HIV Breakthroughs and Risky Sexual Behavior*

Dana Goldman
RAND and NBER

Darius Lakdawalla
RAND and NBER

Neeraj Sood
RAND

March 23, 2004

Abstract

Recent breakthroughs in the treatment of HIV have coincided with an increase in infection rates and an eventual slowing of reductions in HIV mortality. These trends may be causally related, if treatment improves the health and functional status of HIV+ individuals and allows them to engage in more sexual risk-taking. We examine this hypothesis empirically using access to health insurance as an instrument for treatment status. We find that treatment results in more sexual risk-taking by HIV+ adults. This relationship implies that breakthroughs in treating an incurable disease like HIV can increase precautionary behavior by the uninfected and thus reduce welfare. We also show that, in the presence of this effect, treatment and prevention are social complements for incurable diseases, even though they are substitutes for curable ones. Finally, there is less over-provision of treatment for an incurable disease than a curable one, because of the negative externalities associated with treating an incurable disease.

*We are grateful to Jean-Jacques Laffont and Vazha Nadareishvili for helpful discussions. Abby Alpert provided exceptional research assistance.

1 Introduction

While the search for a safe, cheap, and effective treatment for HIV is far from over, remarkable strides have been made over the past ten years. During the mid-1990s, highly active antiretroviral treatments (HAART) for HIV — combinations of three or more drugs that include a protease inhibitor — became available. These treatments substantially improved the health status of HIV patients and represented a remarkable advance over earlier HIV treatments. The numbers of new AIDS cases and deaths due to AIDS have fallen sharply since 1996 (Center for Disease Control, 1998; Palella et al., 1998). This trend in declining mortality rates has been attributed primarily to the increased use of HAART. The effectiveness of these new therapies at lowering mortality and morbidity has been demonstrated both in clinical trials (Hammer et al., 1997; Staszewski et al., 1999) and observational studies on patients receiving care for HIV in outpatient settings (Detels et al., 1998; Palella et al., 1998; Lucas et al., 1999; Vittinghoff et al., 1999; Lucas et al., 2003). It has also been shown to be a cost-effective way of improving outcomes (Freedberg et al., 2001). There is thus a substantial body of evidence that HAART improves outcomes both in clinical settings and in society.

However, breakthroughs in treatment alone are not enough to guarantee health. Both treatment and risk taking (or lack of preventive) behavior determine rates of infection, and there are reasons to believe that advances in treatment may have encouraged more risk-taking and less precaution. Figure 1 illustrates an important

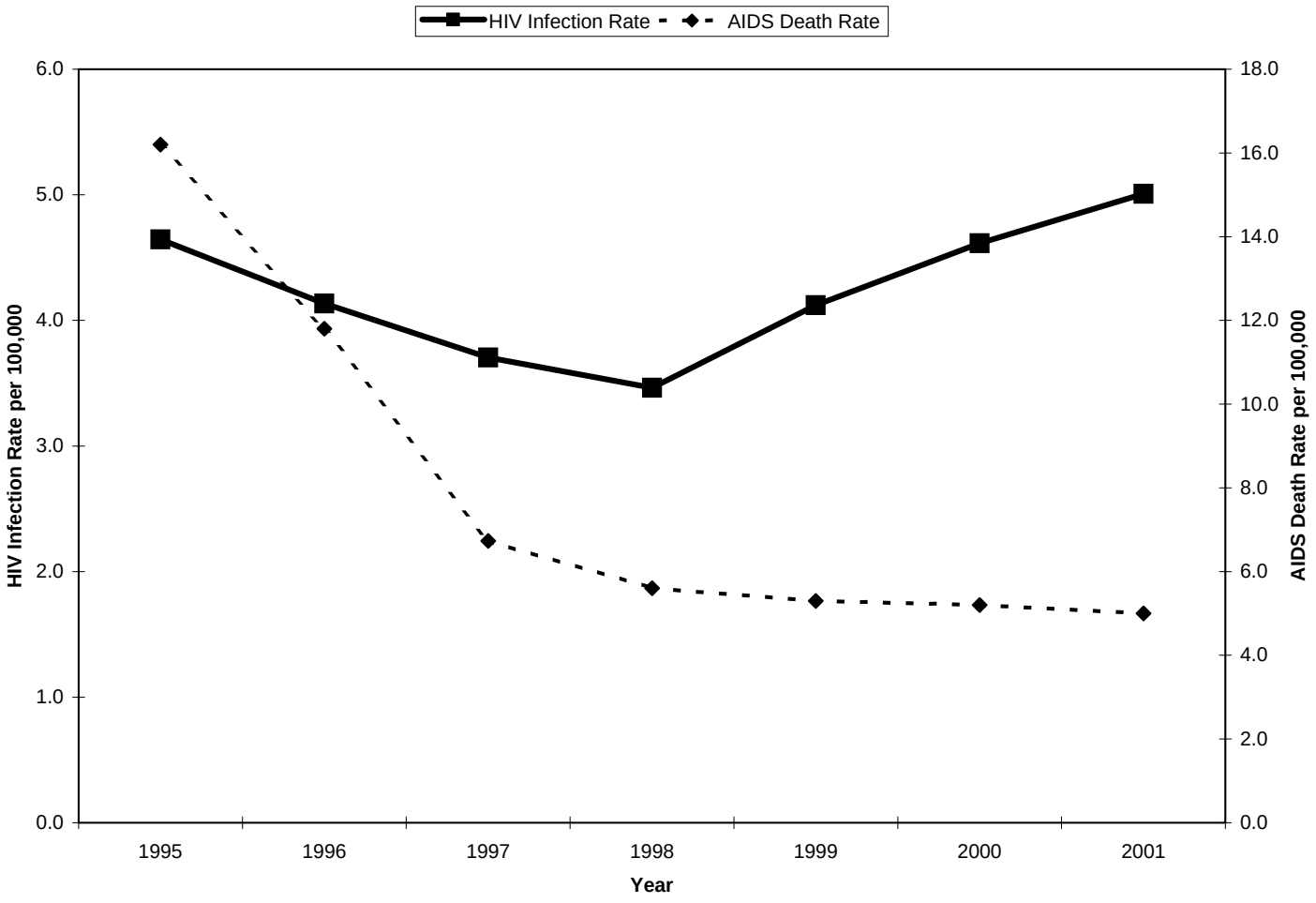
trend in HIV since the development of HAART. The period immediately following the introduction of HAART, around 1995 or 1996, saw a dramatic decline in the overall rate of death from AIDS. However, death rates reached a plateau in 1998, where they have remained roughly level since. Moreover, the plateau appeared to coincide with a steady and significant increase in HIV infection rates, and thus new HIV cases.

