INTERNATIONAL ECONOMIC REVIEW
Vol. 62, No. 3, August 2021 DOI: 10.1111/iere.12505

INNOVATION AND DIFFUSION OF MEDICAL TREATMENT*

By Barton H. Hamilton, Andrés Hincapié, Robert A. Miller, and Nicholas W. Papageorge

Olin Business School, Washington University in St. Louis, U.S.A.; Department of Economics, University of North Carolina at Chapel Hill, U.S.A.; Tepper School of Business, Carnegie Mellon University, U.S.A.; Department of Economics, IZA and NBER, Johns Hopkins University, U.S.A.

We develop and estimate a dynamic structural model of demand in a setting where product characteristics endogenously evolve in response to aggregate consumer choices. The direction and speed of innovation are inefficient because individuals do not account for their influence on innovation, creating an externality. Our application focuses on drugs invented to combat human immunodeficiency virus; they differ in their efficacy and propensity to cause side effects. We find that the externality is quantitatively important; temporarily subsidizing the experimental treatment would have increased average social welfare by improving average health and would have reduced inequality in lifetime utility across health groups.

1. INTRODUCTION

Economists have long recognized that innovation, including the entry of new products and the exit of obsolete ones, is not only determined by science and luck, but also responds to latent consumer demand (Hicks, 1932). Sometimes referred to as demand-pull innovation (Schmookler, 1966; Scherer, 1982), the responsiveness of innovation to demand generates an externality because the benefits an individual indirectly confers upon all (other) future individuals through his effect on innovative activity are not reflected in the price he pays for the product in the decentralized economy (Jovanovic and MacDonald, 1994; Waldfogel, 2003; Finkelstein, 2004). This article develops and estimates a dynamic structural model of demand that endogenizes how consumer choices affect product innovation; we quantify the magnitude of this externality in a rapidly evolving medical treatment market for human immunodeficiency virus (HIV), where innovation was largely driven by experimental clinical trials, and conduct a counterfactual analysis of subsidizing such trials to improve social welfare.

The data set for this study extracts observations from a biennial panel from four American cities that tracks a replenished panel of individuals along with the path of innovations in HIV treatments for over 20 years, from when this market emerged around 1984, until it matured. During this period, frequent incremental innovations in medication were punctuated by sporadic breakthroughs, and occasionally new inferior treatments. Up to seven treatments entered the market in any given time, giving consumers a choice between multiple commercially available treatments with differential characteristics, or participating in clinical trials offering experimental treatments. The data include an objective, continuous measure of health

*Manuscript received April 2020; revised November 2020.

This article benefited from presentations in economics departments at Arizona State, Central Florida, Houston, KU Leuven, Minnesota, NYU, Sydney, Toronto, UCLA, Washington, Wisconsin - Madison, and Yale, and in business schools at Carnegie Mellon, Harvard, Stanford, UNC, and Washington University in St. Louis, as well as conference presentations at the NBER, AHEW, SEA, SITE, and the Federal Reserve Board NYC. Please address correspondence to: Andrés Hincapié, Department of Economics, University of North Carolina at Chapel Hill, U.S.A. E-mail: *andres.hincapie@unc.edu*.

based on immune system status obtained from blood tests administered with the survey every six months. They show that consumers exhibit substantial heterogeneity measured in terms of their health and demographics, including age, race, education, labor force participation, and income. These factors affect their treatment of choice and hence aggregate demand. Section 2 provides a brief historical background and motivates the model structure with patterns observed in the data.

Section 3 describes the generic model of dynamic demand and product innovation upon which our application is based. In the model, consumers choose between taking a commercially available treatment, joining a clinical trial to access an experimental treatment, or not being treated. Their choices reflect a trade-off between multiple treatment attributes: efficacy, which improves long-run health, and current-period side effects. Each period, new treatments become available and treatments in low demand exit the market. The number and type of new treatments are governed by a stochastic process parameterized by the shares of consumers who opt for available treatments and those who opt for clinical trials. When making private medication choices, individual consumers do not take into account their impact on aggregate demand, generating an externality. This externality complicates the computation of equilibrium: it cannot be recast as a dynamic programming problem solved by a social planner, and in addition is rendered nonstationary by ongoing technological progress.

Section 4 adapts the generic model to our empirical application, by modeling the uncertainty consumers face and how it is resolved, along with the parameterization used in estimation. The empirical model includes the choice to participate in clinical trials; the treatment characteristics of clinical trials evolve over time with technological progress. Given the myriad of commercial treatments available each period, we use a clustering algorithm to group treatments with similar characteristics into clusters whose composition changes over time, as new drugs enter the market and others are withdrawn for lack of demand. Consumers opting for a commercially available treatment randomly select from their cluster of choice, knowing only the distributional characteristics of the cluster. They learn the treatment characteristics of the specific treatment they select after one (six month) period of experience; therefore, in the choice set, we distinguish between the first time a consumer takes a given treatment, and repeat prescriptions for that particular treatment. Our clustering approach allows for aggregate demand to speed and sway future innovation while addressing the challenge of a large and changing choice set in a finely differentiated market, a challenge faced by consumers but also by researchers in marketing and industrial organization. We use the model to focus on technological progress in treating HIV/acquired immunodeficiency syndrome (AIDS), and, in particular, the role of clinical trials.

The parameter estimates of the model are presented in Section 6. The model tracks the effects of technological progress on the various components of demand quite well; the model interprets highly active antiretroviral therapy (HAART), the major advance in medical treatment of HIV within the sample frame, as a relatively unlikely event. Our results reveal that a strong distaste for experimentation slows the diffusion of new, superior treatments as well as the development of future treatments in clinical trials.

We quantify the magnitude of the demand-pull externality for experimental treatments at two different points in time, before and after HAART was developed. We compute for both 1991 and 1996, the one-period subsidy, and its lump-sum total cost, which attains the market share of experimental treatments that maximizes expected utility averaged across the HIV+ population. We find that optimally subsidizing experimental treatment after HAART produces considerably greater social benefits than beforehand, and the improvement in value of life from subsidizing experimental treatment in 1996 constitutes about two-thirds of the improvement in average value of life due to technological progress between 1991 and 1996. We also find that the sickest benefit the most from correcting the externality.

This study contributes to a literature on dynamic demand under uncertainty. Following Petrin (2002), treatments are modeled as bundles of characteristics, with dynamic impacts on consumers, as in Gowrisankaran and Rysman (2012). Several other studies apply dynamic

models of demand to health-related choices, and like ours have: learning (Crawford and Shum, 2005; Chan and Hamilton, 2006; Fernandez, 2013; Darden, 2017; Dickstein, 2018); consumer experimentation with new products (Chan and Hamilton, 2006; Fernandez, 2013); side effects, including the effect on labor force participation (Papageorge, 2016); and the equilibrium implications of individual choices that affect the health of others (Chan et al., 2016). Previous work also investigates how demand affects innovation. For example, Finkelstein (2004) shows that policies promoting vaccine use accelerate the development of vaccines; Acemoglu and Linn (2004) and Dubois et al. (2015) relate market size to pharmaceutical innovation using the size of the U.S. market and a measure of the size of the global market, respectively. Dranove et al. (2014) identify a social value of pharmaceutical innovation, showing that Medicare Part D spurred the development of some drugs. Waldfogel (2003) describes the mechanism through which market shares can influence products, thus benefitting individuals with similar tastes. Bolton and Harris (1999) argue that a free-riding problem emerges if experimentation accelerates innovation. Our contribution is to model the externality arising when an evolving choice set is endogenous to consumer demand, provide structural estimates quantifying its importance, and evaluate counterfactual policies designed to mitigate the effects of the externality. Our structural approach allows for counterfactual measurement of the effect of consumer choice on future innovation. Within this structured context, the emphasis we place on clinical trials as instruments driving innovation is also new to the literature.

Our empirical strategy builds on Hotz and Miller (1993), Hotz et al. (1994), and Altuğ and Miller (1998) in using conditional choice probabilities (CCPs), and forward simulation techniques to incorporate how individuals form expectations about future innovations. In our context, the individual's choice set evolves stochastically as a function of endogenous product exit and entry. Exit occurs when demand falls below threshold bounds, whereas entry is determined by the innovation process that contains two components: unexpected, aggregate supply shocks and a systematic component, endogenous to aggregate demand, captured by a multidimensional reference point for innovation.

Not modeling the primitives that drive firm decisions severely limits the scope for analyzing counterfactuals. Although we model an equilibrium supply curve that includes the relevant state variables determining the supply response in equilibrium, we cannot predict how firms would react to different policies when given the opportunity. As mentioned above we do, however, predict the size of the externality and how the market would react to an unanticipated temporary shock to policy.

Two recent papers estimating structural models of firm rivalry analyze how market structure affects technological advance. Goettler and Gordon (2011) estimate a model of duopoly in the market for microprocessors assuming unidimensional product quality, that innovations are a fixed positive amount, and allowing for up to two innovations to be introduced each period. Igami (2017) studies the market for hard disk drives, often considered a durable good, but modeled here as a single, nondifferentiated, unidimensional product for which demand, perfectly forecast by firms, does not depend on past purchase behavior; products are upgraded at most once per period; each period firms move in a predetermined invariant order observing all the past moves, as part of a perfect information game with a fixed finite horizon fully anticipated by all the players.

The complexity of the HIV treatment industry and the nature of our data are critical factors determining our modeling approach. Table 1 presents evidence of this complexity: at least 11 firms produced HIV drugs over the period, some of the firms resulting from mergers.[1] Most treatments are combinations using product components developed by up to four pharmaceutical firms.[2] Product quality is multidimensional; the size of innovations varies between

[1] For example, the first product component (AZT) was introduced by Burroughs–Wellcome in 1987, which became Glaxo–Wellcome in 1995, GlaxoSmithKline in 2000, and transferred its HIV assets to the joint venture ViiV created with Pfizer in 2009. By 1995, at least six firms had introduced product components and had valid patents (Glaxo–Wellcome, Bristol–Myers Squibb, Hoffmann-La Roche, Abbott, Merck, and Boehringer Ingelheim).

[2] The long name and chemical composition of product components are displayed in Table A.2 in Appendix A.1.

TABLE 1
TREATMENTS

Treatment	Companies	Entry	Exit	Treatment	Companies	Entry	Exit
AZT	BW	87-1	–	DDI/D4T/NVP	BMS/BMS/BI	97-2	-
IFNs (α/β)/AZT	G/BW	87-2	95-2	DDI/3TC/NFV	BMS/GW/AG	97-2	-
AL-721		87-2	91-2	DDI/D4T/EFV	BMS/BMS/DP	98-2	08-1
AZT/ACV	BW/BW	89-2	00-1	3TC/ABC/EFV	GW/GW/DP	98-2	-
ACV	BW	89-2	00-1	AZT/NVP/3TC/ABC	GW/BI/GW/GW	99-1	-
AZT/ACV/DDI	BW/BW/BMS	90-1	97-1	AZT/3TC/ABC/EFV	GW/GW/GW/DP	99-1	-
ACV/DDI	BW/BMS	90-1	00-1	AZT/3TC/EFV	GW/GW/DP	99-1	-
AZT/DDC	BW/H	90-1	00-2	AZT/3TC/ABC	GW/GW/GW	99-1	-
AZT/DDI	BW/BMS	90-1	04-2	D4T/3TC/EFV	BMS/GW/DP	99-1	06-1
DDI	BMS	90-1	–	NVP/3TC/ABC	BI/GW/GW	99-2	-
AZT/DDC/ACV/DDI	BW/H/BW/BMS	91-1	97-1	D4T/3TC/LPV/RTV	BMS/GSK/AB/AB	01-1	06-1
AZT/DDC/ACV	BW/H/BW	91-1	99-2	3TC/LPV/RTV/ABC	GSK/AB/AB/GSK	01-2	-
AZT/DDC/DDI	BW/H/BMS	91-1	95-2	AZT/3TC/LPV/RTV	GSK/GSK/AB/AB	01-2	-
DDC/ACV	H/BW	91-1	97-2	AZT/3TC/LPV/RTV/ABC	GSK/GSK/AB/AB/GSK	02-1	-
DDC	H	91-1	99-1	3TC/ABC/EFV/TDF	GSK/GSK/BMS/GI	02-1	-
D4T	BMS	93-1	–	AZT/3TC/ABC/TDF	GSK/GSK/GSK/GI	02-1	-
AZT/ACV/3TC	GW/GW/GW	94-2	00-1	AZT/3TC/LPV/RTV/TDF	GSK/GSK/AB/AB/GI	02-1	-
AZT/3TC	GW/GW	95-1	–	NVP/3TC/TDF	BI/GSK/GI	02-1	07-1
ACV/D4T/3TC	GW/BMS/GW	95-2	00-1	3TC/LPV/RTV/TDF	GSK/AB/AB/GI	02-1	-
AZT/3TC/SQV	GW/GW/H	96-1	05-1	LPV/RTV/EFV/TDF	AB/AB/BMS/GI	02-1	-
D4T/3TC	BMS/GW	96-1	–	3TC/EFV/TDF	GSK/BMS/GI	02-1	-
AZT/3TC/SQV/RTV	GW/GW/H/AB	96-2	–	AZT/3TC/LPV/RTV/ABC/TDF	GSK/GSK/AB/AB/GSK/GI	02-2	-
AZT/ACV/3TC/IDV	GW/GW/GW/M	96-2	00-1	DDI/LPV/RTV/TDF	BMS/AB/AB/GI	02-2	-
ACV/D4T/3TC/IDV	GW/BMS/GW/M	96-2	00-1	DDI/EFV/TDF	BMS/BMS/GI	02-2	-
AZT/3TC/RTV/IDV	GW/GW/AB/M	96-2	06-2	ABC/EFV/TDF	GSK/BMS/GI	02-2	-
D4T/3TC/RTV/IDV	BMS/GW/AB/M	96-2	06-2	LPV/RTV/ABC/TDF	AB/AB/GSK/GI	02-2	-
D4T/3TC/SQV/RTV	BMS/GW/H/AB	96-2	04-2	3TC/RTV/ABC/ATV	GSK/AB/GSK/BMS	03-2	-
DDI/D4T/IDV	BMS/BMS/M	96-2	04-2	EFV/TDF/FTC	BMS/GI/GI	03-2	-
D4T/3TC/IDV	BMS/GW/M	96-2	08-1	RTV/EFV/TDF/FTC/ATV	AB/BMS/GI/GI/BMS	04-1	-
AZT/3TC/IDV	GW/GW/M	96-2	–	3TC/RTV/ABC/TDF/ATV	GSK/AB/GSK/GI/BMS	04-1	-
D4T/NVP/3TC	BMS/BI/GW	97-1	–	DDI/RTV/TDF/ATV	BMS/AB/GI/BMS	04-1	-
AZT/NVP/3TC	GW/BI/GW	97-1	–	RTV/TDF/FTC/ATV	AB/GI/GI/BMS	04-1	-
AZT/3TC/NFV	GW/GW/AG	97-1	–	NVP/TDF/FTC	BI/GI/GI	04-1	-
DDI/D4T/NFV	BMS/BMS/AG	97-1	05-2	LPV/RTV/TDF/FTC	AB/AB/GI/GI	04-2	-
D4T/3TC/NFV	BMS/GW/AG	97-2	–	RTV/TDF/FTC/FPV	AB/GI/GI/GSK	05-1	-

NOTES: Entry and exit dates are displayed in format year-semester (YY-S). Many treatments had not exited by the end of the sample. Entry and exit dates are obtained using the algorithm in Appendix A.1.1. The *Companies* column displays the firms who owned each drug component at the moment of treatment entry. Company acronyms are: Burroughs–Wellcome (BW), Glaxo (G), Bristol–Myers Squibb (BMS), Hoffmann–La Roche (H), Glaxo Wellcome (GW), Abbott (AB), Merck (M), Boehringer Ingelheim (BI), Agouron (AG), DuPont (DP), GlaxoSmithKline (GSK), and Gilead (GI). The long name and chemical composition of product components are displayed in Table A.2 in Appendix A.1.

firms in any given period and over time within a firm; there is aggregate uncertainty about the direction of the quality vector. Finally, our data contain detailed information about individual demographics, treatment choices, and their effects, but much less information about how firms make their research, development, and marketing decisions. A challenge for future researchers seeking to investigate counterfactuals more extensively is to estimate a structural model that combines a rich dynamic demand structure derived from consumer preferences estimated off panel data with an equilibrating supply side mechanism based on the evolving technologies that define the firm primitives.[3]

2. DATA

Our empirical application focuses on the market for HIV treatments that came into existence around 1984 with the beginning of the HIV pandemic, which had caused over 613,000 deaths in the United States by 2009.[4] HIV infection leads to a reduction in the ability of the immune system to fight off routine infections, a condition known as AIDS. In developed countries, where access to medication is widespread and often subsidized, technological advancement has transformed HIV infection into a manageable condition with treatments whose side effects are fairly mild. This was not always the case. In the early years of the epidemic, available treatments were not only largely ineffective, but also had uncomfortable, painful and even deadly side effects. Over time many innovations appeared, most of them small, but in the mid-1990s, a new set of treatments collectively known as HAART was introduced, transforming HIV from a virtual death sentence into a chronic condition.[5] Within two years, mortality rates fell by over 80% among HIV-infected (HIV+) men (Bhaskaran et al., 2008). While the first versions of HAART included drugs that were highly toxic, driving some people to refrain from using them to avoid often intolerable side effects, innovations after the mid-1990s reduced the adverse side effects.

We extract data from the Multi-center AIDS Cohort Study (MACS), an ongoing longitudinal survey of HIV infection in men who have sex with men (MSM) that began in 1984. It is conducted at four large U.S. sites: Baltimore, Chicago, Pittsburgh, and Los Angeles. We cannot formally establish that the sample is geographically representative of the market of MSM in the United States. However, the urban nature of the sample captures the fact that gay men, a group largely overlapping with the set of MSM, were mostly located in selected urban areas (Black et al., 2000, 2002). Within urban areas where gay men concentrate, Los Angeles and Chicago ranked high in both the concentration of gay couples and the seroprevalence of HIV, whereas Baltimore and Pittsburgh tended to rank low (Holmberg, 1996; Black et al., 2002). Although the MACS data are drawn from the United States, previous work suggests that the United States is often the primary target market for pharmaceutical innovation, in particular for HIV drugs (Goldman, 2018), because of its size (OECD, 2017; Sarnak et al., 2017), and lack of price restrictions (Danzon et al., 2005; Kyle, 2007). Concordantly, at least 80% of the product components in Table 1 were first launched in the United States.[6]

At each semiannual visit, survey data are collected on HIV+ men's health status, their treatment decisions (including their participation in clinical trials), out-of-pocket expenditures for prescription medication (antiretrovirals or otherwise), and physical ailments (which can

[3] Chen et al. (2013) calibrate, for the U.S. automobile market, a model of new and used goods in a steady-state Markov perfect equilibrium where consumers are forward-looking and firms with the same unit costs produce durable goods that depreciate over time.

