distributed on $[0,1]$; $k_j = 1/8$ for $j \in \{p,t\}$; $c_j = s_j = 0$ for $j \in \{p,t\}$; $h = 1$; $e_p = 1$; and $e_t = 5/8$. For a treatment, we have $\Pi_t = e_t x - k_t = (5/8)(1/2) - 1/8 = 3/16 = WE_t = WF_t$.

For a preventive,

$$\begin{aligned}\Pi_p &= \max_{p\in[0,\infty)} \left\{p_p \bar{F}(\hat{x}(p_p))\right\} - k_p \\ &= \max_{p\in[0,\infty)} \left\{p_p(1-p_p)\right\} - k_p \\ &= 1/8;\end{aligned}$$

$p_p^* = 1/2$; $WE_p = \int_{p_p^*}^1 x_i dx_i - k_p = 3/8 - 1/8 = 1/4$; $WF_p = x - k_p = 1/2 - 1/8 = 3/8$. Thus, $\Pi_t = 3/16 > 2/16 = \Pi_p$, but $WE_p = 4/16 > 3/16 = WE_t$, and $WF_p = 6/16 > 3/16 = WF_t$. *Q.E.D.*

**Proof of Proposition 9:** Suppose $y_i$ is independent of $x_i$. Then $\pi_p$ equals

$$\max_{p\in[0,\infty)} \left\{\int_{p/\bar{y}}^1 \left[\int_{p/x_i}^{\bar{y}} p\, dF_Y(y_i)\right] dF_X(x_i)\right\} \tag{A31}$$

$$\leq \max_{p\in[0,\infty)} \left\{\int_{p/\bar{y}}^1 \max\left[0, \int_{p/x_i}^{\bar{y}} p\, dF_Y(y_i)\right] dF_X(x_i)\right\} \tag{A32}$$

$$\leq \max_{p\in[0,\infty)} \left\{\int_0^1 \max\left[0, \int_{p/x_i}^{\bar{y}} p\, dF_Y(y_i)\right] dF_X(x_i)\right\} \tag{A33}$$

$$\leq \int_0^1 \left\{\max_{p\in[0,\infty)} \left\{\max\left[0, \int_{p/x_i}^{\bar{y}} p\, dF_Y(y_i)\right]\right\}\right\} dF_X(x_i) \tag{A34}$$

$$= \int_0^1 \left\{\max_{p\in[0,\infty)} \left[\int_{p/x_i}^{\bar{y}} p\, dF_Y(y_i)\right]\right\} dF_X(x_i) \tag{A35}$$

$$= \int_0^1 \left\{\max_{p'\in[0,\infty)} \left[\int_{p'}^{\bar{y}} p' x_i\, dF_Y(y_i)\right]\right\} dF_X(x_i) \tag{A36}$$

$$= x \max_{p'\in[0,\infty)} \left[p' \bar{F}_Y(p')\right] \tag{A37}$$

$$= \pi_t. \tag{A38}$$

Equations (A31) and (A37) hold by applying the independence condition to the formulae (4) and (5) and noting $\pi_j = \Pi_j + k_j$, $j \in \{p,t\}$. The rest of the steps are algebraic manipulations. The inequality in (A34) is strict if there is nontrivial heterogeneity in the distribution of $x_i$. *Q.E.D.*

**Proof of Proposition 10:** Let $\pi_p$ and $\pi_t$ be producer surpluses in the model with no income heterogeneity and $\pi'_p$ and $\pi'_t$ be producer surpluses when income heterogeneity which is independently distributed from disease-risk heterogeneity has been added to the model. Then $\pi'_p$ equals

$$p_z^* \Pr(z_i \geq p_z^*) \tag{A39}$$

$$\geq p_x^* p_y^* \Pr(x_i y_i \geq p_x^* p_y^*) \tag{A40}$$

$$\geq p_x^* p_y^* \Pr(x_i \geq p_x^*) \Pr(y_i \geq p_y^*) \tag{A41}$$

$$= \pi_p \pi'_t / \pi_t, \tag{A42}$$

where

$$\begin{aligned}p_x^* &= \operatorname*{argmax}_p [p \Pr(x_i \geq p)] \\ p_y^* &= \operatorname*{argmax}_p [p \Pr(y_i \geq p)] \\ p_z^* &= \operatorname*{argmax}_p [p \Pr(z_i \geq p)].\end{aligned}$$