The slowdown in deaths from HIV, coupled with an increase in infection rates, suggest that increases in risky behavior might be working at cross-purposes with advances in treatment. Better treatment of HIV does not eliminate the infection, and it may make the infected person healthier and more efficient at spreading the disease.

In this paper we investigate this hypothesis by estimating the causal effect of treatment with HAART on the sexual activity of HIV+ persons. Simple correlations between treatment and sexual activity reveal that treated individuals engage in less risky behavior, but these estimates suffer from the problem that treated people are likely to be sicker than untreated people. To overcome this problem, we exploit state-level variation in the availability of public insurance for HIV-sufferers. Since treatment with HAART is expensive — costing on average about 13,000 dollars per year — it is likely that people with insurance are more likely to get HAART. We find that HIV-sufferers who get treatment because they live in better insured states are more likely to engage in risky behavior.

A relationship between treatment and risky behavior by the infected has important positive and normative implications. Breakthroughs in the treatment of an incurable

Figure 1: HIV Incidence and Deaths from AIDS.



disease can increase preventive behavior among the uninfected, because treatment increases the number of infected people capable of spreading the disease. As a result, breakthroughs can reduce welfare among the uninfected. This leads to a further normative implication unique to an incurable disease like HIV: treatment and prevention are social complements. This is of particular significant in developing countries that might receive gifts of treatment or prevention resources. A sudden increase in the level of treatment raises the marginal utility of prevention; the effect is exactly opposite for a curable disease, where treatment lowers the marginal utility of investing in prevention. In addition, because treatment increases the risk faced by the uninfected, it involves a unique negative externality in the case of incurable disease. There is thus less under-provision of treatment by the private market for an incurable disease than a curable one.

We first describe the data used for this analysis and then present our joint model of treatment and sexual activity. Subsequently, we compare the results from our model with a reduced form model that treats HAART as exogenous. The final sections discuss the positive and normative implications of our findings for preventive behavior among uninfected and the effect of HAART on incidence of HIV. We conclude by discussing the implications of our results for public subsidies for expanding treatment and subsidies for encouraging medical innovations for the treatment of HIV and other chronic but incurable diseases.

2 Data

We use data from a nationally representative study of HIV+ patients in care-the HIV Costs and Services Utilization Study (HCSUS). The HCSUS employed a multi-stage national probability sample design to identify HIV+ patients over 18 years old, who made at least one visit for regular care in the contiguous United States in January or February of 1996. It does not include HIV+ patients whose only contact with the health care system was through military, prison, or emergency department facilities, or who have not made contact with the health care system for their HIV. HCSUS collected data between March 1996 and July 1997 - a period when HAART entered clinical practice and disseminated widely. However, questions about sexual activity were only asked to a random sample of 1,396 respondents. We use this subsample of HCSUS respondents for our analysis.

We construct analytic weights to adjust the sample to the reference population. A respondent's analytic weight, which may be interpreted as the number of people in the population represented by that respondent, is the product of three patient-specific quantities - the sampling weight, the multiplicity weight, and the non-response weight. The sampling weight adjusts for oversampling (of women, for example); the multiplicity weight adjusts for patients who could potentially enter the sample via multiple providers; and the non-response weight adjusts for differential cooperation (Duan et al. 1999). All analyses presented in this paper use these weights.