[4] For comparison, over the same period in the United States, there were 508,000 homicides and U.S. deaths in World War II were just under 420,000. Globally, the number of deaths due to HIV/AIDS stands at roughly 35 million. Currently, roughly 50,000 new infections and 13,000 deaths per year in the United States are attributed to HIV/AIDS.

[5] Two crucial clinical guidelines that comprise HAART became commonly accepted in 1996. First, the usage of protease inhibitors (made widely available toward the end of 1995) as an effective HIV treatment. Second, the usage of several antiretroviral (ARV) drugs simultaneously to delay the onset of AIDS indefinitely.

[6] Based on evidence from PharmaProjects (a longitudinal drug development database) and Internet web searches we conducted, 16 out of 20 of the product components in Table 1 were first launched in the United States.

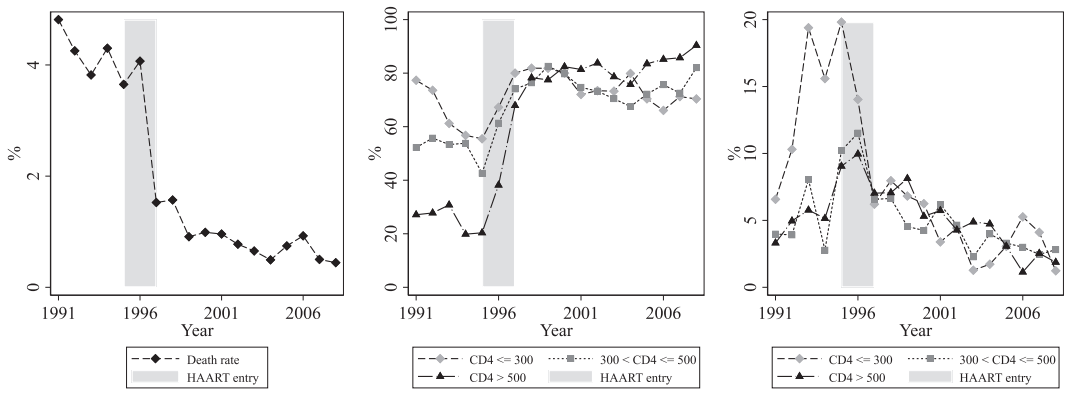
reflect drug side effects), along with sociodemographic information such as labor supply, income, race, and education. The survey instrument is updated every six months, so any new treatment is immediately included. Hence, we use the treatment choice data to construct the life cycle of each treatment, including its sample market share. The main treatments and their life cycles are shown in Table 1. Most are combinations comprising a collection of product components that very often are produced by different pharmaceutical firms. Table 1 also shows that our data capture the onset of the market with its first drug, AZT in 1987, as well as the development of combinations still in use, such as Atripla (EFV/TDF/FTC) and Truvada (TDF/FTC).

Our measure of health, the cluster of differentiation 4 (CD4) count, is based on the immune system; it is defined as the number of white blood cells per cubic millimeter of blood, obtained from blood tests that are administered at each visit. Absent HIV infection, a normal count ranges between 500 and 1,500. For HIV+ individuals, a count below 500 indicates that the immune system has begun to deteriorate. However, such individuals may remain asymptomatic. When the CD4 count drops below a threshold of between 200 and 250, a patient is said to suffer from AIDS: his immune system fails to fight off routine infections, compromising his survival probability. Few data sets contain such objective, continuous measures of health and detailed treatment data, along with economic information. This commends the MACS data set to analyzing demand-pull innovation in the medical treatment market. A drawback of the MACS data is that it lacks information on treatment prices; our empirical work approximates the cost of treatment using out-of-pocket expenditures after controlling for health and other observables.

The MACS data set, replenished over time due to attrition, contains information on 6,972 subjects at 49 semiannual visits for a total of 111,271 observations between 1984 and 2005, in the form of subject visits. We limit our attention to the observations on HIV+ individuals without missing information in relevant variables. Lacking data on gross income and out-of-pocket expenditures at earlier visits, we use two samples. The larger one (20,142 observations) covers visits 6 (87-1, year 1987 semester 1) to 49 (08-2), and only includes health status, ailments, and product usage; the smaller sample (16,851 observations) starting at visit 14 (91-1) contains all the variables. The smaller sample comprises 1,719 males, 68% white, 22% black, and the rest Hispanic; 86% received some secondary education or more, and 23% attended graduate school. Underscoring the gravity of HIV infection, about 40% of the subjects in our sample die prior to the end of the sample period. Appendix A.1 describes how both samples are constructed. In the remainder of this section, we describe the panel data further and document key patterns that motivate the main features of our framework.

2.1. *HAART Has a Dramatic Impact.* Figure 1(a) shows that when HAART is introduced death rates plunge, and continue to fall until 2007, as smaller innovations occurred that made drugs incrementally more effective and less toxic. Table 2 shows that improvements in survival coincide with improvements in immune system health as measured by the CD4 count. Similarly, the proportion of the population experiencing aliments declines 4 percentage points. The sample average age rises from 41 to 47 years old due to aging and lower mortality rates. Accompanying the changing composition of this aging sample, gross income falls from about 19,000 in the pre-HAART era to 16,500 after 1995 in real $U.S. indexed to 2000. There is substantial variation in labor supply both within and between individuals; 74% (68%) of them work (do not work) in at least one period and labor force participation falls by 12 percentage points in the post-HAART era.

2.2. *Individuals Respond to Technological Change.* The sample response to the introduction of HAART is equally noteworthy. On the one hand, Table 2 shows that the share of observations (subject-visits) of individuals consuming a commercially available treatment increases 50% from the pre-HAART era to post-HAART. On the other hand, the share of



NOTES: Left panel shows the probability of dying between periods t and $t+1$ conditional on surviving until t. More than 1,500 surveyed individuals died for AIDS-related causes during our analysis period. The middle and right panels show consumption by health status.

FIGURE 1

SURVIVAL AND CONSUMER DEMAND OVER TIME

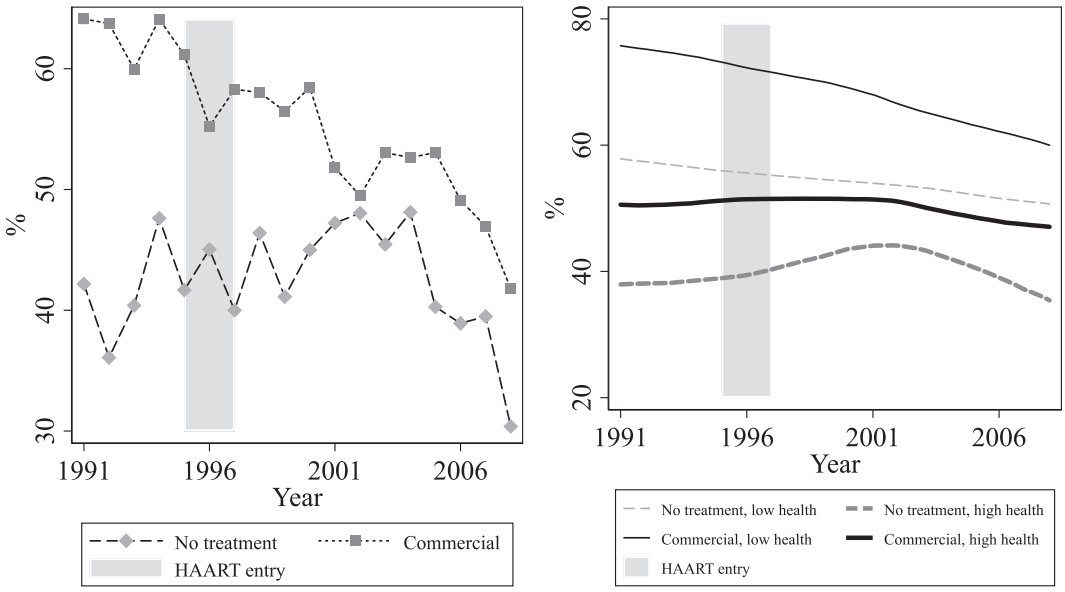
TABLE 2
SUMMARY STATISTICS: SUBJECT VISITS, 1990–2007

	Sample	Pre-HAART ($t <$ 96-2)	Post-HAART ($t \geq$ 96-2)
Ailments	0.43	0.45	0.41
Commercial treatment	0.65	0.49	0.76
Experimental treatment	0.07	0.09	0.05
Labor participation	0.63	0.70	0.58
Age	44.48	40.89	47.01
	(8.03)	(6.99)	(7.75)
CD4	475	407	524
	(297)	(298)	(287)
Gross income	17567	19036	16531
	(8787)	(8733)	(8677)
Out-of-pocket expenditures	266	179	327
	(706)	(598)	(767)
Observations	16851	6972	9879

NOTES: Standard deviation in parentheses. Gross income and out-of-pocket expenditures are semestral and measured in real $U.S. indexed to 2000. Post-HAART starts in the second semester of 1996 ($t \geq$ 96-2). All differences between pre- and post-HAART eras are statistically significant at the 0.01 level. Table A.1 in Appendix A.1 presents descriptives by treatment chosen.

observations of individuals consuming experimental treatments almost halves. Mean out-of-pocket medical expenditures almost doubles, despite the lower incomes on average.

Treatment varies across individuals, and also over time for the same individual; 83% of unique individuals are observed using a commercial treatment at least once and 24% participate in at least one clinical trial. Some of this variation is tied to the individual health and the quality of different treatments. For example, Figure 1(b) shows consumption of commercially available treatments differed across health levels prior to the introduction of HAART. Individuals with low CD4 counts were more likely to use commercially available, relatively ineffective treatments, whereas healthier individuals often avoided treatment altogether. Demand for treatment increased and converged across health levels in response to the introduction of more effective products after HAART.



(a) By Treatment Usage (b) By Treatment Usage and CD4

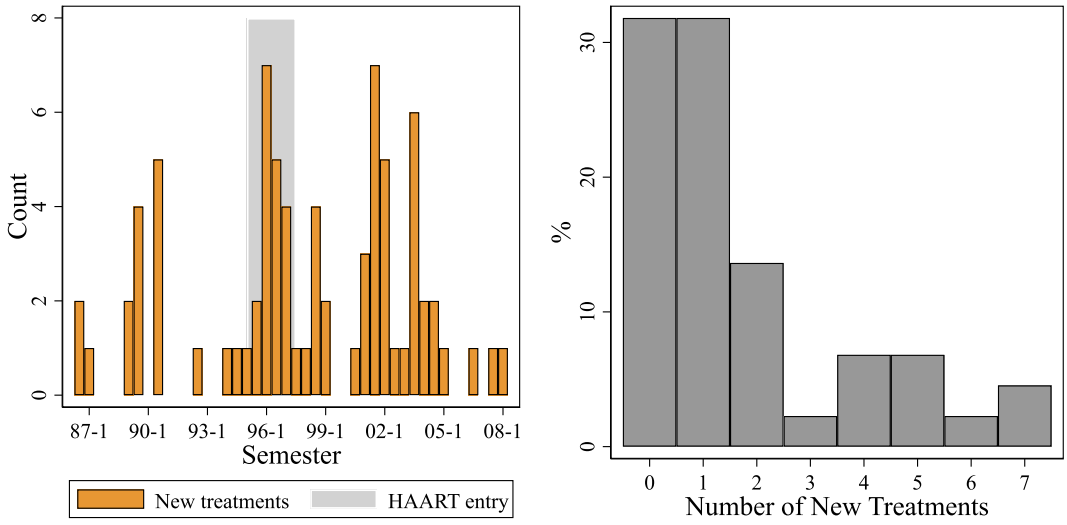
NOTES: This figure displays the mean of the ailments indicator over time. *Commercial* refers to individuals consuming a commercial treatment. *High (low) health* refers to individuals with CD4 counts greater than or equal to (less than) 250. Right panel uses local polynomials to smooth the series.

FIGURE 2

PHYSICAL AILMENTS BY TREATMENT USAGE AND CD4

A striking feature of Figure 1(c) is the spike in trial treatment around the time HAART was introduced. Early trial participation is driven largely by individuals with low CD4 counts, but once effective treatments become available, their participation almost evaporates. Numerous studies document the difficulties recruiting HIV drug trial participants (see, e.g., Mills et al., 2006), especially once commercially available treatments improved after HAART was introduced (Brown et al., 2006; Malani and Philipson, 2011), generating widespread concern that selective participation or attrition could bias study results or slow innovation. Malani and Philipson (2011) provide evidence that the precipitous drop in trial participation among HIV+ men is indeed due to shifts in the demand for trial slots. Increased NIH HIV research funding over these years, a proxy for available trial slots, which we document in Figure A.1 in Appendix A.1.2, corroborates this view. We conclude the drop in trial participation is likely to be due to a decline in the demand for, instead of the supply of, trial slots.

2.3. *Product Characteristics Are Multidimensional.* Figure 1(b) shows that immediately following the introduction of HAART, the proportion of HIV+ individuals taking commercially available treatments rises but it climbs only to roughly 80%, even among those who are most severely affected by HIV. Given its life-sustaining benefits, this seems odd: the treatments are costly, but out-of-pocket costs for medical care do not vary much across treatment alternatives. Figure 2(a) provides support for another explanation, that individuals avoid effective medications because of their side effects. It shows that those who consume a commercial treatment suffer more physical ailments (such as nausea or cramping). Figure 2(b) shows that this relationship holds even after controlling for the underlying immune system health. The consumption patterns are consistent with the notion that treatments are multiattribute products: drug efficacy that improves underlying health, and the propensity for a drug to cause side effects compromising quality of life.



(a) Time Series (b) Empirical Distribution

NOTES: Dates are in format year-semester. Panel 3(a) displays the number of new treatments over time. Panel 3(b) displays the empirical distribution of the number of new treatments obtained from the time series.

FIGURE 3

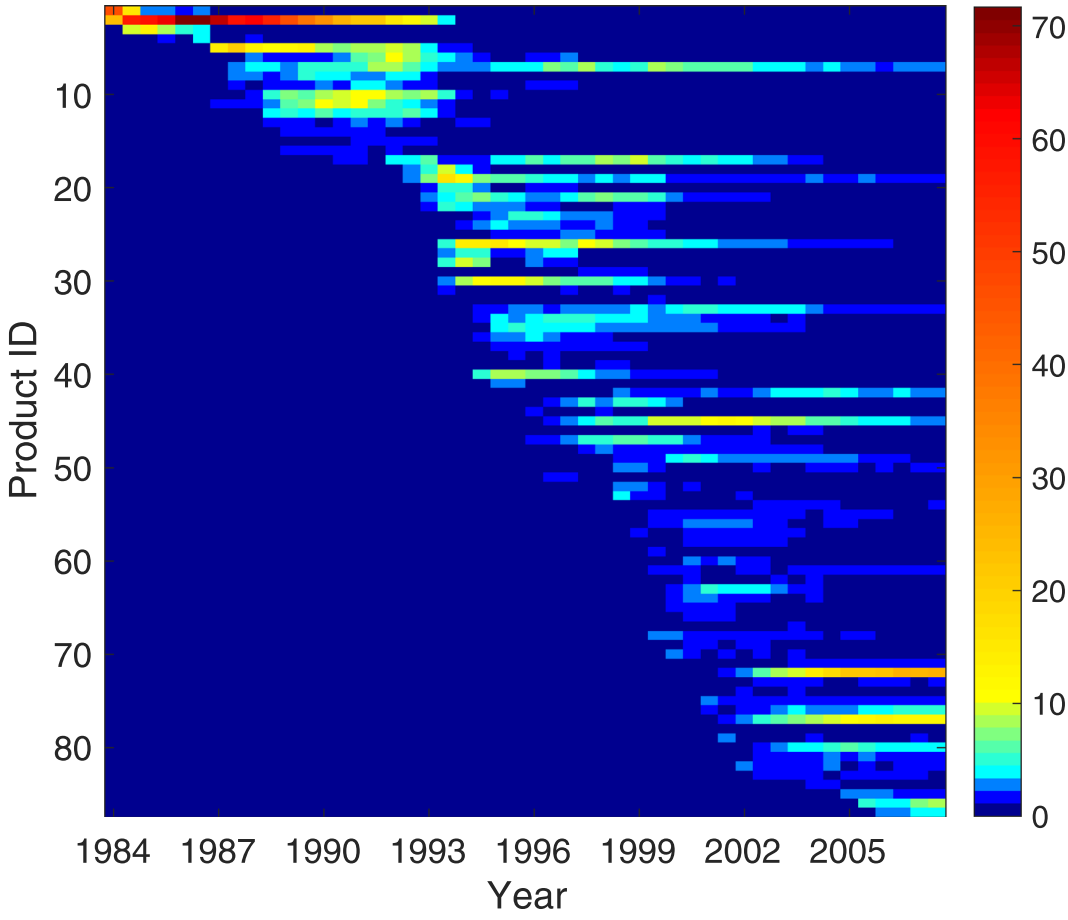
NUMBER OF NEW TREATMENTS
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

2.4. *The Number of New Treatments Fluctuates over Time.* We define a treatment as a combination of product components. (See Table 1 and Appendix A.1.) For instance, AZT and the combination AZT/3TC/ABC are distinct treatments. This definition corresponds to the nature of the market, where large treatment innovations such as HAART are themselves combinations of product components. Because there are complementarities between product components, the physiological response from treatment AZT/DDI is not equal to the sum of the physiological responses of AZT and DDI when taken separately. By this definition, 86 treatments are introduced to the market over the sample period with substantial variation in the number of new treatments introduced each period shown in Figure 3(a). Figure 3(b) shows that the unconditional probability of observing more than one treatment entering in a given period is more than 30% .