Equation (A39) follows from equation (4). Condition (A40) follows because $p_x^*$, as an argmax, produces a higher value for $p \Pr(z_i \geq p)$ than $p_x^* p_y^*$. Condition (A41) follows since $\Pr(x_i y_i \geq p_x^* p_y^*) \geq \Pr(x_i \geq p_x^*) \Pr(y_i \geq p_y^*)$. Equation (A42) follows because $\pi_p = p_x^* \Pr(x_i \geq p_x^*)$ by equation (1), $\pi_t = hx$ by equation (2), and $\pi'_t = x p_y^* \Pr(y_i \geq p_y^*)$ applying the independence assumption to equation (5). Conditions (A39)–(A42) together imply $\pi_p/\pi_t \leq \pi'_p/\pi'_t$. If the distributions of $x_i$ and $y_i$ are continuous, then the inequality in (A41) would be strict. *Q.E.D.*

**Proof of Proposition 11:** Suppose $y_i = g(x_i)$, where $g$ is some increasing function. Let $p_p^*$ be the optimal preventive price. Preventive demand equals $\bar{F}_Z(p_p^*) = \bar{F}_Y(\hat{y})$ for $\hat{y}$ given by the solution to $g^{-1}(\hat{y})\hat{y} = p_p^*$. Hence

$$\pi_p = p_p^* \bar{F}_Y(\hat{y}) = g^{-1}(\hat{y}) \hat{y} \bar{F}_Y(\hat{y}). \tag{A43}$$

Turning to producer surplus from a treatment,

$$\pi_t \geq \hat{y} \int_{\hat{y}}^{\bar{y}} g^{-1}(y_i)\, dF_Y(y_i) \tag{A44}$$

$$\geq \hat{y} \int_{\hat{y}}^{\bar{y}} g^{-1}(\hat{y})\, dF_Y(y_i) \tag{A45}$$

$$= g^{-1}(\hat{y}) \hat{y} \bar{F}_Y(\hat{y}) \tag{A46}$$

$$= \pi_p. \tag{A47}$$

Equation (A44) holds because the producer surplus at the optimal treatment price $\pi_t$ at least weakly exceeds producer surplus from a treatment sold at price $\hat{y}_i$ on the right-hand side. To see that the right-hand side is the correct expression for this producer surplus, note that all types $y_i > \hat{y}$ buy the drug if they contract the disease. Each contracts the disease with probability $x_i = g^{-1}(y_i)$. Integrating over types gives the producer-surplus expression. Equation (A45) holds because $g^{-1}$ is an increasing function, so $x_i \geq g^{-1}(\hat{y}_i)$ for $y_i \geq \hat{y}_i$. Equation (A46) is a straightforward calculation. Equation (A47) follows from (A43). The inequality in (A45) is strict if there is nontrivial heterogeneity in the distribution of $x_i$ for preventive consumers. *Q.E.D.*

**Proof of Proposition 12:** Applying the change of variables used in the proof of Proposition 2, $\hat{x} = h p_p$, to (6) yields

$$\pi_p = \max_{\hat{x}} h \hat{x} \bar{F}_X(\hat{x}). \tag{A48}$$

Cross multiplying (7) and (A48) and substituting the definitions of $R_X$ and $R_H$, we have that $\pi_p > \pi_t$ if and only if $R_X/x > R_H/h$. The reverse inequality is proved similarly. *Q.E.D.*

**Proof of Proposition 13:** Suppose consumers are homogeneous in disease risk ($x$) and heterogeneous in income ($y_i$). Let $h_i = g(y_i)$ for some measurable function $g$. First, compute preventive revenue. Consumer $i$ buys the preventive if $E(xy_i h_i | y_i) \geq p_p$, implying $y_i g(y_i) \geq p_p/x$. Thus the quantity of preventive sold is

$$\left|\left\{y_i | y_i g(y_i) \geq p_p/x\right\}\right|,$$

where $|\cdot|$ denotes the measure of a set. Preventive revenue is thus

$$\pi_p = \max_{p_p \in [0,\infty)} \left\{ p_p \left| \left\{ y_i | y_i g(y_i) \geq p_p / x \right\} \right| \right\} \quad \text{(A49)}$$

$$= \max_{\tilde{p} \in [0,\infty)} \left\{ x\tilde{p} \left| \left\{ y_i | y_i g(y_i) \geq \tilde{p} \right\} \right| \right\}, \quad \text{(A50)}$$

using the change of variables $\tilde{p} = p_p / x$. Second, compute treatment revenue. Consumer $i$ buys the treatment if $y_i h_i \geq p_t$, implying $y_i g(y_i) \geq p_t$. Thus the quantity of treatment sold is