We use two different polychotomous variables as our outcome variables. The first

variable indicates whether the respondent had no sexual partners, one sexual partner, or more than one sexual partner. Similarly, the second variable indicates whether the respondent had no new sexual partner in the past 12 months, one new partner, or more than one new partner. The main explanatory variable we are interested in is treatment with HAART, which is derived from the HCSUS data. Table 1 presents descriptive statistics for all model variables.

Most of the variables are self-explanatory. In some models, we include measures of the lowest ever CD4+ t-lymphocyte cell count, a critical measure of the function of a patient's immune system. A depletion in these cells correlates strongly with the worsening of HIV disease and physical health (Fauci et al., 1998). In this paper, we categorize CD4+ counts into four categories, as shown in Table 1. Patients with CD4+ lymphocyte counts below 50 have a very poor prognosis in general; while those with counts above 500 are considered much healthier. Table 1 shows that patients in poor health are more likely to be treated with HAART.

3 Empirical Model

Let T_i^* represent an index function that measures the treatment propensity for HIV+ patient i.

$$T_i^* = \beta_1 X_i - \epsilon_{T,i} \tag{3.1}$$

Table 1: Weighted descriptive statistics by HAART status (N=1396).

Variable	HAART (N =823)	No HAART (N = 573)
Demographics		
Age (years)	39	39
Non-white (%)	65	56
Female (%)	33	25
Education		
Less than HS degree (%)	31	25
High school degree (%)	31	29
Some college or more (%)	28	27
College (%)	11	19
Lowest ever CD4 count (cells/μl)		
>500 (%)	11	2
200-499 (%)	46	33
50-200 (%)	26	37
0-50 (%)	17	27
AIDS (%)	33	46
State instruments		
Medically-needed threshold	47% of FPL	49% of FPL
AFDC threshold	180% of FPL	180% of FPL
SSI threshold <65% of FPL	93	91
Number of partners		
No partners (%)	35	33
1 partner (%)	38	41
2 or more partner (%)	27	26
Number of new partners		
No partners (%)	65	68
1 partner (%)	14	12
2 or more partner (%)	17	18

The vector X_i represents observed exogenous covariates that determine treatment propensity, such as age, gender, health and education and our state policy instruments that are correlated with insurance and treatment with HAART. Treatment is also assumed to depend on a random error component $\epsilon_{T,i}$ that is uncorrelated with X_i.

We define T_i as an indicator variable that represents whether patient i actually received treatment with HAART:

$$T_i = \begin{cases} 0 & \text{if } T_i^* \leq 0, \\ 1 & \text{if } T_i^* > 0. \end{cases} \tag{3.2}$$

Similarly, let S_i^* represent the sexual activity propensity for HIV+ patient i

$$S_i^* = \beta_2 Z_i - \epsilon_{S,i} \tag{3.3}$$

Z_i is a set of exogenous variables determining sexual activity propensity, and $\epsilon_{S,i}$ represents the error term.

We define S_i as a polychotomous indicator variable that measures the sexual activity of patient i. An individual can have no partner ($S_i = 0$), one partner ($S_i = 1$), or more than one partner ($S_i = 2$), according to the following expression:

$$S_i = \begin{cases} 0 & \text{if } S_i^* \leq \delta_1 \ , \\ 1 & \text{if } S_i^* \in (\delta_1, \delta_2], \\ 2 & \text{if } S_i^* > \delta_2. \end{cases} \tag{3.4}$$

To complete the model and allow for correlation between mortality and insurance choices, we need to assume a joint distribution for the error terms. We assume that the errors are jointly distributed as bivariate standard normal with correlation coefficient ρ. This assumption implies a probit model for treatment and an ordered probit model for sexual activity. It is useful to think about the correlation between treatment propensity and sexual activity propensity as unobserved health. That is, patients with poor unobserved health are more likely to get treatment and they are also less likely to be sexually active.

We use maximum likelihood to estimate the parameters of our model. We have 6 possible outcomes for the dependent variables in our sample: (treatment/no treatment) x (no partner/one partner/2 or more partners). The contribution of patient i to the likelihood function is thus given by:

$$l_i = \Pr(T_i^* > 0, S_i^* \leq \delta_1 | X_i, Z_i)^{[T_i=1][S_i=0]} \times \Pr(T_i^* \leq 0, S_i^* \leq \delta_1 | X_i, Z_i)^{[T_i=0][S_i=0]} \times$$
$$\Pr(T_i^* > 0, S_i^* \in (\delta_1, \delta_2] | X_i, Z_i)^{[T_i=1][S_i=1]} \times \Pr(T_i^* \leq 0, S_i^* \in (\delta_1, \delta_2] | X_i, Z_i)^{[T_i=0][S_i=1]} \times$$
$$\Pr(T_i^* > 0, S_i^* > \delta_2 | X_i, Z_i)^{[T_i=1][S_i=2]} \times \Pr(T_i^* \leq 0, S_i^* > \delta_2 | X_i, Z_i)^{[T_i=0][S_i=2]} \tag{3.5}$$

Finally we obtain the weighted log-likelihood function by summing the log-likelihood across individuals:

$$\ln(\Gamma) = \sum_{i=1}^{N} \ln(l_i) w_i \tag{3.6}$$

Γ is the vector of model parameters; w_i are the analytic weights and N is the sample size. Because it is difficult to interpret the magnitude of the parameter estimates directly, we also report the mean marginal effect of treatment on the probability of having - no partners, one partner, 2 or more partners.