2.5. *Market Concentration Fluctuates with Innovation.* Substantial variation in the number of new treatments (Figure 3) along with consumer preferences for multiple dimensions of drug quality is reflected in both innovation and market concentration. Figure 4 shows innovation and diffusion of new products over time using a heat map: dark colors correspond to low (or 0) market shares, whereas warmer colors indicate higher market shares. In the mid-1980s a few treatments command high shares. As time passes new treatments strip market share from incumbents and less popular treatments exit. Low market shares are common after HAART is introduced around 1995: many new treatments are introduced, and most of them are effective but have strong side effects. As the market matures, effective treatments with fewer side effects become commercially available, mitigating the trade-off between the two treatment qualities, and increasing market concentration once again.

2.6. *Innovation Is Guided by Demand.* We also find evidence that innovation responds to consumer demand. Figure 5 illustrates the process of innovation with snapshots of the evolving market captured from our animated appendix.[7] Each snapshot plots treatment

[7] https://www.dropbox.com/s/2icr4dxrpx9metk/treatmentevolutionNew.mp4?dl=0.



NOTES: HIV treatments from 1984 to 2008. Each ID—or row—represents a treatment. Color indicates the share of the market that the treatment captures. Shares are conditional on individuals who consumed a treatment.

FIGURE 4

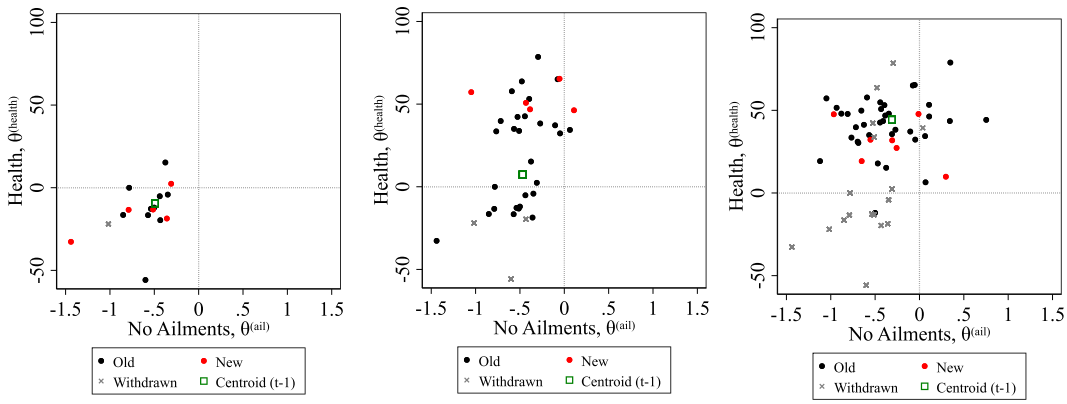
DIFFUSION OF TREATMENTS OVER TIME
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

characteristics, effectiveness, and lack of side effects, on the two axes, indicating new, preexisting, and withdrawn treatments at three different times.[8] Also plotted is the lagged centroid, a summary measure of current market demand defined by the average of commercially available treatment characteristics weighted by their market share.[9] Depending on what treatment mix the treated population collectively choose, the centroid could be any point in the convex hull formed from commercially available treatment characteristics. As the figure suggests, and as our estimates in Section 5 establish more rigorously, the characteristics of new treatments are distributed around the centroid offset by slight upward trend (improving treatment characteristics on both dimensions). This is evidence that future technologies are based on the treatments that are currently in most demand.

Our animated appendix (and its snapshots in Figure 5) shows that in our sample, both the centroid and innovations advance along the efficacy dimension first, and then on the side effects dimension agglomerating initially in the bottom left quadrant, then the top left, and

[8] We measure efficacy as the marginal contribution of a treatment to CD4 count and lack of side effects as the marginal contribution of a treatment to the log odds ratio of not causing ailments versus causing ailments. See Appendix A.8 for more details.

[9] The centroid is formally defined in Section 4 where our empirical model is laid out.



(a) Treatments: 90-1 (b) Treatments: 96-1 (c) Treatments: 01-1

NOTES: The figure shows snapshots of the evolution of the state of the product market at different dates in year-semester (YY-S) format. Products are two-dimensional. On the x-axis is a measure of a treatment's ability to not cause side effects, $\theta^{(ail)}$. On the y-axis is a measure of its contribution to underlying health, $\theta^{(health)}$. Dimensions are measured in different scales. Incumbent products are shown in black. New products are shown in red. Withdrawn products are shown as x. The green square is a measure of the prevalent technology in the previous period.

FIGURE 5

TREATMENT EVOLUTION
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

finally expanding into the top right. For example, comparing the second and third panels of the figure, notice that highly effective treatments exit the market, whereas some treatments that are less effective, but with fewer side effects, remain. A possible explanation for these trends is based on dynamic demand considerations that arise when future utility is geometrically discounted: as new treatments reduce mortality, life expectancy increases, the marginal value of living even longer declines, and consequently, the relative value of taking treatments with fewer ailments rises too.

2.7. *Individuals Do Not Instantaneously Adopt Treatments with Superior Characteristics.* Another feature of this market illustrated in Figure 5 is that consumers do not seem fully informed and seem to learn gradually. Many drugs well inside the frontier on both dimensions of treatment quality were consumed throughout the entire era. Nevertheless, consumers tended to drop treatments of inferior quality over time. The latter fact also provides evidence that other unobserved characteristics of the products were less important.

3. A MODEL OF DEMAND-PULL INNOVATION

This section develops a generic model of demand-pull innovation to explain how consumers can drive technological innovation. We describe the setup, define a rational expectations equilibrium (REE) for the model, and then analyze the nature of demand-pull externalities.

3.1. *Setup.* Consider a perishable differential product market that evolves over discrete time $t \in \{0, 1, \ldots\}$ with consumer tastes and technological change. We assume the characteristics $\theta_k \in \Theta$ of each product $k \in \{1, 2, \ldots\}$ are fixed over time and products are labeled in the order they are introduced. Let $\mathbf{K}_t$ be the set of products available at t, which includes the option of not buying any product ($k = 0$), a set that fluctuates over time because of both entry and exit.[10] We denote by s_{kt} the market share of k at t, the proportion of consumers buying

[10] The ordering of multiple products introduced in the same period is immaterial. Moreover, since some products enjoy a longer shelf life than others, the current set of available products does not correspond to the most recently introduced ones.

k in the total population, and define $\mathbf{s}_t \equiv \{s_{kt}\}_{k\in\mathbf{K}_t}$. We assume current technology $\theta_t \equiv \{\theta_k\}_{k\in\mathbf{K}_t}$ and the market shares of each product stochastically determine the course of innovation: θ_{t+1} is a random variable generated by θ_t and $\mathbf{s}_t$, with transition probability $G(\theta_{t+1}|\theta_t, \mathbf{s}_t)$. This approach incorporates state variables that firms supplying the market would base their decisions on, finessing issues researchers must directly confront when explicitly modeling firms behavior, but for that reason cannot predict how firms react to counterfactual changes that affect firm behavior.

Each consumer makes a product choice every period, setting the choice indicator $d_{kt} = 1$ if he chooses k at t, and 0 otherwise, implying $\sum_{k\in\mathbf{K}_t} d_{kt} = 1$ for all t. Consumers are represented by their period t individual characteristics: a state variable denoted by $z_t \in \mathbb{R}^Z$ with transition probability $F_k(z_{t+1}|z_t)$, and a disturbance vector denoted by ϵ_t, independent and identically distributed over t with cumulative density function $F_\epsilon(\epsilon_t)$ formed from elements ϵ_{kt} for $k \in \mathbf{K}_t$. We assume the population mass is constant, and denote the endogenously determined probability distribution of individual state variables within the population at t by $H_t(z)$. The consumer derives current utility $u_k(z_t) + \epsilon_{kt}$ from choosing $k \in \mathbf{K}_t$ at t, subjectively discounts future periods at the geometric rate of $\beta \in (0, 1)$, and hence attains a lifetime utility of:

$$\sum\nolimits_{t=0}^{\infty} \sum\nolimits_{k\in\mathbf{K}_t} \beta^t d_{kt}[u_k(z_t) + \epsilon_{kt}]. \tag{1}$$

3.2. *Rational Expectations Equilibrium.* At time $\tau \in \{0, 1, \ldots\}$ the consumer faces three sources of uncertainty about future periods $t > \tau$, namely the aggregate technology θ_t, his choice-specific disturbances ϵ_t, and his individual state variables z_t. We assume he maximizes the expected value of (1) choosing d_t at each t, a vector of indicator variables with elements d_{kt} for all $k \in \mathbf{K}_t$, given his current information $(z_t, \theta_t, \epsilon_t)$. Denote by $V_t(z_t, \theta_t)$ the ex ante value function at the beginning of period t just before ϵ_t is revealed. Appealing to Bellman's principle:

$$V_t(z_t, \theta_t) = \int \left\{ \max_{d_t} \sum\nolimits_{k\in\mathbf{K}_t} d_{kt}[v_{kt}(z_t, \theta_t) + \epsilon_{kt}] \right\} dF_\epsilon(\epsilon_t), \tag{2}$$

where $v_{kt}(z_t, \theta_t)$ and the conditional valuation function for choosing k at t given (z_t, θ_t), is recursively defined as:

$$v_{kt}(z_t, \theta_t) \equiv u_k(z_t) + \beta \int V_{t+1}(z', \theta') dG(\theta'|\theta_t, s_t) dF_k(z'|z_t) \tag{3}$$

and s_t denotes the REE market shares formed from the elements s_{kt} for each $k \in \mathbf{K}_t$. The optimal choice rule $d_{kt}(z_t, \epsilon_t, \theta_t)$ solves:

$$d_{kt}(z_t, \epsilon_t, \theta_t) \equiv \prod\nolimits_{j\in\mathbf{K}_t} 1\{\epsilon_{jt} - \epsilon_{kt} \le v_{kt}(z_t, \theta_t) - v_{jt}(z_t, \theta_t)\} \tag{4}$$

for each k, and the REE is defined by market clearing conditions that equate ex post market shares s_t, with product demand s_t^e induced by the optimal choices. Formally, it is a fixed-point solving:

$$s_{kt}^e = \int \int d_{kt}(z, \epsilon_t, \theta_t) dF_\epsilon(\epsilon_t) dH_t(z) \tag{5}$$

for all $(k, t, z, \theta) \in \mathbf{K}_t \times \{0, 1, \ldots\} \times \mathbb{R}^Z \times \Theta$, where the distribution of consumer characteristics $H_t(z)$ follows the law of motion:

$$H_{t+1}(z_{t+1}) = \sum_{k \in \mathbf{K}_t} \int_{z_t \in \mathbb{R}^Z} \int_{z \leq z_{t+1}} \int d_{kt}(z_t, \epsilon_t, \theta_t) dF_\epsilon(\epsilon_t) dF_k(z|z_t) dH_t(z_t) \tag{6}$$

In particular given the ex ante value function $V_{t+1}(z_{t+1}, \theta_{t+1})$, we can construct K_t equations in K_t unknowns, where K_t is the number of products at t and $s_t^e \in \Delta^{K_t}$, by successively substituting the right side of (3) for $v_{kt}(z, \theta_t)$, which explicitly depends on s_t through $G(\theta|\theta_t, s_t)$, into (4), and the resulting expression for (4) into (5). Recursively solving the entire system yields an REE.[11]

3.3. *Externality.* In an REE, consumers impose an externality on each other by not accounting for their individual effects on the rate and direction of innovation. To characterize this inefficiency and provide a theoretical basis for quantifying its effects, we decompose the individual optimization problem into two steps: The first step partitions the support of ϵ_t, for any $p_t \in \Delta^{K_t}$, into regions defined by a cutoff rule. The inversion theorem of Hotz and Miller (1993) implies that the optimal decision rule satisfies this cutoff rule. The second step substitutes the set of rules satisfying this necessary condition back into the original problem: choosing shares p_t for each product $k \in \mathbf{K}_t$ in the submarket defined by z_t then yields

$$p_{kt}(z_t, \theta_t) \equiv \int d_{kt}(z_t, \epsilon_t, \theta_t) dF_\epsilon(\epsilon_t) \tag{7}$$

alternatively interpreted as the CCP for choosing k at t given (z_t, θ_t).

In the first step, denote by D_t the space of functions that map ϵ_t into $\mathbf{K}_t$, and define for any vector $w_t \in \mathbb{R}^{K_t}$ the mapping $\psi_t(w_t) : \mathbb{R}^{K_t} \to \Delta^{K_t}$ as:

$$\psi_t(w_t) \equiv \int \left[\arg\max_{d_t \in D_t} \sum\nolimits_{k \in \mathbf{K}_t} d_{kt}(w_{kt} + \epsilon_{kt} - \epsilon_{0t}) \right] dF_\epsilon(\epsilon_t) \tag{8}$$

and normalize w_{0t} to 0. By Proposition 1 of Hotz and Miller (1993), $\psi_t(w_t)$ is invertible, so for all CCPs $p_t \in \Delta^{K_t}$ and $\epsilon_t \in \mathbb{R}^{K_t+1}$, we can also express the optimal decision for this static problem as a mapping of (p_t, ϵ_t) and the conditional expectation of the selected associated disturbance as a mapping of p_t:

$$\Upsilon_{kt}(p_t, \epsilon_t) \equiv \prod\nolimits_{j \in \mathbf{K}_t} 1\left\{ \epsilon_{jt} - \epsilon_{kt} \leq \psi_{kt}^{-1}(p_t) - \psi_{jt}^{-1}(p_t) \right\}, \tag{9}$$

$$\Lambda_{kt}(p_t) \equiv p_{kt}^{-1} \int (\epsilon_{kt} - \epsilon_{0t}) \Upsilon_{kt}(p_t, \epsilon) dF_\epsilon(\epsilon_t). \tag{10}$$

In our model, $\Upsilon_{kt}[p_t(z_t, \theta_t), \epsilon_t] = d_{kt}(z_t, \epsilon_t, \theta_t)$. Intuitively, (9) partitions the disturbance space into K subspaces optimally defining the cutoff values as a function of the CCPs.[12] Each element of the partition corresponds to the support for a selected disturbance term as a function of the CCPs, used when calculating the expectation displayed by (10).

[11] The existence of an REE can be established by imposing regularity conditions on the primitives, such as placing bounds on $u_k(z_t)$.

[12] For example, if k is the optimal choice given ϵ, then k is optimal for all ϵ^* satisfying $\epsilon^* = \epsilon + (\delta_0, \ldots, \delta_K)'$ with $\delta_k \geq 0$ and $\delta_{k'} \leq 0$ for $k' \neq k$.

The second step exploits (9) and (10) to transform the original problem of choosing d into one of choosing p_t. For any (fixed) $w_t \in \mathbb{R}^{K_t}$:

$$\int \left[\max_{d_t \in D_t} \sum\nolimits_{k \in \mathbf{K}_t} d_{kt}(w_{kt} + \epsilon_{kt} - \epsilon_{0t})\right] dF_\epsilon(\epsilon_t) \tag{11}$$

$$= \int \left[\max_{p_t \in \Delta^{K_t}} \sum\nolimits_{k \in \mathbf{K}_t} \Upsilon_{kt}(p_t, \epsilon_t)(w_{kt} + \epsilon_{kt} - \epsilon_{0t})\right] dF_\epsilon(\epsilon_t)$$

$$= \max_{p_t \in \Delta^{K_t}} \sum\nolimits_{k \in \mathbf{K}_t} \int [w_{kt}\Upsilon_{kt}(p_t, \epsilon_t) + (\epsilon_{kt} - \epsilon_{0t})\Upsilon_{kt}(p_t, \epsilon_t)] dF_\epsilon(\epsilon_t)$$

$$= \max_{p_t \in \Delta^{K_t}} \sum\nolimits_{k \in \mathbf{K}_t} p_{kt}[w_{kt} + \Lambda_{kt}(p_t)]. \tag{12}$$

The first equality is justified by (9); it implies that although the *arg-max* d resulting from the first line depends on ϵ_t, the *arg-max* p_t resulting from the second line does not because p_t is not a function of ϵ_t. Hence, the maximum and integral operators can be exchanged on the third line; the last line (12) then follows from (9).[13]

Setting $v_{kt}(z_t, \theta_t) - v_{0t}(z_t, \theta_t)$ for w_{kt} in (12) and differentiating for each $k \in \mathbf{K}_t$, the CCPs for the REE are a root to:

$$v_{kt}(z_t, \theta_t) - v_{0t}(z_t, \theta_t) + \Lambda_{kt}(p_t) + \sum\nolimits_{k' \in \mathbf{K}_t} p_{k't} \partial \Lambda_{k'}(p_t) / \partial p_{kt} \tag{13}$$

in p_t for each (z_t, θ_t). The objective in a social planning problem (SPP) is to maximize the expected value of (1) integrated over the distribution of population characteristics $H_t(z)$. The first step of the REE and SPP problems is identical. Appealing to (3), the CCPs for the SPP are chosen to maximize:

$$\int \sum_{k \in \mathbf{K}_t} p_{kt} \begin{Bmatrix} u_k(z_t) - u_0(z_t) + \Lambda_{kt}(p_t) \\ +\beta \int V_{t+1}(z', \theta')[dF_k(z'|z_t) - dF_0(z'|z_t)] dG(\theta'|\theta_t, s_t) \end{Bmatrix} dH_t(z) \tag{14}$$

with respect to p_t for each (z_t, θ_t) in the second step. Again, with reference to (3), the necessary first-order condition equates (13) with:

$$-\beta \sum_{k' \in \mathbf{K}_t} p_{k't} \int V_{t+1}(z', \theta')[dF_{k'}(z'|z_t) - dF_0(z'|z_t)] \frac{\partial g(\theta'|\theta_t, s_t)}{\partial s_{kt}} d\theta' \tag{15}$$

instead of 0, where $g(\theta'|\theta_t, s_t) \equiv \partial G(\theta'|\theta_t, s_t)/\partial \theta'$ is the density function of θ'. From (15), the SPP product allocated to each consumer accounts for the effect of innovation on every other consumer. For example, if the consumption share of a particular product boosts innovation and next period's SPP ex ante value function increases with that share, the planner is prompted to allocate that product to those with a lower value of their product-specific disturbance than otherwise, thus increasing its overall share above the REE share.