$$x \left| \left\{ y_i | y_i g(y_i) \geq p_t \right\} \right|,$$

implying treatment revenue is

$$\pi_t = \max_{p_t \in [0,\infty)} \left\{ x p_t \left| \left\{ y_i | y_i g(y_i) \geq p_t \right\} \right| \right\},$$

the same as (A50). *Q.E.D.*

# Appendix B: Extension Allowing for Entry of Competitors

The text focused on the case of a monopoly manufacturer of perfectly safe, effective, and costless pharmaceuticals. In this appendix, we first extend the model to allow for more general product characteristics. We then allow for the possibility of entry of competitors. We do this in the context of an oligopoly model incorporating some realistic institutional features. In particular, the patent system in the model provides only temporary monopoly power to a firm that develops a new product, after which there is generic entry. The main results in the text largely carry over to this extension. We do find a new source of bias against preventives which is engendered by competition.

**General Parameters:** It is useful to first relax the assumption that all products are perfectly safe and effective and costless to manufacture. This makes it possible to consider the case in which the social benefit of preventives and treatments may differ. We will show that the key welfare results from Section 3 continue to hold in this more general setting. Let $c_j \in [0,\infty)$ be the present discounted value of the marginal cost of manufacturing product $j \in \{p,t\}$ and administering it to a consumer. Let $e_j \in [0,1]$ be the efficacy of product $j$—the probability that product $j$ prevents the consumer from experiencing harm from the disease. Let $s_j \in [0,1]$ be the expected harm of side effects from product $j$—the probability that a consumer experiences side effects multiplied by the present discounted value of the harm from the side effects conditional on experiencing them.

**Proposition 14.** *The key welfare results from Section 3 continue to hold for general values of the parameters $c_j \in [0,\infty)$, $e_j \in [0,1]$, and $s_j \in [0,\infty)$.*

*i. The firm never develops a preventive in preference to a treatment unless it is socially efficient to do so. There exist cases in which the firm develops a treatment but it would have been socially efficient to develop a preventive.*

*ii. $1-\pi_p/\pi_t$ provides a tight upper bound on social cost $\sup_{k_j,c_j,e_j,s_j}[(WF-WE)/D]$.*

*iii. There exist parameters $c_j \in [0,\infty)$, $e_j \in [0,1]$, and $s_j \in [0,\infty)$ and distributions of disease risk such that $\pi_p/\pi_t$ can be made arbitrarily close to zero.*

*Proof.* To prove part (i), a treatment is always developed if it is socially efficient to do so because a treatment extracts 100% of social surplus. The proof of Proposition 8 provides a case in which a treatment is developed but it would have been socially efficient to develop a preventive. The proof of part (ii) is similar to Proposition 3 with the added fact that the supremum is generated by setting $c_j = s_j = 0$ and $e_j = 1$ for $j \in \{p,t\}$, the values that happen to be assumed in Proposition 3. Part (iii) follows immediately from Proposition 5. *Q.E.D.*

**Modeling Entry:** To allow for generic entry, we extend the model of Section 2 to an overlapping-generations setting. In period 0, $N$ firms with the research capacity to develop new products sequentially decide whether to expend fixed cost $k_j$ and develop one product $j$ or not to enter. Each period $t = 1,2,\ldots$ thereafter, the old generation from $t-1$ ($O_{t-1}$) dies, the young generation from $t-1$ ($Y_{t-1}$) becomes old ($O_t$), and a young generation ($Y_t$) with distribution of disease risk $F(x_i)$ is born. To simplify the analysis, we will focus on one source of heterogeneity, disease risk, and abstract away from other sources of heterogeneity such as income. Consumers have the following life cycle: young consumers first learn of their disease risk, decide whether or not to purchase the preventive if one is available, and then turn old; old consumers contract the disease or not, decide whether or not to buy a treatment if infected, and then die. Let $\delta \in [0,1]$ be the per-period discount factor.

The first firm to develop a product enjoys patent protection for one period. After product $j$ goes off patent, a fringe of generic manufacturers enter, and price falls to marginal cost $c_j$. Besides delaying generic entry, the patent prevents others of the $N$ research-capable firms from developing the same product. (Even if a second firm were able to invent a "me-too" substitute around the first firm's patent for product $j$, in equilibrium the second firm would not develop the "me-too" product if competition between them were intense enough to reduce producer surplus below the development cost $k_j$.) Thus, we can restrict attention to at most a first and second mover, which must develop different products.