4 Identification

We use state Medicaid policies as our instrumental variables to explain treatment status but not sexual activity or physical health (except via treatment status). Medicaid is the most common form of insurance for the HIV+ population in care, covering 46% of the population. HIV+ patients can qualify for Medicaid through three distinct pathways. First, patients who meet the state's income eligibility and family composition requirements for Aid to Families with Dependent Children (AFDC) as they existed on July 16, 1996 qualify for Medicaid coverage. Second, Supplemental Security Income (SSI) beneficiaries are automatically eligible for Medicaid in 38 states. The other states have different standards for eligibility either as a 209(b) state or a waiver state. Section 209(b) of the Social Security Amendments Act of 1972 allows

States to include more restrictive definitions of "disability" and lower income and assets standards for Medicaid eligibility. Medicaid eligibility is also available through a "medically needy" program for individuals who meet Medicaid's disability criteria but have incomes that exceed the financial eligibility limit. The program allows individuals to "spend-down" to Medicaid eligibility by deducting medical-related expenses from their reported income. States have the option to but are not required to establish a medically needy program. In addition, states vary in their income eligibility levels for the medically needy program. For each patient, we define the following three variables based on Medicaid eligibility rules in the state in which the patient is sampled.

- "Medically Needy Threshold" is the state's income eligibility threshold for the medically needy program expressed as a percentage of the federal poverty line
- "AFDC Threshold" is the State's 1996 income eligibility threshold for Aid to Families with Dependent Children (AFDC) expressed as a percentage of the federal poverty line.
- "$SSI < 65\%\ FPL$" is an indicator variable for whether the state's income eligibility threshold for Medicaid eligibility through the "SSI" category was at least 10 percentage points lower than the federal guideline of 75 percent of the federal poverty line.

As in all IV-based studies, the credibility of our study rests on the believability of our instruments. Our state policy instruments could fail in at least two ways. First, the estimators perform poorly if the instruments are only weakly correlated with the treatment variable—i.e., receipt of HAART (Nelson and Startz, 1990; Bound et al., 1995; Staiger and Stock, 1997). Thus, we report Wald statistics for the joint significance of our instruments in predicting treatment status. Second, our instruments might be correlated with unobserved determinants of sexual activity (like unmeasured health status variables). The assumption that an instrumental variable is uncorrelated with the outcome measure cannot be directly tested. For these reasons, some researchers have argued that IV estimates in this context should be viewed with caution (Bound et al., 1995). However, in our application, it seems clear that patients have little direct influence at an individual level on state policies, so our state level instruments seem valid.

This argument is not enough to establish exogeneity, however, if there are unobserved state-level variables that determine both sexual activity or physical health and treatment status. In that case, state policies would be endogenous in our model despite the lack of control by patients over these policies. In order to address this issue, we develop some indirect evidence that our instruments are not simply picking up differences in unobserved health or sexual activity across states.

In our first test of this assumption, we estimate a logit model of one-year mortality using data from the Medicare Current Beneficiary Survey (MCBS). On the right-hand

side, this mortality model includes a sparse set of health status indicators, such as measures of Activities of Daily Living (ADLs) and a general health index, and our state-policy instruments. If our instruments are correlated with unobserved health, then one would expect to find that our state-level instruments predict the health of patients even in a non-HIV population. Since this elderly or disabled population is by definition insured by Medicare, our instruments should not predict their mortality unless they proxy for unobserved state-level effects. Table 2 reports the regression results and shows that our instruments are not statistically significant in the model, with odds ratios near one. Of course, these results do not prove that unobserved state effects are unimportant in the HCSUS population, but they are certainly suggestive. There is no good reason to expect that such effects should be present for HIV+ patients when they are not present for the elderly or disabled.

In our next test, we estimate whether our key instrument – medically needy threshold – predicts the sexual activity of unmarried respondents in the National Longitudinal Survey of Youth (NLSY). We restrict our sample to data from years 1983 to 1985 – a period when HAART was not available. In theory we should expect that sexual activity of respondents in the pre-HAART era is uncorrelated with state Medicaid policy. However, if our instrument is correlated with propensity for sex, then one would expect that it predicts sexual activity even in the pre-HAART era. Table 3 reports the regression results. The dependent variable is whether the respondent was sexually active in the month before the survey and the independent variables include

Table 2: Logit results of 1-year mortality of Medicare beneficiaries.