4. PARAMETERIZING THE MODEL

Our application to medical treatment focuses on two key characteristics of HIV treatments, their effects on the CD4 count, and the probability a consumer does not experience side effects, or the lack of ailments. Thus, $\Theta \equiv \mathbb{R}^2$ in our application. We add several features to the

[13] See also Aguirregabiria and Magesan (2018) for an alternative proof.

generic framework. Since HIV+ individuals have the option of joining a clinical trial and receiving an experimental treatment, our empirical framework includes this alternative. In contrast to the commercial treatments whose characteristics are fixed upon entry, the characteristics of the experimental treatment evolve over time with technological change and market preferences for health and ailments. For tractability reasons, we constrain the choice of commercial treatments: instead of allowing consumers to select any commercial treatment, they can repeat exactly the same treatment as last period, or choose between several clusters that partition treatments according to their characteristics; upon choosing a cluster, consumers are randomly assigned a treatment from within the cluster. We model variation in labor supply participation and its effects on future participation, mainly because of the documented relationship between health and work. Given the high mortality rates in this population, we also model survival, taking account of demographic heterogeneity throughout the population. We now explain these features in more detail.

4.1. *Innovations in Treatments.* The characteristics of new commercial treatments, and also experimental treatments, are modeled as random draws from a distribution around a *centroid*, a weighted average of the characteristics of treatments commercially available in the previous period, where the weights are market shares. We define $\omega_t \equiv (\omega_t^{(health)}, \omega_t^{(ail)})'$, the centroid for period t, as:

$$\omega_t \equiv \frac{\sum_{k\in\mathbf{K}_{t-1}} s_{k,t-1}\theta_k}{\sum_{k\in\mathbf{K}_{t-1}/\{0\}} s_{k,t-1}}. \tag{16}$$

where $\theta_k \equiv (\theta_k^{(health)}, \theta_k^{(ail)})$. Also let s_{et} denote the share of consumers opting for an experimental treatment through a clinical trial. We assume that the characteristics of new and experimental treatments are random variables, respectively, determined by:

$$\begin{aligned} \theta_k - \omega_{t-1} &= \phi_{0v} + \phi_{1v}\cdot s_{e,t-1} + \nu_k \qquad \text{if } k \in \mathbf{K}_t \text{ and } k \notin \mathbf{K}_{t-1}, \\ \theta_{et} - \omega_t &= \phi_{0v} + \phi_{1v}\cdot s_{e,t-1} + \nu_{et}, \end{aligned} \tag{17}$$

where ν_k and ν_{et} are independent and identically distributed with density function $f_v(v)$ and mean 0.[14] Equation (17) shows that new commercially available treatments in period t are innovations around the previous-period centroid ω_{t-1}, whereas the experimental treatment available through clinical trial participation at t is an innovation around the current-period centroid ω_t. Thus, (17) captures recency in experimental treatment innovations relative to new treatments launched commercially. Also since the primary purpose of clinical trials is to produce innovations, we allow the characteristics of new and experimental treatments to be systematically affected by the amount of experimentation undertaken in the previous period.

4.2. *Entry and Exit.* Both entry and exit are driven by market demand in our application. We assume that the number of new treatments in period t, N_t, is a negative binomial random variable, whose mean depends on the market share of experimental treatments, $s_{e,t-1}$, and the magnitude of innovations in the previous period, denoted by κ_{t-1}:

$$E[N_t|\kappa_{t-1}, s_{et-1}] = \exp(\phi_1^N \kappa_{t-1} + \phi_2^N s_{et-1}). \tag{18}$$

[14] Consistent with this setup, we test and cannot reject the hypothesis that the coefficient on the centroid in Equation (17) is equal to 1, that is, that new product characteristics are drawn from a distribution centered on the centroid. We also find that conditioning on the centroid captures the relationship between experimental treatments at t and the characteristics of new treatments entering the market at $t+1$. In Appendix A.2.1, we test whether the innovation shocks, ν, are serially correlated and we cannot reject the null hypothesis that they are independent.

We quantify the magnitude of current innovations by $\kappa_t \equiv \delta_1 \kappa_t^{(health)} + \delta_2 \kappa_t^{(ail)}$ where (δ_1, δ_2) is a vector of scaling weights to achieve comparability across the different treatment characteristics, $\kappa_t^{(health)}$ is defined analogously to $\kappa_t^{(ail)}$, and:

$$\kappa_t^{(ail)} \equiv \max_{\{k:\ k \in \mathbf{K}_t \text{and } k \notin \mathbf{K}_{t-1}\}} \left\{ \theta_k^{(ail)} - \omega_{t-1}^{(ail)} \right\}. \tag{19}$$

The distribution for N_t captures two empirical patterns. First, if a smaller proportion of the market consumes experimental treatments, fewer new treatments are likely to appear, and for the same reason, we expect less innovation in treatment characteristics. Second, a relatively large number of new treatments tend to follow large breakthroughs, because they spur rival suppliers.

The exit rule is defined by the dyad $\{\underline{s}, \bar{s}\}$. We can split the market share of treatment k into new $\underline{s}_{kt}$ and repeat $\bar{s}_{kt}$ consumers, where $\underline{s}_{kt} + \bar{s}_{kt} = s_{kt}$. If $\underline{s}_{kt}$ dips below the critical number $\underline{s}$, the treatment is no longer available for new consumers; when $\bar{s}_{kt}$ dips below $\bar{s}$, the treatment is withdrawn altogether.

4.3. *Choice Set.* Consumers' knowledge of the market is limited. At each t, they know the product spectrum and market shares (θ_t, s_t). Supposing that i took a commercial treatment in the previous period and that the treatment has not been withdrawn, he can order a repeat prescription. In this case, he knows its characteristics. Let $r_t \in \{0, 1\}$ denote a period t indicator variable for whether the consumer took a commercial treatment in period $t-1$ that is still on the market ($r_t = 1$) or not; for $r_t = 1$, denote by θ_{rt} the treatment characteristics of last period's commercial treatment. The experimental treatment is also in a consumer's choice set; at period t, he knows the innovation process, that is, $f_v(v)$, along with its previous-period market share $s_{e,t-1}$, and can therefore infer the probability distribution generating θ_{et} but does not know the realization of θ_{et}. He also retains the option of being untreated.

A consumer cannot, however, directly select any (other) treatment from $\mathbf{K}_t$, but only choose one of J clusters that partition $\mathbf{K}_t$; he is randomly assigned to a treatment within the selected cluster. Let $\mathbf{K}_{jt}$ comprise the treatments belonging to the jth cluster at time t, and denote by $q_{kjt}(k|\mathbf{K}_{jt})$ the probability that treatment $k \in \mathbf{K}_{jt}$ is assigned when cluster j is chosen at t.[15] We denote by θ_{jt} the vector of treatment characteristics from choosing j at t; it is drawn from a probability density induced onto the jth cluster at t, defined by:

$$f_{jt}(\theta|\mathbf{K}_{jt}) = \sum_{k \in \mathbf{K}_{jt}} q_{kjt}(k|\mathbf{K}_{jt}) I\{\theta_k = \theta\}. \tag{20}$$

Each period t commercial treatments with similar characteristics are regrouped into J clusters using an algorithm denoted by $c(\mathbf{K}_t)$.[16] We assume that the consumer knows $f_{jt}(\theta|\mathbf{K}_{jt})$, the distribution of treatment characteristics within each cluster.

Using a clustering algorithm shrinks huge choice sets driven by a plethora of similar products competing in a finely differentiated market, a challenge for consumers, and also empirical researchers in the fields of marketing and industrial organization. Basing the algorithm on the characteristics of treatments captures, albeit in a reduced form way, the role intermediaries such as doctors and pharmacies play in writing prescriptions; their expertise helps consumers identify and choose a cluster of treatments with their preferred probability distribution of characteristics. In our model, treatments are an experience product: consumers learn their treatment characteristics after one period. This implies that they sometimes purchase new inferior treatments, an empirical feature of our data. Clustering also induces differential

[15] We set $J = 3$ but test the algorithm with two, three, and four clusters. Results are shown in Section 5.

[16] The clustering rule c in this model is known as the k-means algorithm; it is used in machine learning. A flexible polynomial based on the characteristics of treatments in the cluster specifies q_{kjt}. See Appendix A.2.1 for details.

or staggered learning over the product life cycle (even) in an REE, because consumers cannot immediately identify exactly which treatments are the most successful from aggregate market data alone.[17]

To incorporate clusters, the experimental treatment, and repeat prescriptions into the choice set, we adopt the following notation: Let $d_{jt} \in \{0,1\}$ for all $j \in \{0,1,\ldots,J,J+1,J+1+r_t\}$, where setting $d_{jt}=1$ means that at t, the consumer chooses: not to purchase a treatment when $j=0$; a commercial treatment from cluster j that differs from last period's treatment when $j \in \{1,\ldots,J\}$; the experimental treatment when $j=J+1$; the same commercial treatment as in the prior period when $r_t=1$ and $j=J+2$. The repetition indicator r_t can be expressed recursively as:

$$r_t = \sum\nolimits_{j=1}^{J} 1\{\theta_{j,t-1} \in \mathbf{K}_t\} d_{jt-1} + 1\{\theta_{r,t-1} \in \mathbf{K}_t\} d_{J+2,t-1} r_{t-1}. \tag{21}$$

4.4. *Health, Survival, and Ailments.* Treatment in period t directly affects consumers through two channels, through the health of the consumer, denoted by $h_t \in \mathbb{R}^+$, and measured by his CD4 count, and physical ailments $y_{1t} \in \{0,1\}$, where $y_{1t}=1$ means there are no side effects from the treatment. We model the health production function as:

$$h_{t+1} = \sum\nolimits_{s=0}^{5} \gamma_s^{(health)} h_t^s + \sum\nolimits_{j=0}^{J+1+r_t} d_{jt} \theta_{jt}^{(health)} + \epsilon_{ht}. \tag{22}$$

The first expression in (22), a higher order polynomial in lagged health, captures its persistence over time. The second is the boost to the CD4 count from treatment in the previous period, whereas ϵ_{ht} is independent and identically distributed (*i.i.d.*) with mean 0. We do not allow for differential treatment characteristics by race or ethnicity but we test this constraint in Appendix A.2.1.[18] We find that efficacy does not vary across race or ethnicity, whereas there is limited evidence that side effects do.

The probability of the event $y_{1t}=0$, suffering a physical ailment in period t, depends on current health, h_t, and the second characteristic of current treatment, captured by the summation of $d_{jt}\theta_{jt}^{(ail)}$ over j. We assume that the probability of having physical ailments in t is:

$$\begin{aligned} &\Pr\left[y_{1t}=0 \middle| h_t, d_t, \left\{\theta_{jt}^{(ail)}\right\}_{j=0}^{J+1+r_{it}} \right] \\ &= \left[1+\exp\left(\sum\nolimits_{s=0}^{5} \gamma_s^{(ail)} h_t^s + \sum\nolimits_{j=0}^{J+1+r_{it}} d_{jt}\theta_{jt}^{(ail)} \right) \right]^{-1}. \end{aligned} \tag{23}$$

Due to the high mortality rates afflicting the sample population, we also model survival. It depends on a polynomial in health, lagged physical ailments, and a demographic vector that includes age (in half year increments), race/ethnicity (black, Hispanic, white), and education level (high school, some college, college, or more than college). Let $b_t=1$ denote survival through to period t, with $b_t=0$ otherwise, and let a_t denote the consumer's corresponding demographics. We assume:

$$\Pr\left[b_t=1 | x_t^{(live)}, b_{t-1}=1 \right], = \left[1+\exp\left(x_t^{(live)} \gamma^{(live)} \right) \right]^{-1} \tag{24}$$

[17] In Miller (1988), for example, in equilibrium, all consumers can deduce product quality two periods after its introduction, by observing the level of repeat buying. More generally, it is well known that rational expectations equilibria are fully revealing unless some other form of aggregate uncertainty or market imperfection is present (Radner, 1979; Grossman and Stiglitz, 1980).

[18] We do not consider individual-specific treatment effects because the sample is too small to obtain reliable fixed effect estimates for the more than 80 observed treatments.

$$\text{where } x_t^{(live)} = \left(1, h_t, \ldots, h_t^5, a_t, y_{1,t-1}\right).$$

4.5. *Labor Supply, Income, and Medical Expenditures.* Labor supply, denoted by $y_{2t} \in \{0, 1\}$, a state variable the individual learns at the beginning of t before making his treatment decision, follows a logit transition probability that depends on current health h_t, demographics a_t, and previous-period labor supply $y_{2,t-1}$:

$$\Pr\left[y_{2t} = 1 | x_t^{(labor)}\right], = \left[1 + \exp\left(x_t^{(labor)} \gamma^{(labor)}\right)\right]^{-1} \tag{25}$$
$$\text{where } x_t^{(labor)} = \left(1, h_t, \ldots, h_t^4, a_t, y_{2,t-1}\right).$$

Gross income, y_{3t}, is governed by the process:

$$y_{3t} = x_t^{(inc)} \gamma^{(inc)} + \eta + \varepsilon_t^{(inc)} \tag{26}$$
$$\text{where } x_t^{(inc)} = \left(1, h_t, \ldots, h_t^7, a_t, y_{1t}, y_{2t}\right),$$

the term η captures person-specific productivity and $\epsilon_t^{(inc)}$ is an *i.i.d.* income shock uncorrelated with $x_t^{(inc)}$ that the consumer observes before making his treatment choice.

Out-of-pocket expenditure on health care y_{4t} is determined by:

$$\left(y_{4t} = \max\left\{x_t^{(spend)} \gamma^{(spend)} + \sigma^{(spend)} \varepsilon_t^{(spend)}, 0\right\}\right) \tag{27}$$
$$\left(\text{where } x_t^{(spend)} = \left(1, h_t, \ldots, h_t^6, a_t, y_{1t}, y_{2t}, d_t\right), \right)$$

and $\varepsilon_t^{(spend)}$ is an *i.i.d.* standard normal random variable. Expenditures increase from purchasing a treatment but may also increase due to underlying health and physical ailments. Since we do not directly observe pharmaceutical prices, (27) differentiates between the costs of not being treated, a commercial treatment and an experimental treatment, but does not account for cost differences between alternative commercial treatments.[19]

4.6. *Preferences.* We model current utility in period t from choosing j as:

$$U_{jt} \equiv \alpha^{(inc)}(y_{3t} - y_{4t}) + \alpha^{(ail)} y_{1t} d_{0t} + \epsilon_{jt} \tag{28}$$
$$+ \sum_{j=1}^{J} d_{jt}\left(\alpha_1^{(health)} h_t + \alpha_1^{(dem)} a_t\right) + \sum_{j=J+1}^{J+1+r_{it}} d_{jt}\left(\alpha_j^{(health)} h_t + \alpha_j^{(dem)} a_t\right),$$

where the choice-specific disturbance ϵ_{jt} is *i.i.d.* as type 1 extreme value (TIEV). Individuals are risk neutral. The coefficient $\alpha^{(inc)}$ is the marginal utility of wealth, $\alpha^{(ail)}$ is the taste for absence of ailments when untreated (that is, when $y_{1it} = d_{0it} = 1$), whereas $\alpha_j^{(health)}$ and $\alpha_j^{(dem)}$ are choice-specific utilities associated with health and demographics. Besides affecting lifetime utility indirectly (through its impact on future health, survival, and outcomes), current health affects utility directly; in particular, α_{jh} captures differences in the time and psychic costs of accessing an experimental treatment by health (for example, if doctors are more willing to suggest experimental treatments to the sickest), and it also captures how individuals

[19] End-users customarily pay a standardized deductible that is a fraction of the brochure price of the drug paid by the insurance company. Mean out-of-pocket expenditures in our sample are \$273 every six months (indexed to 2000 \$U.S.). In 1995, a semester of ARV drugs cost between \$488 and \$2,315, and a semester of HIV treatment (primary ARVs, adverse events, laboratory monitoring, and prophylaxis) cost between \$1,054 and \$5,504 (Gable et al., 1996).

may be more willing to try a new treatment from a cluster when in poor health. We restrict the cluster coefficients on health and demographics to be the same, implying $\alpha_j^{(health)} = \alpha_1^{(health)}$ and $\alpha_j^{(dem)} = \alpha_1^{(dem)}$ for all $j \in \{1, \ldots, J\}$, but allow for individuals to derive different utility from consuming an experimental treatment or from repeating consumption of a commercially available treatment.

4.7. *Connecting the Generic Model to the Parameterization.* Summarizing, the law of motion for the supply of available treatments, $G(\theta_{t+1}|\theta_t, \mathbf{s}_t)$ in the generic model of the previous section, is determined by (16) and (17) in the parameterization that define the characteristics of new commercial treatments as well as experimental treatments, by (18) and (19) determining the number of new commercial treatments, and by the dyad $\{\underline{s}, \bar{s}\}$ defining the withdrawal of treatments. The personal state variables for this application, z_{it}, are: survival, b_{it}; health, h_{it}; the characteristics of the treatment taken last period if a commercial treatment was consumed, denoted by θ_{rit}; demographic variables, a_{it}; an individual fixed effect, η_i; and labor supply participation y_{2it} stochastically determined by (25). Taking the set of available treatments and the distribution of consumer characteristics as given, the consumer with personal state z_{it} makes a treatment choice, setting $d_{jit} = 1$ for some $j \in \{0, 1, \ldots, J, J+1, J+1+r_{it}\}$. Appendix A.7.2 presents the value function of the parameterization.