In this model, competition between a preventive and a treatment is asymmetric. Competition from a preventive does not reduce the profits of the treatment patenter, which makes its profits from sales to the infected among the initial old generation $O_1$. It is too late for these consumers to receive a preventive, and they will die before generic version of treatments become available. On the other hand, competition from a treatment does reduce the profits of the preventive patenter. The preventive patenter makes its profits from sales to the initial young generation $Y_1$. The treatment is a substitute product for these consumers: rather than buying the preventive, they can wait to see if they become infected and buy the treatment. This competition effect is amplified because the generation $Y_1$ consumers will not only have access to the patented treatment but also will benefit from competition from generics that follow, driving treatment prices to marginal cost.

**Equilibrium:** To derive the equilibrium of this model, first consider the firm's profit from developing a treatment. Let $\Pi_t$ be the single-period monopoly profit from a treatment. Extending (2) to allow for general parameter values, it can be shown that $\Pi_t = (e_t h - s_t - c_t)x - k_t$. In the competition model, the firm earns $\Pi_t$ as well, whether its rival produces a preventive or does not enter. The firm earns this $\Pi_t$ by serving the infected in generation $O_1$. It earns zero flow profit serving subsequent generations because of generic entry.

A firm's profit from developing a preventive depends on what its rival does. If its rival does not enter, the present value of its profit stream, denoted $\Pi_{p0}$, has the same functional form as $\Pi_p$ from equation (1), but where the cutoff type indifferent between buying and not changes from $\hat{x}(p_p) = p_p/h$ to $\hat{x}(p_p) = (p_p + s_p)/(\delta e_p h)$. The preventive developer earns this $\Pi_{p0}$ from selling to consumers in generation $Y_1$. The discount factor $\delta$ inserted in the new formula for $\hat{x}(p_p)$ reflects the fact that the benefit to consumers in generation $Y_1$ from consuming the preventive is the harm avoided in the next period when they become generation $O_2$. The preventive developer earns zero flow profit serving subsequent generations because of generic entry. If the rival develops a treatment rather than not entering, the preventive developer's profit is lower because consumers in generation $Y_1$ anticipate cheap generic treatments will be available when they become generation $O_2$. The present value of the preventive developer's profit stream, denoted $\Pi_{pt}$, again has the same functional form as $\Pi_p$ in equation (1), but now the formula for the cutoff type is

$$\hat{x}(p_p) = \frac{p_p + s_p}{\delta e_p [c_t + s_t + (1 - e_t)h]}. \tag{B1}$$

Equation (B1) comes from equating the surplus the marginal preventive consumer in generation $Y_1$ obtains if he/she buys the preventive to that if he/she waits until the next period and buys the treatment at price $c_t$ if he/she becomes infected. Equation (B1) accounts for the fact that a consumer of the preventive has the option of getting the treatment in the next period if the preventive turns out to be ineffective. Again, the preventive developer earns zero flow profit serving subsequent generations because of generic entry.

Entry decisions in the subgame-perfect equilibrium can be characterized as follows. If $\Pi_{pt} > \Pi_t > 0$, the first mover develops a preventive and the second mover a treatment. If $\Pi_t > \Pi_{pt} > 0$, the first mover develops a treatment and the second mover a preventive. If $\Pi_t > 0 > \Pi_{pt}$, the first mover develops a treatment and the second mover does not enter. If $\Pi_{p0} > 0 > \Pi_t$, the first mover develops a preventive and the second mover does not enter. If $0 > \max(\Pi_t, \Pi_{p0})$, neither firm enters. Ignoring knife-edge cases $\Pi_t = 0$, $\Pi_{p0} = 0$, and $\Pi_{pt} = 0$, equilibrium entry decisions can be neatly summarized: a treatment is developed (either alone or together with a preventive) if and only if $\Pi_t > 0$; a preventive is developed (either alone or together with a treatment) if and only if (a) $\Pi_{pt} > 0$ or (b) $\Pi_{p0} > 0 > \Pi_t$.

The next proposition formalizes the notion that competition adds a new effect biasing firms in favor of treatments and against preventives.