Variables	Odds Ratio	95% Conf. Interval	
State Instruments		Lower Limit	Upper Limit
Medically-Needy Threshold	1.001	0.997	1.004
AFDC Threshold	1.001	0.997	1.003
SSI Threshold < 65% of FPL	1.14	0.816	1.592
Average Firm Size	0.985	0.924	1.05
Number of Activities of Daily Living (ADL) in which limited			
No Limitation	Ref. Cat	Ref. Cat	Ref. Cat
1 ADL	2.375	1.803	3.129
2 ADL	2.503	1.819	3.443
3 ADL	3.296	2.307	4.723
4 ADL	3.376	2.331	4.889
5 ADL	5.857	4.347	7.893
6 ADL	7.589	5.786	9.954
Self reported Health Status			
Excellent	Ref. Cat	Ref. Cat	Ref. Cat
Very Good	1.625	1.042	2.536
Good	1.839	1.205	2.805
Fair	2.415	1.574	3.703
Poor	4.283	2.7726	6.618

our instrument and a sparse set of demographics. The results clearly show that the medically needy threshold does not predict sexual activity and is thus unlikely to be correlated with unobserved propensity for sex.

5 Results

To illustrate the consequences of selection bias, we estimate a "naïve" ordered probit model where HAART is treated as an exogenous variable. Table 4 reports the results from the naïve models. In order to demonstrate the importance of including information on health status, we include two sets of estimates; one set with controls for disease progression and one without.

In both regression models without controls disease progression, treatment with HAART is associated with reduced sexual activity, although the coefficients are not statistically significant. In other words, the models predict that respondents on HAART have fewer sexual partners and are also less likely to have new partners. This "anomalous" finding persists even after including controls for disease progression. However, in general the coefficients are substantially smaller in the models with severity of illness controls. We attribute these findings to a spurious positive correlation between severity of illness and treatment with HAART for HIV patients.

The parameter estimates for the structural model are shown in Table 5, along with the correlation between HAART and sexual activity. The first and third columns in Table 5 show the results for the treatment equation. The results show that the

Table 3: The Medically Needy Threshold and Sexual Activity in the NLSY Population, 1983-85.

Variables	Coefficients	95% Conf. Interval	
		Lower Limit	Upper Limit
Medically Needy Threshold	0.00004	-0.0004012	0.000481
Education			
less than or equal to 12 years	-0.016	-0.049	0.017
13 to 14 years	-0.051	-0.083	-0.018
15 to 16 years	-0.089	-0.140	-0.038
greater than or equal to 56 years	--	--	--
Age	0.009	0.004	0.015
Sex	0.07	0.05	0.09
Constant	0.41	0.29	0.52
Number of Observations	15590		

Note: Model includes data from 1983 to 1985 and includes year fixed effects.

state level instruments are highly correlated with receipt of HAART and are jointly significant ($p < 0.01$). As expected, we find that more generous medically needy eligibility rules are associated with higher likelihood of receipt of HAART. The other two instruments are individually insignificant. They are also jointly insignificant: using a likelihood ratio test or a Wald test, we are unable to reject the null hypothesis that the other two instruments do not affect HAART treatment. In an alternate specification in which we only included the medically needy threshold, we find our results to be almost entirely unchanged. However, in the interest of considering the most general set of policy variables, all the results we report come from a model with all three instruments.

Table 4: Results for the Naïve Models.

Coefficient	Partners		New partners	
	No severity controls	With severity controls	No severity controls	With severity controls
HAART	-0.46	-0.01	-0.14	-0.11
Demographics				
Age	-0.04 **	-0.04 **	-0.35	-0.35 **
Non-white	0.02	0.02	-0.12 *	-0.11 *
Female	-0.37 **	-0.39 **	-0.56	-0.57 **
Education				
Less than HS degree	-0.53 **	-0.54 **	-0.69 **	-0.70 **
High school degree	-0.41 **	-0.43 **	-0.65 **	-0.65 **
Some college	-0.38 **	-0.39 **	0.35 **	-0.35 **
College or more	--		--	
Lowest ever CD4 count (cells/µl)				
>500	--	0.38 **	--	0.23
200-499	--	0.24 **	--	0.27 *
50-200	--	0.27 **	--	0.15
0-50	--		--	
AIDS	--	-0.07	--	0.06
Ordered Probit Cut-offs				
Cut1	-2.51 **	-2.35 **	-1.62 **	-1.40 **
Cut2	-1.38 **	-1.20 **	-1.15 **	-0.93 **

Note: Standard errors account for clustering at the state level
* Statistically significant at 90% confidence level
* Statistically significant at 95% confidence level

Table 5: Estimates from the Structural Models.

	Partners		New partners	
Parameters	HAART	# of partners	HAART	# of new partners
HAART	--	1.294 **	--	1.303 **
Demographics				
Age (years)	-0.003	-0.029 **	-0.003	-0.022 **
Non-white	-0.183	0.096	-0.163	0.021
Female	-0.161 **	-0.211 **	-0.126 **	-0.317 **
Education				
Less than HS degree	-0.266 *	-0.283 **	-0.268 *	-0.336 **
High school degree	-0.237 **	-0.204 *	-0.270 **	-0.297 **
Some college	-0.247 **	-0.173 *	-0.259 **	-0.103
College or more	--	--	--	--
Lowest ever CD4 count (cells/μl)				
>500	-1.150 **	0.855 **	-1.094 **	0.733 **
200-499	-0.432 **	0.397 **	-0.440 **	0.430 **
50-200	-0.379	0.211 *	-0.032	0.119
0-50	--	--	--	--
AIDS	0.047	-0.093	0.054	-0.007
State instruments				
Medically-needed threshold	0.002 *	--	0.002 **	--
AFDC threshold	0.003	--	0.004	--
SSI threshold >65% of FPL	-0.074	--	-0.159	--
Wald test	8.990 **		27.330 **	
Correlation RHO	-1.365 **		-1.625 **	-1.625 **
Ordered Probit Cut-offs				
Delta1		-0.676	0.331	0.257
Delta2		0.871 **		0.331 **