Then his ailments status, y_{1it}, is drawn from (23), his gross income, y_{3it}, is drawn from (26), and his medical out-of-pocket expenditures, y_{4it}, are drawn from (27), all conditional on his personal state variables z_{it} and his treatment choice. Consequently, when the consumer makes his choice, he does not know U_{jit}, but only $E[U_{jit}|\epsilon_{jit}, z_{it}, j]$. When maximizing his expected lifetime utility, the consumer takes into consideration his health transition determined by (22), aging, and aggregate transitions determined in part by the equilibrium market shares s_t, and he discounts the future not only by β, but also by the survival probability given by (24). The REE for this medical treatment model is solved the same way as described for the generic model, and the externality we described for the generic model takes the same form.[20]

Intuitively, the dynamic aspects of the consumer choice problem are motivated by the trade-off between health and ailments, as well as the trade-off between current expenditure and health. Aging changes the balance, shortening lifetime horizon, and thus increasing the likelihood of treatments based on more palliative care, but experience with medical professionals may overcome an initial reluctance to seek help. Dynamically, the prospects of successful innovations encourages consumers to endure more hardship now in the hope of a healthier life in years to come. Consumers can wait for new treatments to enter the market. However, their patience only resolves some of the uncertainty as new treatments are grouped in clusters with old treatments, and consumers only know the distributional characteristics of clusters $f_{jt}(\theta|\mathbf{K}_{jt})$ given by (20). A consumer resolves his uncertainty regarding a new treatment after one period; thus, the characteristics of repeat prescriptions are known to him. Consequently, the market is the main source of consumer learning. Alternatively, consumers can take an experimental treatment hoping to obtain early access to the next breakthrough; such choice always entails facing the uncertainty associated with the distribution of new treatment shocks $f_v(v)$.

5. ESTIMATION

We identify and estimate the model sequentially: (i) treatment characteristics, individual transitions for mortality, health and labor supply, and processes for ailments, income, and medical expenses; (ii) entry, exit, and innovations in medical treatment; (iii) clustering; and (iv) individual utility. Drawing upon parameter estimates obtained in the earlier stages where

[20] Given the TIEV assumption for the disturbance term, $\Lambda_{jt}(p)$, in Equation (10) equals $-\ln p_{jt}$, and hence, the expression for the selection correction to $E[U_j(h_t, y_t)|z_t, j]$ reduces to $-p_{jt}^{-1}$. (See Hotz and Miller, 1993.)

appropriate, each piece is estimated separately. The entire procedure is repeated 500 times to obtain bootstrapped estimated standard errors that account for the sequential process.

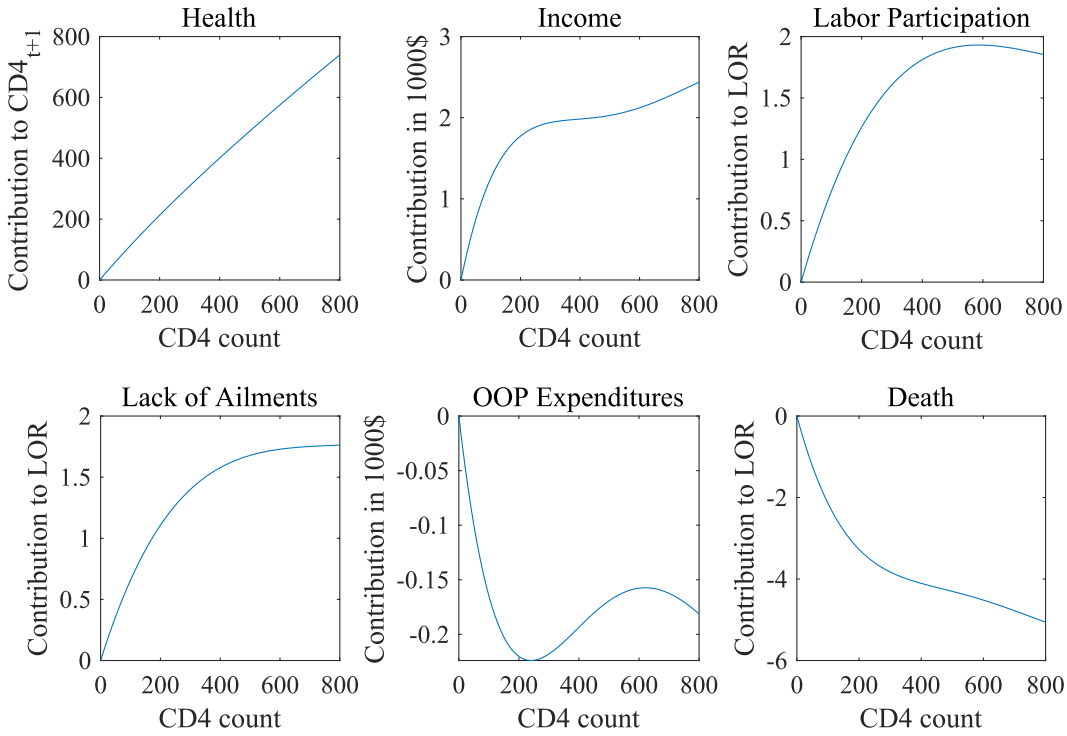
In the first step, treatment characteristics θ_k are identified and estimated using observed patient health outcomes for given treatment choices: effectiveness, $\theta_k^{(health)}$, is estimated using the process for future health, h_{it+1}, in (22); the (lack of) side effects $\theta_k^{(ail)}$ is estimated using the process for physical ailments, y_{1it}, in (23). The transition function for health along with other state transitions and outcomes and survival probabilities are identified and estimated using their sample analogues. The second step is to estimate centroids for innovation and the magnitude of previous innovations for each t using treatment characteristics and Equations (16) and (19). Then we estimate Equation (17) that governs innovation. The residuals of this equation are used to nonparametrically estimate the two-dimensional distribution of innovation shocks f_v. The third step uses the estimated treatment characteristics to form clusters using the clustering rule $c(\mathbf{K}_t)$ described in Appendix A.8. The characteristics of the treatments in each cluster and treatment shares yield the distribution of characteristics induced onto the jth cluster at t given by Equation (20).

The last step, described below, estimates the utility function with a CCP estimator using the optimality conditions that arise from dynamic discrete choice. (See Appendix A.8.4.) The forward-simulation procedure we implement is a multistage algorithm that modifies the approach of (Altuğ and Miller, 1998) to our context, where the choice set evolves over time: (a) Estimate flexible parametric CCPs from observed consumer behavior that control for the aggregate state as well as person-specific state variables to predict treatment choices given possible counterfactual future choice sets drawn from the endogenous, stochastic processes of entry and exit estimated in previous steps; (b) simulate a collection of aggregate paths describing the evolution of a hypothetical market for each observation (i, t) (i.e., the evolution of available products and population characteristics), using the CCPs and the entry and exit processes; (c) given the aggregate paths constructed in step (b) and the CCPs simulate a choice and transition path for each observation (i, t); (d) set $\beta = 0.95$ and estimate the utility function parameters in (28) using a Generalized Method of Moments (GMM) estimator; the estimator has a closed form because simulated lifetime utility is linear in the remaining parameters. Appendix A.8 further elaborates these steps.

5.1. *Life-Cycle Processes and Treatment Characteristics.* Tables A.10–A.14 in Appendix A.15.1 report the point estimates of the individual processes; almost all of the coefficients are statistically significant at the 5% level. Figure 6 depicts the estimated relationship between current-period health and other state variables and outcomes. The relationship between current and one-period-ahead health is nearly linear; health deteriorates over time but is somewhat persistent. The effect of current health on other processes is highly nonlinear. When the CD4 count falls below about 250 (the AIDS threshold), sharp changes occur; mortality, ailments, and medical expenditures increase; and labor supply and income decline. These large shifts illustrate the well-known fact that a declining CD4 count caused by HIV infection has only a small impact on observed health until very low levels are reached, at which point it becomes catastrophic.

The estimated equations for individual-level outcomes and transitions exhibit additional patterns that have been found elsewhere. Survival is higher for black men and for those not suffering physical ailments. Labor supply increases with education, past participation in the job market, and up until age 40. Gross income decreases with ailments (poor physical health reducing productivity), increases with employment and education, and is concave in age. Black and Hispanic men earn less on average than white men, suggesting that racial/ethnic inequality in labor outcomes, evident in many other samples, extends to the HIV+ population.

Out-of-pocket medical expenditures increase with age, education, and ailments (controlling for treatment usage), perhaps due to expenditures on other health conditions. Controlling for health status, education, labor supply, and treatment, minorities spend less out-of-pocket.



NOTES: CD4 count measured in hundreds of cells per microliter. LOR stands for log odds ratio. OOP stands for out-of-pocket. Semestral income and expenditures measured in real $U.S. indexed to 2000.

FIGURE 6

EFFECT OF CURRENT HEALTH ON FUTURE HEALTH AND OUTCOMES
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

Employment increases expected expenditures, possibly reflecting different pricing schemes for public versus private insurance. Finally, we find that out-of-pocket prescription expenses are incurred when taking experimental treatments, but 27% less than when taking commercially available treatments; Santolaya Perrín and García López (2008) find a similar gap in out-of-pocket prescription expenses of 33%. A caveat of our model is that it does not incorporate the role of insurance on out-of-pocket expenditures that could lead to biased estimates of the nonpecuniary costs of taking treatment. We show in Appendix A.7.1 that insurance is not strongly correlated with either labor participation or health, but it does decrease out-of-pocket expenses for individuals taking commercial treatment. Since 88% of our sample is covered by insurance, we do not think that this issue is critical and leave for future research the exploration of the role of insurance on demand externalities in innovation.

5.2. *Innovation and Entry.* Table 3 presents our coefficient estimates associated with the systematic components of (17), the innovation process that determines the distribution of characteristics of new commercial treatments and experimental treatments.[21] The positive parameter estimates for $\phi_{1\nu}^{(health)}$ and $\phi_{1\nu}^{(ail)}$ indicate that the market share of the experimental treatment has a positive effect on the mean characteristics of new treatments next period. The negative constants imply that in the absence of any clinical trials, the mean characteristics of new treatments would be inferior to the current share-weighted mean. Thus, expected effectiveness innovations are positive for lagged shares of the experimental treatment above 5.6%,

[21] Table A.15 in Appendix A.15.1 reports our estimates of the treatment characteristics, obtained from the health and ailments processes.

TABLE 3
INNOVATION COMPONENTS

Health Innovation				*Ailments Innovation*			
coef.	Variable	est.	se	coef.	Variable	est.	se
$\phi_{1\nu}^{(health)}$	s_{et-1}	433.11	(21.04)	$\phi_{1\nu}^{(ail)}$	s_{et-1}	1.93	(0.35)
$\phi_{0\nu}^{(health)}$	*Constant*	−24.14	(1.56)	$\phi_{0\nu}^{(ail)}$	*Constant*	−0.15	(0.03)

NOTES: Estimates from (17). In the table, "coef." stands for coefficient and "est." stands for estimate; "se" stands for standard error, in parentheses, computed using subsampling with 500 subsamples.

and expected innovations on the ailments dimension are positive for lagged shares of the experimental treatment above 7.7%. Since the mean share of the experimental treatment is 7% in our sample (Table 2), new treatments are more effective than the prevalent technology on average, but have more side effects. Therefore, the improvement seen on the mean ailments over time is due to the equilibrium selection of consumers favoring treatments with fewer ailments.

Figure 7 depicts the estimated distribution of innovation shocks, $f_\nu(\nu)$, formed from the residuals of (17). Conditional on the previous share of the experimental treatment, the density is unimodal and approximately bell shaped: small innovations are the norm. In addition, there is positive correlation of 0.24 between the two quality dimensions: shocks improving efficacy tend to be accompanied by fewer side effects. The second component of the innovation process is the distribution of the number of new commercial treatments.[22] Table A.16 in Appendix A.15.1 reports the estimates of the coefficients ϕ_1^N and ϕ_2^N in (18); they imply that the expected number of new treatments increases with both the size of previous innovations and the previous share of the experimental treatment. Figure A.3 in Appendix A.15.1 plots the estimated unconditional distribution against the relative frequencies of entry, illustrating the satisfactory fit between model and data.

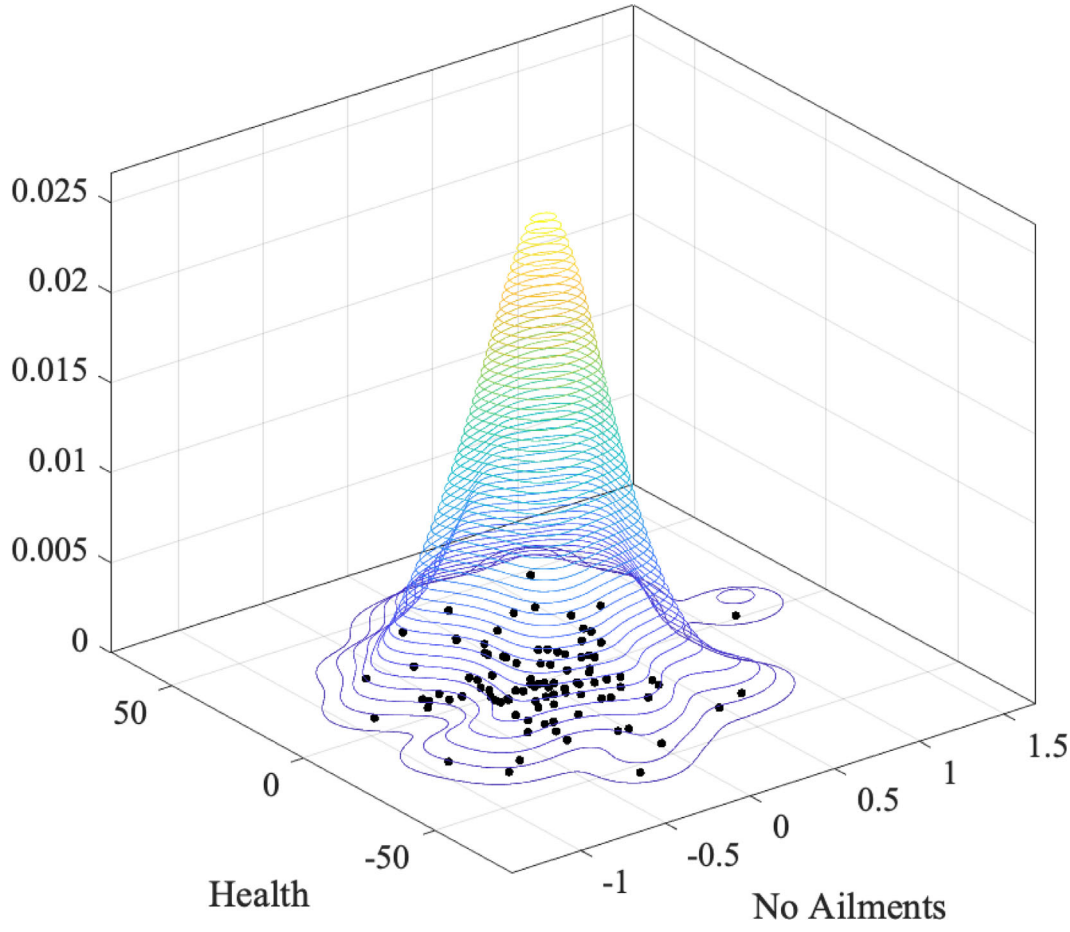
5.3. *Clusters.* The clustering algorithm minimizes a distance metric between treatment characteristics within the cluster, without regard to vintage. We tested the algorithm with two, three, and four clusters. As Figure 8 shows, when $J = 3$, one cluster typically picks out older products in the lower left quadrant, whereas another one generally captures treatments in the upper right quadrant. Using only two clusters often yielded a vertical ranking on effectiveness, erasing trade-offs between efficacy and propensity to cause ailments. As shown in Table A.9 in Appendix A.8, using four clusters dramatically decreases statistical power by severely reducing the number of individuals per cluster at any given time.

In the estimated model, individuals are randomly assigned a treatment from their preferred cluster according to $q_{kjt}(k|\mathbf{K}_{jt})$ given in (20). Point estimates reported in Table A.17 in Appendix A.15.1 indicate that treatments with relatively harsh side effects within their cluster are also less likely to be assigned.

5.4. *Preferences.* We estimate the preference parameters and corresponding standard errors under the constraint that utility is nondecreasing in income, in other words $\alpha^{(inc)} \geq 0$.[23] Setting $\beta = 0.95$ yields the lowest value of the econometric criterion function among the limited set $\{0.8, 0.9, 0.95\}$, so we report the remaining parameter estimates when $\beta = 0.95$ in

[22] The third component of the law of motion of the set of available treatments is the exit rules. We set the exit thresholds $\underline{s}$ and $\bar{s}$ using their sample counterparts, the minimum values observed in the data, at 0.0047 and 0.0012, respectively.

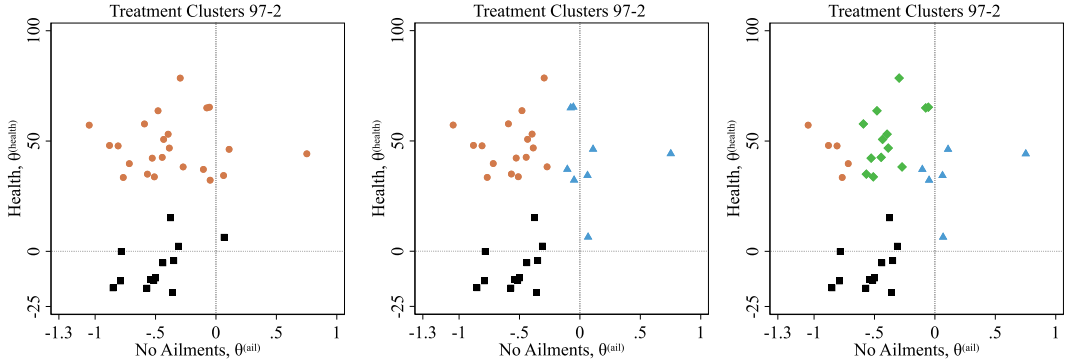
[23] We also estimate the model and obtain standard errors without imposing this constraint. Because the constraint is not binding, the point estimates are identical. Only the standard errors change and the only qualitative change is that the net income utility coefficient becomes insignificant.



NOTES: $f_v(v)$ is estimated nonparametrically off the residuals from (17).

FIGURE 7

DISTRIBUTION OF INNOVATION SHOCKS, $f_v(v)$.
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]



(a) Two Clusters (b) Three Clusters (c) Four Clusters

NOTES: Treatments commercially available in the second semester of 1997 grouped into two to four clusters.