**Proposition 15.** *The existence of $N \geq 2$ competing firms in the model enlarges the set of parameters for which a treatment is developed and reduces the set of parameters for which a preventive is developed compared to a model in which a single research-capable firm makes both sequential development decisions.*

*Proof.* Compare the present model involving competition between preventives and treatments, which we will label Model 1, to the monopoly model laid out in the statement of the proposition, which we will label Model 2. We begin by proving two facts that will be useful later in the proof. Fact 1 is that $\Pi_b$, the monopolist's profit from developing both products, equals $\Pi_t + \Pi_{pt}$. Conditional on developing both, the monopolist's optimal pricing strategy is to charge a treatment price maximizing profit from sales to generation $O_1$, yielding marginal profit $\Pi_t$, and charging a price for the preventive that maximizes profit from sales to generation $Y_1$ given generics will enter the treatment market, yielding marginal profit $\Pi_{pt}$. Fact 2 is that $\Pi_b \leq \Pi_t + \Pi_{p0}$. This holds because $\Pi_{p0} \geq \Pi_{pt}$ because of the business-stealing effect between preventives and treatments due to their substitutability.

Suppose the parameters are such that a treatment is not developed in equilibrium in Model 1. According to the paragraph preceding the proposition, we must have $\Pi_t < 0$. (We ignore knife-edged cases such as $\Pi_t = 0$ throughout the proof for simplicity. It is easily seen that the proof holds for these cases as well.) But $\Pi_t < 0$ implies $\Pi_b < \Pi_{p0}$ by Fact 2, in turn implying $\max(\Pi_t, \Pi_b) < \max(\Pi_{p0}, 0)$, and so a treatment would not be developed in equilibrium in Model 2.

Suppose the parameters are such that a preventive is developed in equilibrium in Model 1. According to the paragraph preceding the proposition, either (a) $\min(\Pi_t, \Pi_{pt}) > 0$ or (b) $\Pi_{p0} > 0 > \Pi_t$. If (a) holds, then by Fact 1, $\Pi_b = \Pi_t + \Pi_{pt} > \Pi_t > 0$. Thus, $\max(\Pi_{p0}, \Pi_b) > \max(\Pi_t, 0)$. Thus a preventive is developed in equilibrium in Model 2. If (b) holds, then again $\max(\Pi_{p0}, \Pi_b) > \max(\Pi_t, 0)$, and so a preventive is developed in equilibrium in Model 2.

The proof is completed by constructing a case in which a treatment is developed in equilibrium in Model 1 but a preventive is developed in equilibrium in Model 2. Let consumers be homogeneous, with $x_i = 1$ for all $i$. Let $\delta = e_p = 1$. Let $c_j = s_j = 0$ for $j \in \{p, t\}$. Let $k_t < e_t h$ and $k_p \in \big((1 - e_t)h, (1 - e_t)h + k_t\big)$. It can be shown that $\Pi_t = e_t h - k_t > 0$, $\Pi_{p0} = h - k_p$, and $\Pi_{pt} = (1 - e_t)h - k_p < 0$. According to the paragraph preceding the proposition, since $\Pi_t > 0 > \Pi_{pt}$, a preventive alone is developed in equilibrium in Model 1. Since $k_p < (1 - e_t)h + k_t$, $\Pi_{p0} > \Pi_t$. Hence $\Pi_{p0} > \Pi_t > \Pi_t + \Pi_{pt} = \Pi_b$, where the last step holds by Fact 1. Thus, a preventive alone is developed in equilibrium in Model 2. *Q.E.D.*

The logic behind Proposition 15 is that a monopolist would internalize the negative business-stealing effect that treatments exert on preventives arising because products are substitutes. There exist cases in which a monopolist would not develop the treatment in order to keep preventive profit high, while a competing firm would develop the treatment since it does not care about preventive profits, and in some of these cases drug entry deters entry of a preventive.

The competition effect identified in Proposition 15 may be socially costly, as the next proposition states.

**Proposition 16.** *In the competitive model, social welfare never falls with a reduction in the cost of developing a preventive, $k_p$, but may fall with a reduction in the cost of developing a treatment, $k_t$.*

*Proof.* All of the direct and indirect effects of reducing $k_j$ on social welfare are non-positive except possibly for one: the possibility of deterring entry by the other product. In the text, we established that a treatment will be developed if $\Pi_t > 0$, independent of the preventive's entry decision, and thus independent of $k_p$. So reducing $k_p$ weakly increases social welfare.