Note: Standard errors account for clustering at the state level
* Statistically significant at 90% confidence level
* Statistically significant at 95% confidence level

Columns 2 and 4 in Table 5 show the results for the sexual activity equations. Table 5 shows several important differences from the naïve results. In contrast to the naïve model, we see that receipt of HAART increases sexual activity among HIV+ patients. This reversal in the effect of HAART is explained by the negative correlation between unobserved sexual propensity and unobserved propensity for treatment with HAART. The most likely cause for this negative correlation is unobserved health — patients with more advanced disease are more likely to receive HAART and less likely to be sexually active. The descriptive statistics in Table 1 and parameter estimates in Tables 4 and 5 show that patients in poorer health are more likely to receive HAART.

Table 6 summarizes the results from the structural model in terms of mean marginal effect of HAART on sexual activity.[1] For instance, the results for number of partners indicate that treatment with HAART reduces the probability of having no partners by 41 percentage points and increases the probability of having 2 or more partners by 30 percentage points. Similarly, receipt of HAART reduces the probability of having no new partners by 39 percentage points and increases the probability of having 2 or more new partners by 33 percentage points. Both of the above marginal effects of HAART on treatment are statistically significant at the 95% level.

[1]Standard errors are calculated by using 1000 bootstrap replications. The bootstrapping accounts for stratified random sampling.

Table 6: Mean Marginal Effect of HAART on Sexual Activity.

	Mean Predicted Probability (Std Dev)		
	0 partners	1 partner	2 or more partners
Number of partners			
HAART	0.21	0.39	0.40
	(0.04)	(0.04)	(0.07)
No HAART	0.62	0.29	0.09
	(0.09)	(0.06)	(0.04)
Marginal Effect	-0.41	0.10	0.31
	(0.13)	(0.03)	(0.10)
Number of new partners			
HAART	0.46	0.12	0.42
	(0.06)	(0.02)	(0.07)
No HAART	0.85	0.06	0.09
	(0.06)	(0.02)	(0.05)
Marginal Effect	-0.39	0.06	0.33
	(0.11)	(0.02)	(0.10)

Note: Mean marginal effects and standard deviations calculated using 1000 bootstrap replications.

6 The Economics of Treatment Breakthroughs

If new HIV treatments save the lives of HIV+ individuals and encourage riskier behavior, the economics of HIV treatment differ from the economics of treating curable disease. First, improving the treatment of an incurable disease can actually increase the level of precautionary behavior, while treatment of a curable disease always lowers such behavior and its associated distortions. The ultimate relationship between treatment breakthroughs and prevalence, however, remains ambiguous, just as it does for a curable disease. From a normative point of view, treatment and prevention of an incurable disease are social complements, but they are substitutes for a curable disease. Since treatment of an incurable disease increases infection risk, it also raises the

marginal social return to prevention. Moreover, since treatment of an incurable engenders negative externalities for the uninfected, there is less private under-provision of treatment in this case.

6.1 Positive Implications for Prevalence and Incidence

For curable diseases, breakthroughs in treatment always lower distortionary preventive behavior, because they lower the cost of being ill and lower the exposure risk by reducing prevalence (cf, Philipson, 2000). For an incurable disease like HIV, however, treatment breakthroughs can actually *increase* distortionary behavior by raising prevalence and exposure risk.

Consider a simple two-stage game. In the first stage, infected people first decide whether or not to receive treatment. Treated individuals survive to the second stage, while untreated individuals do not. In the second period, surviving infected people interact with uninfected ones and thus expose them to the risk of infection.

In the case of a curable disease, treatment lowers the prevalence of the disease in the second stage; in this simple game, prevalence goes to zero, as all the treated individuals are rid of the disease. When a disease is incurable, however, increases in the availability of treatment raise the prevalence of disease faced by uninfected individuals in the second period.

Suppose the share of infected individuals in the initial period is π^1. The proportion treated decreases in the price of treatment p, according to $\gamma(p)$.[2] In period 2, the

[2]For example, there may be heterogeneity in the value of future life that creates a downward

share of infected individuals among the survivors is given by:

$$\pi^2(p, \pi^1) \equiv \frac{\pi^1 \gamma(p)}{1 - \pi^1(1 - \gamma(p))} \tag{6.1}$$

The effect of treatment price on second period prevalence is given by:

$$\pi^2_p(p, \pi^1) = \frac{\pi^1 \gamma'(p)(1 - \pi^1)}{1 - \pi^1(1 - \gamma(p))^2} < 0 \tag{6.2}$$

In the case of an incurable disease, declines in the price of treatment (i.e., breakthroughs in treatment) increase the exposure risk π^2 faced by uninfected individuals in the second period.