FIGURE 8

VARYING THE NUMBER OF CLUSTERS
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

TABLE 4
UTILITY PARAMETERS, u_t

coef.	Variable	est.	se
$\alpha^{(inc)}$	$NetIncome_t$ $(y_{3t} - y_{4t})$	0.057	(0.023)
$\alpha^{(ail)}$	$NoAilments_t \cdot NoTreatment_t$ $(y_{1t} d_{0t})$	1.019	(1.120)

		Cluster $j = 1, 2, 3$		*Experimental* $j = J + 1$		*Repeat* $j = J + 2$	
coef.	variable	est.	se	est.	se	est.	se
$\alpha_j^{(dem,w)}$	*White*	−3.546	(0.703)	−1.468	(0.309)	0.502	(0.459)
$\alpha_j^{(dem,b)}$	*Black*	−4.190	(0.751)	−2.553	(0.351)	0.276	(0.477)
$\alpha_j^{(dem,h)}$	*Hispanic*	−3.967	(0.825)	−1.585	(0.363)	0.707	(0.391)
$\alpha_j^{(dem,a)}$	Age_t	0.043	(0.012)	0.032	(0.006)	0.009	(0.007)
$\alpha_j^{(health)}$	$h_t/10^3$	−2.021	(0.417)	−2.461	(0.189)		

NOTES: Estimation of (28). In the table, "coef." stands for coefficient and "est." stands for estimate; "se" stands for standard error, in parenthesis, computed using subsampling with 500 subsamples; h_t is defined as the number of white blood cells per cubic millimeter of blood.

Table 4. Similar to results in Chan and Hamilton (2006) and Papageorge (2016), estimated utility increases with net income, and untreated individuals value no ailments more highly (since $\alpha^{(ail)} > 0$). We find that all individuals lose utility from being treated, African Americans and Hispanics the most; African Americans obtain the largest disutility from consuming experimental treatments.[24] Age mitigates the utility costs of new treatments, both commercial and experimental. Consuming new treatments, especially experimental, is less costly for the most unhealthy, who, on average, have found their previous treatments unsuccessful. Finally, the utility of remaining on a treatment is positive relative to trying a new treatment, but not significantly different from the baseline no-treatment option.

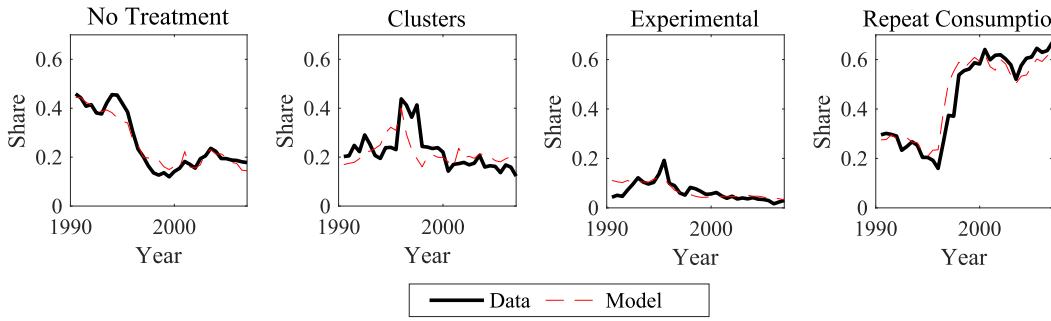
6. TECHNOLOGICAL PROGRESS IN MEDICAL TREATMENT

The final section of our analysis draws upon the model and our parameter estimates to address the following four questions: How well does the model track the aggregate features of the data? Given the specification and the estimates of the model, how likely were the key innovations that were observed over this period? What is the quantitative impact of the externality evaluated at the REE? How much would subsidizing experimental treatments improve matters?[25]

6.1. *Fitting the Aggregate Data.* Figure 9 plots actual treatment choices over time along with those generated by the estimated model given the state at every point in time. The latter captures the main trends in the data, including the rise in repeated usage as treatments improve over time and the decline in the share of individuals not consuming any treatment. The model also captures shifts over time in the share of individuals trying something new, either

[24] The greater disutility among African Americans from consuming experimental treatments may be due to treatment costs, different expected health outcomes, and/or greater distrust in the medical system resulting from events such as the infamous Tuskegee experiment (Harris et al., 1996; Alsan and Wanamaker, 2018).

[25] In addition, in Appendix A.15.3, we study the evolution of treatment quality if consumers had less influence over the process of innovation, restricting the role of demand pull. We find that eliminating the effects of repeat consumption improves health and survival, but leads to more physical ailments. The reason is that once the process of innovation has delivered a baseline level of effectiveness, individuals switch toward medical treatments with fewer side effects despite the detrimental impact on their survival.



NOTES: Simulated and empirical choice shares over time.

FIGURE 9

GOODNESS OF FIT
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

by consuming experimental treatments or by choosing a cluster that entails assignment to a new treatment.[26]

6.2. *How likely Was HAART?.* We assess the likelihood of a large innovation such as HAART by comparing simulated paths the model generates, plotted as gray lines in Figure 10, with the realized path observed in the data, the black line. We simulated 100 paths starting at two distinct dates, namely. the first semester of 1991, prior to the breakthroughs when overall health was declining, and the second semester of 1996, shortly after the introduction of HAART when the trend in average health was reversed. The six panels of Figure 10 display average population CD4 count, ailments, and survival along with paths for the two start dates until the end of the sample period in 2008.[27]

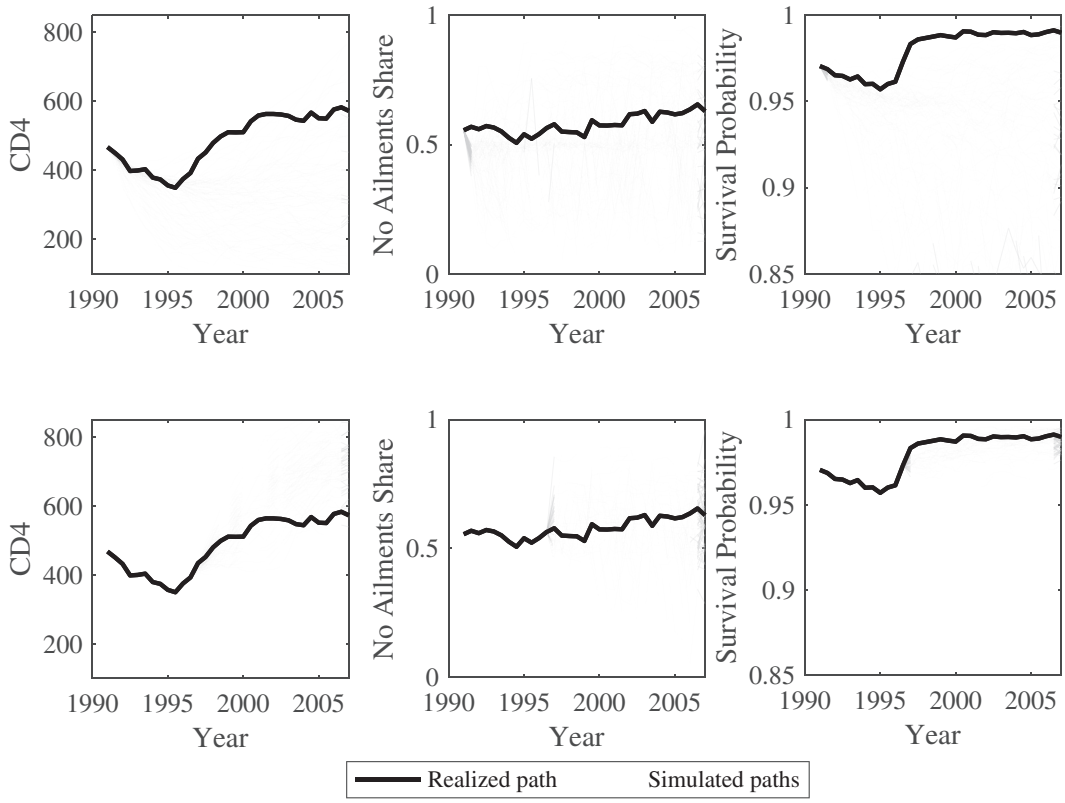
Strictly interpreted within the confines of the estimated model, the top three panels show that the improvements in health and survival rates immediately following the introduction of HAART in 1996 are extraordinary, but the reduction in ailments is much closer to central tendencies of the simulation process. The abrupt departure by the historical path from the mass covered by the simulation process should not be construed as a model misspecification. HAART was widely hailed as a revolutionary discovery at the time, underlining our opinion that it is useful to treat technological progress as a nonstationary process allowing for breakthroughs.

The bottom panels in Figure 10 displaying simulations beginning in 1996 convey a more nuanced picture. Consumers do not expect great improvements in abating ailments, but nevertheless entertain the possibility that on this dimension, they might get much better, or much worse: in reality, modest improvements occur. However, the model also predicts that given the state of technology at the end of 1996, in the 12 years, following health will improve more than, but survival rates less than, what actually occurs. Not only was HAART unpredictable (falling outside the 95% confidence band), but so were the years following, to a lesser degree, even though innovations after 1996 proceeded in relatively small steps.

6.3. *The Social Benefit of Clinical Trials.* From (13) and (15), the externality associated with the REE distorts all the choices. Nonetheless, we focus on just one alternative, experimental treatments. Aside from implementing an optimal subsidy for experimental treatments, social welfare can be further improved only by changing the market shares of commercially

[26] In the years just prior to HAART introduction, the efficacy of commercially available treatments had increased, pushing up the reference point for innovation and thus attracting more individuals into clinical trials. See our animated appendix.

[27] Other simulated quantities can be found in Figure A.4 in Appendix A.15.2.



NOTES: One hundred simulated paths conditional on the state of the world at 91-S1 and 96-S2.

FIGURE 10

DISTRIBUTION OF TECHNOLOGY PATHS: INDIVIDUALS

available products. Although the entire medical treatment industry is heavily regulated, experimental treatments are among the most regulated, giving greater scope for public health authorities to affect their usage. In addition, because less healthy consumers are more likely to participate in trials, policies that induce innovation by subsidizing experimental treatments tend to favor the most disadvantaged.

Table 5 summarizes two measures of the externality: a marginal measure evaluated at the REE, which we present in this subsection, and a second measure capturing the welfare gains from a subsidy policy, which we present in the next subsection. Both measures are evaluated at two points in time: 1991 (first semester) and 1996 (second semester). Recall that from Figure 1(c), these points bracket the spike that occurred in experimental treatments when HAART was discovered; the share of the experimental treatment in these two semesters is comparable at 0.10 and 0.09, respectively.

As a point of reference, we provide benchmark utilities for the REE in the first panel of Table 5. As a consequence of the breakthrough, average lifetime utility, normalized to wealth, increased by \$9,000 between these two points in time. The second panel shows the differences in lifetime utility between the *sicker*, those with a CD4 count below two hundred ($h_t < 200$), and the *healthier* ($h_t \geq 200$). The improvement for sicker consumers, \$39,000, is much more pronounced than for healthier consumers, \$12,000. Technology alone closes the gap in lifetime utility between the healthier and the sicker by about 28%.

The third line in the first panel of Table 5 presents our marginal measure of the externality. Increasing the market share of the experimental treatment by a 10th of a percentage point raises expected lifetime utility by \$323 in 1991 (from a base of \$344,000) and by \$238 in 1996

TABLE 5
OPTIMAL DEMAND FOR EXPERIMENTAL TREATMENTS

	Policy Introduced at	
	91-1	96-2
REE share of experimental treatment, s_{et}	0.102	0.092
REE average lifetime utility	344	353
Change on average lifetime utility from adding 0.001 to s_{et}	0.323	0.238
REE average lifetime utility (*sicker*)	266	305
REE average lifetime utility (*healthier*)	361	373
Gap (*healthier - sicker*)	95	68
SPP share of experimental treatment, s^*_{et}	0.113	0.175
SPP average lifetime utility	345	359
Flat subsidy required to decentralize s^*_{et}	2.5	15
Flat tax required to fund the subsidy	0.282	2.629
SPP average lifetime utility (*sicker*)	267	315
SPP average lifetime utility (*healthier*)	363	377
Gap (*healthier - sicker*)	96	62

NOTES: Social planning problem solved at the first semester of 1991 and the second semester of 1996. REE stands for rational expectations equilibrium; SPP stands for social planning problem. Monetary values in thousands of real $U.S. indexed to 2,000. *Sicker* (*Healthier*) individuals are those with CD4 counts below (at or above) 200.

(from a base of $353,000).[28] Both positive numbers imply that on the margin, there are social gains from increasing consumption of experimental treatments in clinical trials at the REE.

6.4. *Subsidizing Consumers to Participate in Clinical Trials.* Policy counterfactuals that affect the primitives generating both the demand and supply curves cannot be analyzed in our model, because we only estimate a reduced form of supply for the equilibrium generating the data. Confining the counterfactual analysis to a one-period shift, however, inoculates our results against this limitation because there is no supply response to unanticipated policy changes that only last one period; the estimated reduced-form supply process applies before and after the temporary intervention. The last exercise we conduct is to estimate the potential welfare gains from subsidizing the consumption of experimental treatment with a flat subsidy at different rates for one period.

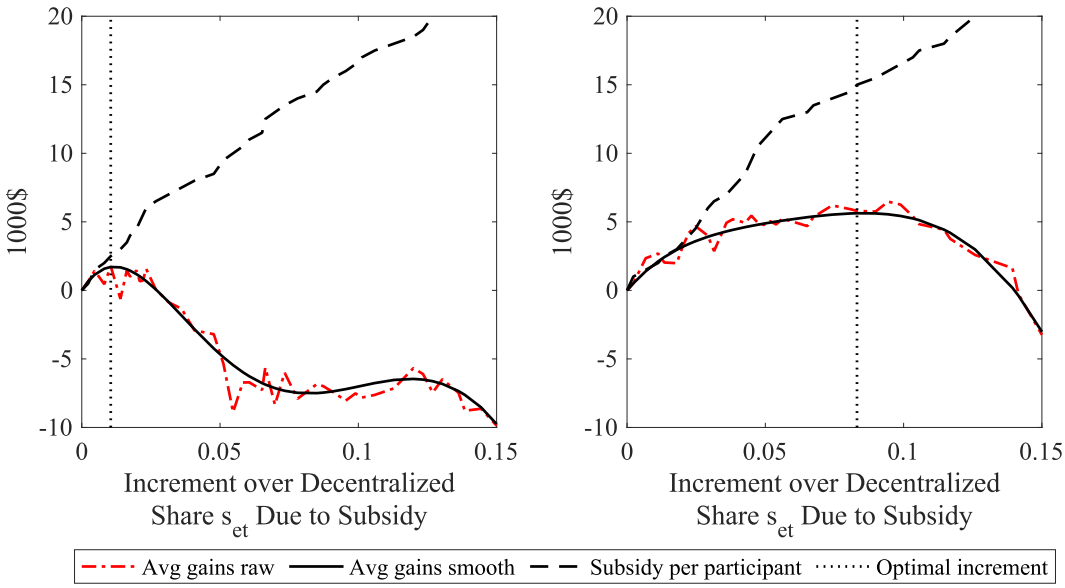
Table 5 and Figure 11 present the results from this exercise at the first semester of 1991 and again at the second semester of 1996.[29] The dashed line in Figure 11 plots the amount of the subsidy per consumer of the experimental treatment required to attain a given increase in the market share of the experimental treatment from the REE (calibrated on the horizontal axis). The subsidy required to achieve a given increase in the level of adoption in 1996 does not differ much from the rate required to achieve a similar increase in 1991.

The solid line displays the average net gains as a function of the increase in market share; it smooths the point-dash line crisscrossing it, the predicted experimental treatment share calculated for each level of subsidy separately.[30] The vertical dotted line marks the optimal one period intervention; the third panel in Table 5 reports the estimates. In 1991, the planner's optimal share of consumers using the experimental treatment is approximately the same as the decentralized share. The utility costs of increased consumption of experimental treatments quickly outweigh the marginal benefits from increasing the share of the experimental treatment (e.g., through speeding up innovation) in a time when individuals are very sick, no good

[28] These numbers are obtained by multiplying the derivative of the average lifetime utility function, evaluated at the REE, by 0.001.

[29] We discretize the share of the experimental treatment in increments of 0.005 units and simulate aggregate lifetime utility 1,000 times for each value.

[30] The solid line in Figure 11 applies a fifth degree local smoothing polynomial over the point-dash line representing the predicted experimental treatment calculated for each level of subsidy. We use the smoothed version when evaluating marginal gains and the optimal increment. Appendix A.15.4 contains further details.



NOTES: On the left (right) panel is the first (second) semester of 1991 (1996). On the x-axis are increments in the experimental treatment's share over the decentralized share s_{et}. The solid line represents average gains in welfare over the decentralized allocation. The dashed line indicates the subsidy per participant necessary to decentralize a given increment. The dotted line indicates the planner's optimal increment over s_{et}.

FIGURE 11

OPTIMAL ASSIGNMENT OF EXPERIMENTAL TREATMENT WITH A FLAT SUBSIDY
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

treatments have been invented and previous innovations have been small. In 1991, the optimal subsidy is $2,500 financed with a lump-sum tax of $282 on HIV+ individuals; it raises average lifetime utility by $1,000. By 1996, large innovations have occurred, our estimates of the innovation process show that further innovations are therefore more probable, and consumer health is rapidly improving. Compared to 1991, the optimal subsidy is six times as large and the flat tax financing it is about 10 times higher in 1996; however, average lifetime utility increases by about $6,000 to $359,000.

Not only are average net benefits, the subsidy rate and financing costs greater in 1996 than in 1991; so is the change in the distribution of benefits in the bottom panel of Table 5. Healthier consumers gain an estimated $2,000 from a 1991 intervention, but sicker consumers only an estimated $1,000, compared to a gain of 10 times that amount from a 1996 intervention, healthier consumers gaining $4,000. Thus, the gap between the two groups would increase slightly in 1991, but shrink by about $6,000 in the event of a 1996 intervention. Note though that the decline in inequality is not at the expense of healthier consumers. Equity increases because the sicker benefit the most from faster innovation. Consequently, the subsidy reduces free-riding by healthier consumers that occurs predominantly at the expense of the sicker.

Figure 11 also shows that in 1996, the optimal experimental treatment share lies 8 percentage points above its REE counterpart, almost twice as high, generating welfare gains that outweigh individual losses due to consumption of experimental treatments. In 1991, however, increasing the share by more than about 1 percentage point yields net losses. These differences are reflected in the subsidy rates: the figure shows that any subsidy less than the optimal one increases welfare, which imparts considerably more flexibility in setting a welfare-improving subsidy rate in 1996 (since the optimal subsidy is $15,000) than in 1991.