The proof is completed by demonstrating a case in which a reduction in $k_t$ reduces social welfare. Let consumers be homogeneous, with $x_i = 1$ for all $i$. Let $e_p = 1$. Let $c_j = s_j = 0$ for $j \in \{p,t\}$. Let $k_p \in \big((1-e_t)h, h\big)$. We will compare the case in which $k_t$ is high, namely $k_t \in (e_t h, \infty)$, to a case in which $k_t$ is low, namely $k_t = 0$. In the first case, $\Pi_t = e_t h - k_t < 0$. Further, $\Pi_{p0} > 0$. But, as noted in the text preceding Proposition 15, $\Pi_{p0} > 0 > \Pi_t$ implies that a preventive alone is developed. The present discounted value of the stream of social welfare in equilibrium is

$$\frac{\delta h}{1-\delta} - k_p. \tag{B2}$$

In the second case, $\Pi_t = e_t h - k_t = e_t h > 0$. Further, $\Pi_{pt} = (1-e_t)h - k_p < 0$. But, as noted in the text preceding Proposition 15, $\Pi_t > 0 > \Pi_{pt}$ implies that a treatment alone is developed. The present discounted value of the stream of social welfare in equilibrium is

$$\frac{e_t h}{1-\delta} - k_t. \tag{B3}$$

The limit as $\delta \to 1$ of the ratio of expression (B2) to (B3) equals $1/e_t$. Thus, for $\delta$ sufficiently close to one, both $k_t$ and social welfare are higher in the first than the second case. *Q.E.D.*

The intuition behind Proposition 16 is that a reduction in $k_t$ increases the incentive to develop a treatment, which may deter the entry of preventives, even some preventives that generate more social surplus than the treatment. As noted, competition between preventives and treatments is asymmetrically tougher on preventives, so preventives do not have a similar competitive effect on treatments.

## References


Acemoglu, D. and J. Linn (2004). "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," *Quarterly Journal of Economics* 119: 1049–1090.

Anderson, O. W., P. Collette, and J. J. Feldman. (1960) *Expenditure Patterns for Personal Health Services, 1953 and 1958: Nationwide Survey*. New York: Health Information Foundation.

Ault, Kevin A. (2007) "Human Papillomavirus Vaccines and the Potential for Cross-protection Between Related HPV Types," *Gynecologic Oncology* 107: S31–S33.

Australian Government Department of Health and Ageing. (2009) *Guidelines for Deeds of Agreement for Pharmaceutical Benefits Scheme (Version 1.3)*. Canberra, Australia.

Barder, O., M. Kremer, and R. Levine. (2005) *Making Markets for Vaccines: Ideas to Action*. Center for Global Development.

Blanchflower, D. G. and A. J. Oswald. (2004) "Well-Being Over Time in Britain and the USA," *Journal of Public Economics* 88: 1359–1386.

Boulier, B. (2006) "A Shot in the Dark: Uncertainty and Vaccine Demand and Supply," George Washington University working paper.

Brito, D. L., E. Sheshinski, and M. D. Intrilligator. (1991) "Externalities and Compulsory Vaccination," *Journal of Public Economics* 45: 69–90.

Centers for Disease Control. (2005) *National Health and Examination Survey (NHANES) 2003–2004*. Washington, D.C.

Centers for Disease Control. (2006a) "Cases of HIV Infection and AIDS in the United States, 2004," *HIV/AIDS Surveillance Report*, vol. 16. Table 8: "Estimated Numbers of Persons Living with HIV/AIDS, by Year and Selected Characteristics, 2001-2004—35 Areas with Confidential Name-Based HIV Infection Reporting" downloaded March 25, 2006 from www.cdc.gov/hiv/topics/surveillance/resources/-reports/2004report/default.htm.

Centers for Disease Control. (2006b) *HIV Prevalence Trends in Selected Populations in the United States: Results from National Serosurveillance, 1993–1997*. Table 3: "HIV Prevalence Among Injection Drug Users Entering Drug Treatment Centers, by Metropolitan Area and Sex, 1993–1997" downloaded April 14, 2006 from www.cdc.gov/hiv/pubs/hivprevalence/selected.htm.

Centers for Disease Control. (2009) "Vaccines & Preventable Diseases: List of Vaccines Used in United States," downloaded October 10, 2009 from www.cdc.gov/vaccines/vpd-vac/vaccines-list.htm.

Clay, K. B., D. S. Sibley, and P. Srinagesh. (1992) "Ex Post vs. Ex Ante Pricing: Optional Calling Plans and Tapered Tariffs," *Journal of Regulatory Economics* 4: 115–138.

Coase, R. J. (1972) "Durability and Monopoly," *Journal of Law and Economics* 15: 143–149.

Corey, E. J., L. Kürti, and B. Czakó. (2007) *Molecules and Medicine*. Hoboken: Wiley.