The effect of treatment on the precautionary behavior of uninfected individuals is governed by two offsetting incentives: reductions in the price of treatment make it less costly to become infected, but increases in exposure raise the expected pay-off to prevention. Uninfected individuals may choose to enjoy risky behavior in period 2, but this risks infection in their subsequent lives. Suppose that the utility of uninfected individuals increases in risky behavior r, according to $u(r)$. Suppose also that subsequent lifetime utility is v if infection is avoided, but $v - c(p)$ if not; c is the utility cost of being infected and is assumed to rise in the price of treatment. The probability of infection rises in exposure risk π^2 and in risky behavior r, according to $\phi(\pi^2, r)$. A given amount of risk-taking is more hazardous when prevalence is higher,

sloping demand for treatment.

so that $\phi_{\pi r} > 0$. Faced with these incentives, the uninfected individual chooses a level of risky behavior that maximizes his forward-looking utility:

$$\max_r u(r) + \phi(\pi^2(\pi^1, \gamma(p)), r)(v - c) + \left(1 - \phi\left(\pi^2(\pi^1, \gamma(p), r)\right)\right) v \tag{6.3}$$

The optimal level of risk-taking sets the marginal utility of risk-taking equal to its expected marginal cost, according to:

$$u'(r) = \phi_r c \tag{6.4}$$

In the case of a curable disease, reductions in the price of treatment always reduce precautions and thus increase welfare among the uninfected. This result changes with an incurable disease. Comparative statics reveals that:

$$\frac{\partial r}{\partial p} = \frac{\phi_{r\pi}\pi_p^1 - \phi_r c_p}{u''(r) - \phi_{rr} c} \tag{6.5}$$

The conventional effect of treatment is embodied in the second term. Treatment lowers the cost of disease and thus discourages risk-taking. However, there is a unique offsetting effect for an incurable infectious disease that is embodied in the first term. Reductions in the price of treatment encourage precaution by raising the risk of exposure. Treatment saves the lives of infected people who then function as carriers. Therefore, breakthroughs in HIV treatment — unlike breakthroughs against other

curable diseases — can lead to welfare-*reducing* precaution by the uninfected. The ultimate relationship between treatment and risk-taking continues to be ambiguous even in the case of an incurable disease.

The relationship between treatment and risk-taking helps characterize the effect of treatment on the eventual prevalence and incidence of the disease. If expansion in treatment leads to more risk-taking, it will actually raise eventual prevalence. Treatment increases the initial risk of infection for an incurable disease by increasing transmissive behavior of the infected. If it also increases risk-taking behavior, incidence and eventual prevalence must also rise (Figure 1 in the introduction illustrates this case). If, on the other hand, improved treatment lowers eventual prevalence, the distortionary impacts of treatment unique to incurable diseases must be quite significant. Of course, both preventive behavior and prevalence could rise if the reduction in risk-taking is insufficient to offset the increase in initial prevalence.

6.2 Normative Implications for Treatment Provision

The positive relationship between treatment and prevalence for incurable diseases has important normative implications. In the case of an incurable disease, treatment and prevention are unambiguously complementary in the social welfare function. However, in the case of a curable disease, they are unambiguously substitutable. This has particular policy relevance in the context of developing economies, which may receive foreign aid in the form of increased HIV prevention or treatment. A government

that receives free or subsidized HIV treatment faces a steeper marginal return to investments in prevention. On the other hand, free or subsidized treatment for a curable disease, e.g., typhoid, *lowers* the marginal social return to prevention for the poor government.

To study this more precisely, we segment the society into initially infected individuals — who comprise the share of society π^1 — and initially uninfected individuals. Suppose we can represent the set of infected individuals as a single representative agent whose utility is increasing and concave in the proportion treated. $\Omega(\gamma, c)$ is the aggregate welfare of infected individuals as a function of the proportion treated, and the cost of being infected (which includes the monetary cost of treatment). Breakthroughs in treatment lower treatment cost and raise the net marginal utility of treatment, so that $\Omega_{\gamma c} < 0$.

Given Pareto weights λ_1 and λ_2, for infected and uninfected individuals respectively, the socially efficient levels of treatment and risk-taking solve:

$$\max_{\gamma, r} \lambda_1 \Omega(\gamma, c) + \lambda_2 \left(u(r) + \phi(\pi^1, r)(v - c) + (1 - \phi(\pi^1, r))v \right) \tag{6.6}$$

Efficient treatment and risk-taking are characterized by the first order conditions:

$$\lambda_1 \Omega_\gamma(\gamma, c) - \lambda_2 \phi_\pi \pi^1_\gamma c = 0 \tag{6.7}$$

$$u'(r) - \phi_r c = 0 \tag{6.8}$$

The effect of increases in treatment on the marginal social return to prevention is given by

$$-\lambda_2 \phi_{r\pi} \pi^1_\gamma c \tag{6.9}$$

where,

$$sign(-\lambda_2 \phi_{r\pi} \pi^1_\gamma c) = -sign(\pi^1_\gamma) \tag{6.10}$$

The latter expression implies that treatment and risk-taking are complementary for a curable disease, where increases in treatment lower the marginal social return to prevention. Conversely, treatment and risk-taking are substitutes for an incurable disease.