One last comparison is illuminating: we find the discovery of HAART improves average welfare by about $9,000; correcting the experimental share for the distortion in the REE only increases welfare by one-ninth of that amount in 1991, but by two-thirds in 1996. These results

suggest that the potential benefits from ameliorating externalities by subsidizing the consumption of experimental treatment vary substantially in the course of developing medical treatments.

7. CONCLUSION

We provide a framework to assess how consumer choices affect technological progress. In our case, aggregate consumer demand affects not only the speed of innovation, but also the direction of innovation in cases where product quality is multidimensional. We apply our framework to study consumer behavior and innovation in the market for HIV drugs. We capture several mechanisms through which consumer demand affects innovation, including experimentation with new drugs by participating in clinical trials, which accelerates the entry and increases the expected quality of new treatments. We show that individually, optimal consumer behavior can slow the process of innovation due to a distaste for experimentation. Moreover, individuals do not internalize the consequences of their treatment choices on other consumers' welfare, implying an externality that arises through the impact on technological progress. Our estimates show that providing incentives for consumption of experimental treatments can improve social welfare and equity.

APPENDIX A

A.1 *Data Appendix.* Data collection for the MACS started in 1984 with 4,954 men enrolled.[31] Two more enrollments have taken place: one in 1987–1991 (668 additional men) and another in 2001–2003 (1,350 additional men). We only use data from the first two enrollments. Since data are semiannual, each period t corresponds to six months. Below we describe the main variables we use in our study:

Health (h_{it}): At every visit, individuals undertake a physical examination that includes a blood sample that provides a measure of underlying health status: the individual's CD4 count. We denote as h_{it} the CD4 count of the individual at the start of period t. According to the official U.S. government's website for HIV:[32]

> The CD4 count is [...] a snapshot of how well your immune system is functioning. CD4 cells (also known as CD4+ T cells) are white blood cells that fight infection. [...] These are the cells that the HIV virus kills. As HIV infection progresses, the number of these cells declines. When the CD4 count drops below 200 [cells per microliter] due to advanced HIV disease, a person is diagnosed with AIDS. A normal range for CD4 cells is about 500-1,500.

Ailments (y_{1it}): Starting at visit 4, individuals are asked about physical symptoms. We focus on unusual bruises lasting at least two weeks, unintentional weight loss of at least 10 pounds, fatigue, diarrhea, fever, night sweats, and tender/enlarged glands. Individuals are asked if they have felt each of the ailments for at least three days during the period. Although individuals are asked explicitly about side effects starting at visit 13, we choose not to use this part of the data because it lacks consistency over time and more importantly, because individuals are most likely unable to correctly distinguish between side effects and symptoms. In our model, y_{1it} takes the value of 1 if an individual reports having any of the problems mentioned above.

Labor supply (y_{2it}): Whether the individual worked full time (35 hours or more per week) during period t.

[31] Data in this manuscript were collected by the MACS with centers (Principal Investigators) at The Johns Hopkins Bloomberg School of Public Health (Joseph B. Margolick, Lisa P. Jacobson), Howard Brown Health Center, Feinberg School of Medicine, Northwestern University, Cook County Bureau of Health Services (John P. Phair, Steven M. Wolinsky), University of California, Los Angeles (Roger Detels), and University of Pittsburgh (Charles R. Rinaldo). The MACS is funded by the National Institute of Allergy and Infectious Diseases, with additional supplemental funding from the National Cancer Institute. UO1-AI-35042, 5-MO1-RR-00052 (GCRC), UO1-AI-35043, UO1-AI-35039, UO1-AI-35040, UO1-AI-35041. Website located at http://www.statepi.jhsph.edu/macs/macs.html.

[32] See https://www.hiv.va.gov/patient/diagnosis/labs-CD4-count.asp

TABLE A.1
SUMMARY STATISTICS BY TREATMENT CHOSEN: SUBJECT VISITS, 1990–2007

	Full Sample	Experimental	Commercial	No Treatment
Observations	16851	1106	10980	4765
Ailments	0.43	0.54	0.46	0.32
Labor participation	0.63	0.60	0.60	0.71
Age	44.5	44.4	45.6	41.9
	(8.0)	(7.5)	(8.0)	(7.6)
CD4	475	385	456	540
	(297)	(304)	(291)	(298)
Gross income	17567	18054	17449	17727
	(8787)	(8677)	(8801)	(8776)
Out-of-pocket expenditures	266	282	337	98
	(706)	(986)	(771)	(360)

NOTES: Standard deviation in parentheses. Gross income and out-of-pocket expenditures are semestral and measured in real $U.S. indexed to 2000.

Income (y_{3it}): Starting at visit 14, individuals answer the question "Which of the following categories describes your annual individual gross income before taxes?" For visit 14, categories are brackets that increase every $10,000, the last category being censored at "$70,000 or more." For visits 15–35, the brackets are censored at $50,000, and for visits 36–41, the brackets are censored at $60,000. We censor at $50,000 to obtain a uniform question over time. Then we assign the middle point to individuals in the bracket. For the highest bracket, we assign the upper limit ($50,000). We divide gross income by two since our periods are half-years. Gross income as well as out-of-pocket expenditures (below) are in real $U.S. indexed to 2000.

Out-of-pocket expenditures (y_{4it}): Starting at visit 14, individuals are asked a version of the following question: "Please, estimate the TOTAL out-of-pocket expenses that you or other personal sources (your lover, family or friends) paid for prescription medications since your last visit." This question is open so values are not categorized.

Demographics (a_{it}): Individuals are either white, black, or Hispanic, and their age increases by half a year every period.

A.1.1 *Products and product components.* Starting at visit 6, individuals are asked about their medication. From visit 13 forward, as the number of treatments available increases, they answer separate survey modules for ARV drugs and non-antiretroviral drugs (NARVs). We focus on ARVs since these are the drugs used to treat HIV infection. Below we provide the empirical definition of experimental and commercial treatments that we use in the article.

Experimental treatment. Individuals are asked to name specifically which drugs they took as well as whether or not they took the drug as part of a research study. In the original data, some of the reported drugs are themselves coded as trials. We regard these instances as individuals consuming an experimental treatment. If an individual consumes at least one of his drugs as part of a clinical trial, we regard the individual as consuming an experimental treatment in that period.

Commercial treatment. We define a commercial treatment as a combination of product components where no component is consumed in a clinical trial. (See Table A.2.) This definition generates 1,835 treatments. We reduce the number of commercial treatments using the following algorithm:

1. We start with the set of treatments that have more than 40 observations in the sample and denote this the set of "core commercial treatments."[33] Our core commercial treat-

[33] We tried different criteria for the minimum number of observations and treatment classification did not change substantially. Since our definition of core commercial treatments can miss treatments appearing near the end of the time period studied, we select the core products using all periods but exclude the last four periods from estimation.

TABLE A.2
CHEMICAL FORMULAS OF PRODUCT COMPONENTS

Component	Short Name	Chemical Formula
Isoprinosine	IAD	$C_{52}H_{78}N_{10}O_{17}$
Ribavirin	RBV	$C_8H_{12}N_4O_5$
Interferons (α/β)	IFNs	
Zidovudine	AZT	$C_{10}H_{13}N_5O_4$
Zalcitabine	DDC	$C_9H_{13}N_3O_3$
Egg lecithin	AL-721	
Acyclovir	ACV	$C_8H_{11}N_5O_3$
Didanosine	DDI	$C_{10}H_{12}N_4O_3$
Stavudine	D4T	$C_{10}H_{12}N_2O_4$
Nevirapine	NVP	$C_{15}H_{14}N_4O$
Lamivudine	3TC	$C_8H_{11}N_3O_3S$
Saquinavir	SQV	$C_{38}H_{50}N_6O_5$
Ritonavir	RTV	$C_{37}H_{48}N_6O_5S_2$
Indinavir	IDV	$C_{36}H_{47}N_5O_4$
Nelfinavir	NFV	$C_{32}H_{45}N_3O_4S$
Lopinavir	LPV	$C_{37}H_{48}N_4O_5$
Abacavir	ABC	$C_{14}H_{18}N_6O$
Efavirenz	EFV	$C_{14}H_9CIF_3NO_2$
Tenofovir	TDF	$C_9H_{14}N_5O_4P$
Emtricitabine	FTC	$C_8H_{10}FN_3O_3S$
Atazanavir	ATV	$C_{38}H_{52}N_6O_7$
Fosamprenavir	FPV	$C_{25}H_{36}N_3O_9PS$
Darunavir	DRV	$C_{27}H_{37}N_3O_7S$
Raltegravir	RAL	$C_{20}H_{21}FN_6O_5$

NOTES: Consulted in PubChem (August, 2020) https://pubchem.ncbi.nlm.nih.gov

ments
are listed in Table 1 that shows that there are 70 core commercial treatments overall with at most five components. Out of 20,142 subject-visit observations of individuals taking commercial treatments, 13,767 are covered by treatments classified as core commercial treatments.

2. We code the remaining 6,375 observations of noncore commercial treatments as core commercial treatments using the steps below. Each step sequentially assigns the remaining observations that were not assigned in previous steps.
   (a) Noncore commercial treatment k is assigned to core commercial treatment k' if k' is the core commercial treatment with the highest number of components that is contained by k. Of the remaining 6,375 observations of noncore commercial treatments, this rule assigns 2,963 uniquely and leaves 3,412 unassigned (1,647 that were assigned to multiple core commercial treatments plus 1,765 that were not assigned to any core commercial treatment).
   (b) If assigned to multiple core commercial treatments in step (a):
      i. First, we use the past history of the individual. If at period t, the individual is consuming noncore commercial treatment k'' that was assigned to both core commercial treatments k and k' in step (a), and he was observed consuming core commercial treatment k in period $t-1$, then his treatment at t is recoded as k. We repeat this procedure until no further gains are obtained. Out of the remaining 1,647 observations assigned to multiple core commercial treatments, 428 are assigned uniquely in this step.
      ii. Second, we use the future history of the individual. If at period t, the individual is consuming noncore commercial treatment k'' that was assigned to both core commercial treatments k and k' in step (a), and he was observed consuming core commercial treatment k' in period $t+1$, then his treatment at t is recoded as k'. We

repeat this procedure until no further gains are obtained. Out of the remaining 1,219 observations assigned to multiple core commercial treatments, 274 are assigned uniquely in this step.

iii. Third, we use the core commercial treatment with the highest share at t. If at period t, the individual is consuming noncore commercial treatment k'' that was assigned to both core commercial treatments k and k' in step (a), and $s_{kt} > s_{k't}$, then his treatment at t is recoded as k. This final step assigns uniquely the remaining 945 observations assigned to multiple core commercial treatments.

(c) If not assigned to a core commercial treatment in step (a): we regard all 1,765 observations as "fringe treatments" since they do not contain any core commercial treatment. We aggregate all fringe treatments that appear at period t into one single "fringe mix," and assign to it all users consuming this product over time. We only consider fringe mixes that have at least 40 users. This reduces the number of observations by 345 (which represents 1.6% of the number of observations of individuals using a treatment). This aggregation leads to 16 fringe mixes that we pool with the set of core commercial treatments, which amounts to a total of 86 commercial treatments overall. (See Table A.3.)

3. In the model, we specified that a treatment gets withdrawn from the market altogether when its share falls below $\bar{s}$ for two consecutive periods. However, in the data, a treatment may have a share below $\bar{s}$ for more than two consecutive periods and then reappear again. Seventy-eight out of 86 core commercial treatments have unique spells without "reappearance." We regard the remaining treatments with multiple spells as measurement error and follow the next procedure to ensure that treatments have unique spells without reappearance. For every core commercial treatment k with reappearance:

(a) We identify all spells that treatment k has in the data. That is, we identify the first spell and all reappearances.

(b) From those spells, we select the one that contains the period t' in which $s_{kt'}$ was the highest. We drop all observations of individuals consuming commercial treatment k in other spells.

Out of 19,797 observations of individuals taking commercial treatments (20,142 minus 345 from step 2(c)), this smoothing procedure drops 42 observations leaving 19,755 observations of individuals taking commercial treatments. Supporting the importance of the spells selected by this procedure, the maximum share in the selected spell is on average 24 times larger than the maximum share in other spells of the same commercial treatment.[34] Tables 1 and A.3 include entry and exit dates implied by this spell smoothing procedure.

A.2 *Model Parameterization Appendix.*

A.2.1 *Evolution of the choice set.* In this section, we provide further details of the law of motion of the set of available treatments as well as its empirical implementation.

Differential treatment effects. In the model, treatment characteristics do not vary across races. Since we have limited data to fully relax this assumption, we explore here the possibility of differential treatment characteristics across races in a limited fashion. Following the same data criteria as in Appendix A.1, we first find the subset of treatments with at least 40 black users and the subset of treatments with at least 40 Hispanic users. We then interact the characteristics of those treatments in Equations (22) and (23) with race/ethnicity. The *Fixed* columns of Table A.4 present the estimates of the model with no interactions. (See also Table A.15.)

[34] In addition to this procedure, we tried (i) selecting the spell with the highest average share and (ii) selecting the spell with the highest sum of shares. All criteria result in very similar entry and exit dates.

TABLE A.3
COMMERCIAL TREATMENTS: FRINGE MIXES

Treatment	Companies	Entry	Exit	Treatment	Companies	Entry	Exit
IAD/RBV/IFNs (α/β)	LP/ICN/G	87-1	97-1	NVP/3TC/LPV/RTV/TDF	BI/GSK/AB/AB/GI	03-1	-
IFNs (α/β)/3TC/SQV/IDV/EFV	GW/GW/H/M/DP	97-1	07-1	3TC/LPV/RTV/ABC/TDF/ATV	GSK/AB/AB/GSK/GI/BMS	04-1	-
NVP/3TC/SQV/RTV/IDV	BI/GW/H/AB/M	97-2	06-2	RTV/TDF/FTC/ATV/FPV	AB/GI/GI/BMS/GSK	04-2	-
NVP/3TC/SQV/RTV/NFV	BI/GW/H/AB/AG	98-1	06-2	SQV/RTV/TDF/FTC/ATV	H/AB/GI/GI/BMS	05-1	-
NVP/SQV/RTV/ABC/EFV	BI/H/AB/GW/DP	99-1	05-2	3TC/RTV/ABC/TDF/ATV/FPV	GSK/AB/GSK/GI/BMS/GSK	05-2	-
NVP/RTV/NFV/ABC/EFV	BI/AB/AG/GW/DP	99-2	–	SQV/RTV/ABC/TDF/FTC	H/AB/GSK/GI/GI	07-1	-
NVP/LPV/RTV/ABC/EFV	BI/AB/AB/GSK/BMS	01-2	08-2	3TC/RTV/TDF/FTC/RAL	GSK/AB/GI/GI/JJ	08-1	-
NVP/3TC/NFV/ABC/TDF	BI/GSK/AG/GSK/GI	02-2	–	RTV/TDF/FTC/DRV/RAL	AB/GI/GI/JJ/JJ	08-2	-

NOTES: When displaying the components of fringe mixes, we only include up to the six most used components in the mix. Entry and exit dates obtained using the algorithm in Subsection A.1.1. Entry and exit dates are displayed in format year-semester (YY-S). Many products had not exited by the end of the sample. The *Companies* column displays the firm who owned the drug component at the moment of treatment entry. Company acronyms are: Burroughs–Wellcome (BW), Glaxo (G), Bristol–Myers Squibb (BMS), Hoffmann–La Roche (H), Glaxo Wellcome (GW), Abbott (AB), Merck (M), Boehringer Ingelheim (BI), Agouron (AG), DuPont (DP), GlaxoSmithKline (GSK), Gilead (GI), Leo Pharmaceutical (LP), ICN Pharmaceuticals (ICN), and Janssen (JJ).