Courty, P. (2003) "Ticket Pricing Under Demand Uncertainty," *Journal of Law and Economics* 46: 627–652.

Courty, P. and H. Li. (2000) "Sequential Screening," *Review of Economic Studies* 67: 697–717.

Danzon P. M. (1998) "The Economics of Parallel Trade," *PharmacoEconomics* 13: 293–304.

Dunne, E. F., *et al.* (2007) "Prevalence of HPV Infection Among Females in the United States," *Journal of the American Medical Association* 297: 813–819.

Euerle, B. and P. H. Chandrasekar. (2012) "Syphilis," in B. A. Cunha, ed., *Medscape Reference*. Accessed August 27, 2012 from emedicine.medscape.com/article/229461.

Fey Cortez, Michelle and Simeon Bennett. (2011) "Gilead HIV Breakthrough of the Year Stymied by $12,000 Cost, Side Effects," *Bloomberg News*, Feb. 28. Downloaded December 29, 2012 from www.bloomberg.com/news/2011-02-28/gilead-s-12-000-a-year-hiv-prevention-pill-fails-to-win-physician-support.html

Francis, P. J. (1997) "Dynamic Epidemiology and the Market for Vaccinations," *Journal of Public Economics* 63: 383-406.

Finkelstein, A. (2004). "Static and Dynamic Effect of Health Policy: Evidence from the Vaccine Industry," *Quarterly Journal of Economics* 119: 527–564.

GEN News Highlights. (2012) "FDA: HIV Numbers Drove Truvada Decision," July 17, article no. 81247053.

Geoffard, P.-Y. and T. Philipson. (1997) "Disease Eradication: Public vs. Private Vaccination," *American Economic Review* 87: 222-230.

Gersovitz, M. (2003) "Births, Recoveries, Vaccinations, and Externalities," in R. Arnott, ed., *Economics for an Imperfect World: Essays in Honor of Joseph E. Stiglitz*, 469–483.

Gersovitz, M. and J. S. Hammer. (2004) "The Economical Control of Infectious Diseases," *Economic Journal* 114: 1–27.

Gersovitz, M. and J. S. Hammer. (2005) "Tax/Subsidy Policy Toward Vector-Borne Infectious Diseases," *Journal of Public Economics* 89: 647–674.

Getzen, T. E. (2000) "Health care is an Individual Necessity and a National Luxury: Applying Multilevel Decision Models to the Analysis of Health Care Expenditures," *Journal of Health Economics* 19: 259–270.

Grady, D. (2012) "F.D.A. Advisory Panel Back Preventive Use of H.I.V. Drug," *New York Times*, May 11: A1.

Greenwood, D. (2008) *Antimicrobial Drugs: Chronicle of a Twentieth Century Medical Triumph*. Oxford: Oxford University Press.

Harpavat, S. and S. Nissim. (2001) *MicroCards: Review Cards for Medical Students*. Philadelphia: Lippincott Williams & Wilkins.

Harris, M. and A. Raviv. (1981) "A Theory of Monopoly Pricing Schemes with Demand Uncertainty," *American Economic Review* 71: 347–365.

Hernandez, B. Y., *et al.* (2008) "Transmission of Human Papillomavirus in Heterosexual Couples," *Emerging Infectious Diseases* 14: 888–894.

Howard, Robin S. (2005) "Poliomyelitis and the Postpolio Syndrome," *British Medical Journal* 330: 1314–1318.

Immunization Action Coalition. (2009) *Print Materials by Diseases and Vaccines* downloaded October 10, 2009 from www.immunize.org/printmaterials/dis_tet.asp.

Kaplan, E. H. (1990) "Modeling HIV Infectivity: Must Sex Acts Be Counted?" *Journal of Acquired Immune Deficiency Syndromes* 3: 55–61.

Kessing, S. G. and R. Nuscheler. (2006) "Monopoly Pricing with Negative Network Effects: The Case of Vaccines," *European Economic Review* 50: 1061–1069.

Klein, B., R. A. Crawford, and A. A. Alchian. (1978) "Vertical Integration, Appropriable Rents, and the Competitive Contracting Process," *Journal of Law and Economics* 21: 297–326.

Kremer, M. and R. Glennerster. (2004) *Strong Medicine: Creating Incentives for Pharmaceutical Research on Neglected Diseases*. Princeton: Princeton University Press.

Kremer, M., C. M. Snyder, and H. Williams. (2012) "Vaccines: Integrated Economic and Epidemiological Models," mimeo, Harvard University.