There is one other point worth noting from equation 6.7. In a simple economy where infected individuals are in control of their treatment levels, the privately efficient level of treatment satisfies $\Omega_\gamma = 0$. This implies that there is private under-provision of treatment for a curable disease, where $\pi^1_\gamma < 0$, but private *over*-provision of treatment for an incurable one where the opposite is true. This result might be too strong, since there may be other positive externalities from treatment that we are not considering here, such as spillovers in the labor market. The more general statement is that incurable diseases involve less under-provision of treatment than curable ones.

Finally, our analysis suggests that the public health goal of reduced prevalence by expanding treatment might be at odds with the goal of improved social welfare, as suggested by Philipson (2000). Improved treatment has the initial effect of rais-

ing prevalence and thus encouraging distortionary precautions by the uninfected. If prevalence ends up falling with treatment, this implies merely that distortionary precautions are significant enough to offset the initial effect of treatment on prevalence. Therefore, reductions in prevalence might suggest worse welfare outcomes, not better ones.

7 Conclusions

Treating an incurable infectious disease involves unique issues for individual behavior and social policy. While improvements in treatment clearly improve the welfare of infected individuals, societies must cope with the reality that treatment can fuel the further spread of the disease. We have presented empirical evidence consistent with this argument and have developed a few of the implications this has for behavior and the optimal provision of treatment. It appears that recent advances in treatment may have allowed HIV-sufferers to engage in more risky behavior and thus further the spread of HIV.

Our results suggest that a cure for HIV, or other incurable infectious diseases, would have unique welfare benefits, because it would break the link between treatment and risky behavior by infected individuals. Curative treatment generates positive external effects by reducing disease exposure for the uninfected. This creates unique R&D externalities in the search for a cure, which involves not just positive externalities, but the *elimination* of negative externalities. The total external effect

is thus larger than the positive externality alone would suggest.

References

Bound, J., D. Jaeger, and R. Baker (1995). "Problems With Instrumental Variables Estimation When the Correlation Between the Instruments and the Endogenous Explanatory Variable is Weak." *Journal of the American Statistical Association* 90(430):443–450.

Center for Disease Control (1998). "Diagnosis and Reporting of HIV and AIDS in States with Integrated HIV and AIDS Surveillance — United States, January 1994 - June 1997." *Morbidity and Mortality Weekly Report* 47(15):309–314.

Detels, R., et al. (1998). "Effectiveness of potent antiretroviral therapy on time to AIDS and death in men with known HIV infection duration. Multicenter AIDS Cohort Study Investigators." *Jama* 280(17):1497–503.

Fauci, A., et al., eds. (1998). *Harrison's Principles of Internal Medicine, 14th Edition.* New York: McGraw Hill.

Freedberg, K. A., et al. (2001). "The cost effectiveness of combination antiretroviral therapy for HIV disease." *N Engl J Med* 344(11):824–31.

Hammer, S. M., et al. (1997). "A controlled trial of two nucleoside analogues plus indinavir in persons with human immunodeficiency virus infection and CD4 cell counts of 200 per cubic millimeter or less. AIDS Clinical Trials Group 320 Study Team." *N Engl J Med* 337(11):725–33.

Lucas, G. M., R. E. Chaisson, and R. D. Moore (1999). "Highly active antiretroviral therapy in a large urban clinic: risk factors for virologic failure and adverse drug reactions." *Ann Intern Med* 131(2):81–7.

Lucas, G. M., R. E. Chaisson, and R. D. Moore (2003). "Survival in an urban HIV-1 clinic in the era of highly active antiretroviral therapy: a 5-year cohort study." *J Acquir Immune Defic Syndr* 33(3):321–8.

Nelson, C. R., and R. Startz (1990). "Some Further Results on the Exact Small Sample Properties of the Instrumental Variable Estimator." *Econometrica* 58(4):967–976.

Palella, F., et al. (1998). "Declining Morbidity and Mortality Among Patients with Advanced Human Immunodeficiency Virus Infection." *New England Journal of Medicine* 338(13):853–60.

Philipson, T. (2000). "Economic Epidemiology and Infectious Diseases." *Handbook of health economics. Volume 1B.* A. J. Culyer and J. P. e. Newhouse, eds. New York: Elsevier Science. 1761–1797.

Staiger, D., and J. H. Stock (1997). "Instrumental Variables Regression with Weak Instruments." *Econometrica* 65(3):557–586.

Staszewski, S., et al. (1999). "Efavirenz plus zidovudine and lamivudine, efavirenz plus indinavir, and indinavir plus zidovudine and lamivudine in the treatment of HIV-1 infection in adults. Study 006 Team." *N Engl J Med* 341(25):1865–73.

Vittinghoff, E., et al. (1999). "Combination antiretroviral therapy and recent declines in AIDS incidence and mortality." *J Infect Dis* 179(3):717–20.