Lau, Brandyn D., Brian L. Pinto, David R. Thiemann, and Christoph U. Lehmann. (2011) "Budget Impact Analysis of Conversion from Intravenous to Oral Medication When Clinically Eligible for Oral Intake," *Clinical Therapeutics* 33: 1792–1796.

Lewis, T. R. and D. E. M. Sappington. (1994) "Supplying Information to Facilitate Price Discrimination," *International Economic Review* 35: 309–327.

Malueg, D. A. (1994) "Monopoly Output and Welfare: The Role of Curvature of the Demand Function," *Journal of Economic Education* 25: 235–250.

Maleug, D. A. and C. M. Snyder (2006) "Bounding the Relative Profitability of Price Discrimination," *International Journal of Industrial Organization* 24: 995-1011.

Mandell, G. L., J. E. Bennett, and R. Dolin. (2009) *Principles and Practice of Infectious Diseases* seventh edition. Philadelphia: Elsevier Churchill Livingstone.

Miravete, E. (1996) "Screening Consumers Through Alternative Pricing Mechanisms," *Journal of Regulatory Economics* 9: 111–132.

Morbidity and Mortality Weekly Report. (various years) "Summary of Notifiable Diseases, United States," Centers for Disease Control and Prevention, downloaded December 20, 2009 from www.cdc.gov/-mmwr/mmwr_nd/index.html

Mueller, Steffen, Eckard Wimmer, and Jeronimo Cello. (2005) "Poliovirus and Poliomyelitis: A Tale of Guts, Brains, and an Accidental Event," *Virus Research* 111: 175–193.

National Cancer Institute. (2009) "BRCA1 and BRCA2: Cancer Risk and Genetic Testing," *National Cancer Institute Fact Sheet.* Retrieved August 16, 2012, from www.cancer.gov/cancertopics/factsheet/-Risk/BRCA.

National Network for Immunization Information. (2009) "Vaccine Information," downloaded October 10, 2009 from www.immunizationinfo.org/vaccineInfo/index.cfm.

Newell, R., A. Jaffee, and R. N. Stavins. (1999) "The Induced Innovation Hypothesis and Energy-Saving Technological Change," *Quarterly Journal of Economics* 114: 907–940.

Newhouse, J. P. (1977) "Medical Care Expenditure: A Cross-National Survey," *Journal of Human Resources* 12: 115–125.

Nowsheen, S., et al. (2012) "ER2 Overexpression Renders Human Breast Cancers Sensitive to PARP Inhibition Independently of Any Defect in Homologous Recombination DNA Repair," *Cancer Research* 72: 4796–4806.

Oster, E., I. Shoulson, K. Quaid, and E. R. Dorsey. (2010) "Genetic Adverse Selection: Evidence from Long-Term Care Insurance and Huntington Disease," *Journal of Public Economics* 94: 1041–1050.

Pecorino, P. (2002) "Should the US Allow Prescription Drug Reimports from Canada?" *Journal of Health Economics* 21: 699–708.

Rockstroh, J. K., *et al.* (1995) "Male-to-Female Transmission of HIV in a Cohort of Hemophiliacs—Frequency, Risk Factors and Effect of Sexual Counseling," *Infection* 23: 29–32.

Rosenberg, E. (1999) "Drug Makers Shy from Work on AIDS Vaccine," *San Francisco Examiner*. March 16.

Royce, R. A., *et al.* (1997) "Sexual Transmission of HIV," *New England Journal of Medicine* 336: 1072–1078.

Snyder, C. M., W. Begor, and E. R. Berndt. (2011) "Economic Perspectives on the Advance Market Commitment for Pneumococcal Vaccines," *Health Affairs* 30: 1508–1517.

Stole, L. A. and J. Zwiebel. (1996) "Organizational Design and Technology Choice under Intrafirm Bargaining," *American Economic Review* 86: 195–222.

Thomas, P. (2002) "The Economics of Vaccines," *Harvard Medical International (HMI) World*. September/October.

UNAIDS. (2000) *Epidemiological Fact Sheets by Country*.

UNAIDS. (2004) *USA: Epidemiological Fact Sheets on HIV/AIDS and Sexually-Transmitted Infections*.

U.S. Food and Drug Administration. (2009) "Complete List of Vaccines Licensed for Immunization and Distribution in the US," downloaded October 10, 2009 from www.fda.gov/BiologicsBloodVaccines/-Vaccines/ApprovedProducts/ucm093833.htm.

World Bank. (2000) *World Development Indicators 2000*. Washington, DC.