TABLE A.4
ALLOWING FOR DIFFERENTIAL TREATMENT CHARACTERISTICS BY RACE

	Health, $\theta^{(health)}$							
	Fixed		Interacted Model					
			Baseline		Black		Hispanic	
Treatment	est.	se	est.	se	est.	se	est.	se
AZT	−12.004	(2.697)	−13.083	(2.836)	11.620	(7.630)	−1.429	(10.132)
AZT/ACV	−12.752	(4.764)	−12.524	(5.270)	−0.580	(14.031)	−7.996	(15.890)
DDI	15.263	(4.574)	10.500	(5.007)	18.206	(13.458)	16.269	(15.758)
D4T	39.776	(6.299)	35.932	(7.420)	28.900	(17.128)	−7.342	(14.854)
AZT/3TC	34.398	(6.227)	33.008	(7.076)	7.570	(15.004)		
AZT/3TC/IDV	65.041	(6.220)	63.981	(6.656)	50.301	(27.399)	−38.594	(17.388)
AZT/NVP/3TC	46.275	(7.123)	48.631	(7.310)	−14.614	(22.904)		
D4T/NVP/3TC	46.846	(9.161)	39.110	(11.207)	12.617	(21.152)		
AZT/3TC/NFV	50.776	(10.417)	55.120	(12.008)	−16.881	(23.696)		
D4T/3TC/NFV	48.018	(10.212)	42.722	(10.330)	31.994	(35.497)		
AZT/3TC/EFV	43.526	(5.327)	50.587	(6.394)	−22.661	(11.775)	−2.925	(23.985)
3TC/ABC/EFV	53.341	(8.516)	53.483	(9.367)	−1.891	(22.681)		
AZT/3TC/ABC	54.824	(10.999)	43.748	(13.699)	32.510	(22.542)		
AZT/3TC/LPV/RTV	49.838	(12.997)	50.021	(14.166)	−0.718	(30.172)		
3TC/EFV/TDF	47.790	(10.024)	46.582	(11.379)	11.952	(24.054)		
3TC/LPV/RTV/TDF	51.672	(11.705)	40.431	(11.734)	35.023	(28.730)		
3TC/ABC/EFV/TDF	31.846	(13.014)	29.723	(15.283)	7.040	(28.846)		
AZT/3TC/LPV/RTV/ABC	9.855	(14.503)	31.029	(27.445)	−29.062	(32.178)		
EFV/TDF/FTC	54.798	(4.453)	56.572	(5.682)	−2.546	(9.855)	−8.747	(12.511)
RTV/TDF/FTC/ATV	53.028	(5.309)	57.060	(6.708)	−11.569	(11.726)	−4.244	(16.299)
LPV/RTV/TDF/FTC	46.723	(7.767)	48.031	(10.202)	−3.566	(15.270)		
	Ailments, $\theta^{(ail)}$							
	Fixed		Interacted Model					
			Baseline		Black		Hispanic	
Treatment	est.	se	est.	se	est.	se	est.	se
AZT	−0.500	(0.041)	−0.570	(0.044)	0.491	(0.126)	0.427	(0.163)
AZT/ACV	−0.539	(0.080)	−0.653	(0.089)	0.685	(0.269)	0.556	(0.280)
DDI	−0.375	(0.071)	−0.403	(0.081)	−0.072	(0.189)	0.520	(0.261)
D4T	−0.717	(0.092)	−0.721	(0.105)	0.083	(0.257)	−0.130	(0.317)
AZT/3TC	0.064	(0.094)	0.056	(0.104)	0.448	(0.263)		
AZT/3TC/IDV	−0.075	(0.081)	−0.030	(0.090)	−0.026	(0.274)	−0.062	(0.351)
AZT/NVP/3TC	0.109	(0.111)	0.069	(0.120)	0.237	(0.311)		
D4T/NVP/3TC	−0.386	(0.121)	−0.511	(0.150)	0.635	(0.274)		
AZT/3TC/NFV	−0.432	(0.122)	−0.635	(0.139)	0.830	(0.291)		
D4T/3TC/NFV	−0.881	(0.130)	−0.967	(0.145)	0.484	(0.352)		
AZT/3TC/EFV	0.342	(0.085)	0.369	(0.106)	−0.101	(0.186)	−0.024	(0.359)
3TC/ABC/EFV	0.108	(0.125)	0.065	(0.139)	0.400	(0.346)		
AZT/3TC/ABC	−0.442	(0.128)	−0.582	(0.158)	0.415	(0.267)		
AZT/3TC/LPV/RTV	−0.655	(0.180)	−0.627	(0.217)	−0.095	(0.387)		
3TC/EFV/TDF	−0.011	(0.154)	0.103	(0.178)	−0.387	(0.366)		
3TC/LPV/RTV/TDF	−0.092	(0.174)	−0.070	(0.212)	−0.084	(0.369)		
3TC/ABC/EFV/TDF	−0.308	(0.178)	0.057	(0.219)	−1.200	(0.398)		
AZT/3TC/LPV/RTV/ABC	0.298	(0.274)	−1.476	(0.506)	2.608	(0.637)		
EFV/TDF/FTC	0.118	(0.067)	0.173	(0.083)	0.000	(0.157)	−0.460	(0.191)
RTV/TDF/FTC/ATV	0.138	(0.090)	0.079	(0.118)	0.384	(0.203)	−0.436	(0.266)
LPV/RTV/TDF/FTC	−0.183	(0.135)	−0.526	(0.165)	0.972	(0.287)		

NOTES: Treatment characteristics are estimated as indicators for treatment usage in (A.12) and (A.13). Only the characteristics of treatments that are interacted with race/ethnicity are shown here, all other estimated coefficients are omitted. The first panel corresponds to health characteristics, and the second panel corresponds to ailments characteristics. In the table, "est." stands for estimate and "se" stands for standard errors, in parenthesis. The *Fixed* columns correspond to the estimates of treatments characteristics not allowing for race interactions (Table A.15). The *Interacted Model* columns correspond to the estimates when treatment characteristics are interacted with race/ethnicity.

TABLE A.5
MEASURES OF ADHERENCE BY RACE

	(1)		(2)		(3)		(4)		(5)	
	est.	se	est.	se	est.	se	est.	se	est.	se
Black	−0.278	(0.027)	0.124	(0.017)	−19.364	(2.713)	−0.165	(0.046)	−0.004	(0.008)
Hispanic	−0.027	(0.043)	−0.042	(0.028)	6.918	(4.353)	0.087	(0.075)	−0.014	(0.012)
Observations	6071		5708		5708		3498		6082	

NOTES: Linear regressions of five measures of adherence on race/ethnicity. Controls for health, education, age, and prior labor participation are also included in the regression but omitted from the table. Measure (1) is a ranking from 1 (never) to 5 (always) indicating how closely individuals follow treatment schedule. Measure (2) is an indicator for whether individuals skipped medication. Measure (3) is the number of days without skipping medication. Measure (4) is an indicator for whether individuals followed special indications, provided that such indications were given. Measure (5) is an indicator for whether individuals took fewer pills than prescribed in any of the antiretroviral drugs prescribed. In the table, "est." stands for estimate and "se" stands for standard errors, in parenthesis.

The *Interacted Model* columns correspond to the estimates when treatment characteristics are interacted with race. All other coefficients that are not interacted are omitted from the table.

The results show that the race interactions of effectiveness, that is, the health characteristic, are not statistically significant for virtually all the treatments in the table. Moreover, the estimated health characteristics of the baseline in the interacted model are very similar to those of the model with no interactions. The results are slightly less clear-cut for the ailments characteristic. About one-third of the treatments in the table have statistically significant interactions with the black indicator. Moreover, out of eight treatments with a sufficiently large number of Hispanic users, four have statistically significant interactions with the Hispanic indicator. Overall, there seems to be no variation in effectiveness across race/ethnicity and only weak variation in side effects. The race gradient of side effects could be caused by behavioral differences such as adherence.

To explore differences in adherence by race, we exploit limited adherence data coming from a battery of questions that were added to the MACS survey in the first semester of 1999. We compare adherence using five measures pertaining to the six months prior to when the questions are asked:

1. *How closely did you follow treatment schedule?* Answers range from 1 (never) to 5 (all the time).
2. *Did you skip medication?* We create an indicator variable such that "Yes" = 1.
3. *Number of days without skipping medication*. For individuals who did not skip medication, this takes the value of 180. For other individuals, we calculate this number from their answer to the question: *When was the last time you skipped medication?*
4. *Did you follow special indications?* This question is only asked among individuals who declared that they were given special indications with their treatment. We create an indicator variable such that "Yes" = 1.
5. *Did you take fewer pills than prescribed in any of the antiretroviral drugs prescribed?* We create an indicator variable such that "Yes" = 1.

Table A.5 presents the results from regressing the measures of adherence on race/ethnicity as well as other components of the individual state such as health, education, age, and prior labor participation. For the first four measures, the black indicator is positive and significant, and for the fifth measure, it is not significant. Focusing on measures 2–4, Table A.5 suggests that black individuals are 12.4% points more likely to skip medication, have on average 19.4 fewer days without skipping medication (the sample average is 100.3 days), and are 16.5% points more likely to not follow special indications, provided that such indications were given. The Hispanic indicator is not significant for any of the measures.

According to the results in Table A.4, most of the ailments interactions with the black indicator that are statistically significant (only a third are significant) are positive. These results

TABLE A.6
AUTOCORRELATION OF INNOVATION SHOCKS

	Ailments, $\bar{\nu}_t^{(ail)}$				*Health*, $\bar{\nu}_t^{(health)}$			
	(1) est.	se	(2) est.	se	(3) est.	se	(4) est.	se
$\bar{\nu}_{t-1}$	−0.101	(0.155)			0.229	(0.154)		
$\bar{\nu}_{-1}$			−0.026	(0.199)			0.321	(0.191)
Observations	43		27		43		27	

NOTES: In the table, "est." stands for estimate; "se" stands for standard errors, in parenthesis. Each regression contains a constant that is omitted for simplicity. Columns (1) and (2) regress the average ailments innovation shock on $\bar{\nu}_{t-1}^{(ail)}$ and $\bar{\nu}_{-1}^{(ail)}$, respectively. Columns (3) and (4) regress the average health innovation shock on $\bar{\nu}_{t-1}^{(health)}$ and $\bar{\nu}_{-1}^{(health)}$, respectively.

are consistent with black individuals displaying lower adherence and mechanically reporting lower side effects. Since adherence data are only available late in our sample period, we do not include this mechanism in the model.

Correlation in innovation shocks. We show empirically that the characteristics of new treatments are conditional displacements around recent prevalent technology (Equation (17)) where the random component is an innovation shock drawn from the distribution f_ν. This relation generates correlation between the characteristics of current and future treatments. Since new technologies draw from previous technologies in ways the centroid may not capture, it is conceivable that the innovation shocks themselves are correlated over time. We show here that this is not the case in our data.

Because in several periods, there are multiples draws for ν, we define $\bar{\nu}_t$ as the average over the innovation shocks of new treatments entering the market at t:

$$\bar{\nu}_t = \frac{1}{N_t}\left(\sum_{k:\ k\in\mathbf{K}_t,\ k\notin\mathbf{K}_{t-1}} \nu_{kt}\right), \tag{A.1}$$

where N_t is the number of new treatments at t. Because in some of the periods, no new treatments are introduced, we follow two different procedures to construct the lagged average innovation shock:

1. We substitute to zero the missing innovation shock averages from periods where there is no shock. We then regress $\bar{\nu}_t$ on $\bar{\nu}_{t-1}$.
2. We drop periods where there was no innovation shock and define $\bar{\nu}_{-1}$ as the last average innovation shock received prior to the current period. We then regress $\bar{\nu}_t$ on $\bar{\nu}_{-1}$.

As shown in Table A.6, in both of these procedures, we find the coefficient of the autoregression to be not significant at the at 10% level. Our assumption that the innovation shocks, ν, are drawn i.i.d. from f_ν is supported by the data.

The distribution of the number of new treatments. The number of new treatments N_t is distributed negative binomial with dispersion in the mean:

$$\begin{aligned} N_t &\sim Poisson\left(\mu_{t-1}^*\right); \qquad \mu_{t-1}^* \sim Gamma\left(1/\alpha_{t-1}^N, \alpha_{t-1}^N \mu_{t-1}\right) \\ \mu_{t-1} &= \exp(\phi_1^N \kappa_{t-1} + \phi_2^N s_{et-1}); \qquad \alpha_{t-1}^N = \exp(\phi_3^N + \phi_4^N \kappa_{t-1}), \end{aligned} \tag{A.2}$$

where the magnitude of previous innovations κ_{t-1} is defined in (19) and the scaling weights, which account for the fact that different characteristics may be measured in different scales, are given by the maximum innovations observed in the data:

$$\delta_r^{-1} \equiv \max_{k:\, k\in\mathbf{K}_\tau, k\notin\mathbf{K}_{\tau-1},\ \forall\tau>0} \left\{\theta_k^{(r)} - \omega_{\tau-1}^{(r)}\right\},\ \text{for } r \in \{ail,\ health\}. \tag{A.3}$$

The end of a treatment's life cycle. The exit rule $\{\underline{s}, \bar{s}\}$ introduced in Section 4 is defined as follows: Recall that the market share of treatment k can be decomposed into new $\underline{s}_{kt}$ and repeat $\bar{s}_{kt}$ consumers ($\underline{s}_{kt} + \bar{s}_{kt} = s_{kt}$). Define the conditional share of commercial consumers as

$$\tilde{s}_{kt-1} \equiv \frac{s_{kt-1}}{\sum_{k'\in\mathbf{K}_{t-1}} s_{k't-1}}, \tag{A.4}$$

and the conditional share for new commercial consumers as

$$\tilde{\underline{s}}_{kt-1} \equiv \frac{\underline{s}_{kt-1}}{\sum_{k'\in\mathbf{K}_{t-1}} \underline{s}_{k't-1}}. \tag{A.5}$$

No new consumers can access treatment k if $\tilde{\underline{s}}_{kt-1}$ falls below the critical number $\underline{s}$ for three consecutive periods. Treatment k reaches the end of its life cycle when $\tilde{s}_{kt-1}$ falls below the critical number $\bar{s}$ for two consecutive periods. The number of consecutive periods for each exit rule is chosen to match the data, where a single period of low demand does not always signal the end of a treatment's life cycle. This relaxation of the exit rule adds two state variables to the aggregate state of the problem, $\mathcal{E}_{t-1}^1$ and $\mathcal{E}_{t-1}^2$, which are indicators of to what extent the conditions for exit are binding:

$$\mathcal{E}_{kt}^1 = \mathbf{I}\{\tilde{\underline{s}}_{kt-1} < \underline{s}\}\left(\mathcal{E}_{kt-1}^1 + \mathbf{I}\{\tilde{\underline{s}}_{kt-1} < \underline{s}\}\right), \tag{A.6}$$

$$\mathcal{E}_{kt}^2 = \mathbf{I}\{\tilde{s}_{kt-1} < \bar{s}\}\left(\mathcal{E}_{kt-1}^2 + \mathbf{I}\{\tilde{s}_{kt-1} < \bar{s}\}\right), \tag{A.7}$$

where $\mathcal{E}_{kt_k}^1 = \mathcal{E}_{kt_k}^2 \equiv 0$. Exit for new consumers binds when $\mathcal{E}_{kt}^1 = 3$ and exit for all consumers binds when $\mathcal{E}_{kt}^2 = 2$.

A tractable choice set. The clustering rule $c(\mathbf{K}_t)$, which allows us to reduce the size of the choice set, is characterized as the solution to a k-means clustering algorithm. At every period t, the clusters $j = 1, \ldots, J$ are chosen to minimize:[35]

$$c(\mathbf{K}_t) = \sum_{j=1}^{J} \sum_{k\in\mathbf{K}_t} \mathbf{I}\{k \in j\} \left\| \theta_k - \theta_j^c \right\|^2, \qquad \theta_j^c \equiv \frac{\sum_{k\in\mathbf{K}_t} \mathbf{I}\{k \in j\}\theta_k}{\sum_{k\in\mathbf{K}_t} \mathbf{I}\{k \in j\}}, \tag{A.8}$$

where $\sum_{j=1}^{J} \mathbf{I}\{k \in j\} = 1$ for all $k \in \mathbf{K}_t$. The within-cluster assignment probability is given by:

$$q_{kjt}\left(k\middle|\mathbf{K}_{jt}\right) = \frac{\exp\left(x_{kt}^{(weight)}\gamma^{(weight)}\right)}{\sum_{k\in j}\exp\left(x_{kt}^{(weight)}\gamma^{(weight)}\right)}, \tag{A.9}$$

[35] See Duda and Hart (1973) and Andrew W. Moore's *K-means and Hierarchical Clustering* tutorial at http://www.cs.cmu.edu/∼awm/tutorials.html. (See Appendix A.8.2 for implementation details.)

TABLE A.7
CORRELATION BETWEEN INSURANCE, LABOR STATUS, AND HEALTH

variable	est.	se
$LaborParticipation_t$, y_{2t}	0.018	(0.005)
$1\{250 < h_{t-1} \leq 500\}$	−0.008	(0.006)
$1\{h_{t-1} > 500\}$	−0.016	(0.006)
Observations	16,521	

NOTES: Independent variable is an indicator for whether an individual is covered by insurance. In the table, "est." stands for estimate; "se" stands for standard errors, in parenthesis. Regression also controls for age, race/ethnicity, and education.

where $x_{kt}^{(weight)}$ includes a constant term, the ranking (within its cluster) of the characteristics of the treatment, the number of members in the cluster, whether the treatment is new, and several interactions. The vector of parameters $\gamma^{(weight)}$ is obtained from a nonlinear regression of within cluster shares $s_{kt|j}$ such that:

$$\mathbb{E}\left[s_{kt|j}|x_{kt}^{(weight)}\right] = \exp\left(x_{kt}^{(weight)}\gamma^{(weight)}\right), \qquad s_{kt|j} \equiv \frac{s_{kt}}{\sum_{k' \in \mathbf{K}_{jt}} s_{k't}}. \tag{A.10}$$

A.2.2 *Health insurance.* We provide here an assessment of the importance of insurance, a mechanism that is not included in our model. Using the insurance data from MACS, we created the indicator *insured* that takes the value of 1 if the individual is covered by at least one of the following: HMO, group private, individual private, Medicaid, Medicare, veteran's administration, champus/champva, and other. Out of 16,851 individual-visit observations in our sample 88% were covered by insurance, 10% were not covered, and the remaining 2% had missing insurance data.

We first test whether insurance is associated with labor or health status by regressing the *insured* indicator on labor market participation, health, and other components of the individual state. The results in Table A.7 indicate that although labor participation is significantly associated with insurance status, the effect is small. Labor participation is associated with a 1.8% point increase in the probability of insurance coverage. Health status is only significantly associated with insurance status for individuals with very high health (CD4 counts above 500). High health is associated with a small decrease in the probability of insurance coverage of 1.6% points.

Finally, we explore the role of insurance on out-of-pocket expenses for prescription medications. The second column of Table A.8 shows the estimates from our out-of-pocket process in equation (27) and the fourth column shows the estimates adding insurance to the regression. We find that insurance is associated with a statistically significant 25% decrease in the expenses of individuals getting commercial ARV treatment and has no significant effect on the ARV-related expenses of individuals getting experimental treatment. Insurance coverage is also associated with a statistically significant increase of 373 dollars in the amount of out-of-pocket expenses for prescription medication (conditional on having positive out-of-pocket expenses). This direct effect suggests that insured individuals may be accessing more NARV medications.

Because we have a rich measure of underlying health, we can confidently back out treatment characteristics. Therefore, we do not consider that our estimated treatment characteristics are affected by the omission of insurance. However, the results here suggest that the gap in nonpecuniary benefits between commercial treatment and no treatment may be overestimated; this would be the case if individuals not getting treatment were doing so due to lack of insurance as opposed to disutility caused by treatment. This would also imply that the disutility from ailments when getting treatment may be overestimated as well. Since the majority of

TABLE A.8
OUT-OF-POCKET EXPENDITURES AND INSURANCE

	Baseline		Interacted	
variable	est.	se	est.	se
h_t	−0.002	(0.0004)	−0.002	(0.0005)
$h_t^2/10^3$	0.009	(0.002)	0.009	(0.002)
$h_t^3/10^7$	−0.133	(0.033)	−0.132	(0.029)
$h_t^4/10^{10}$	0.090	(0.031)	0.090	(0.022)
$h_t^5/10^{14}$	−0.266	(0.134)	−0.265	(0.072)
$h_t^6/10^{18}$	0.279	(0.221)	0.278	(0.084)
Age_t	0.037	(0.004)	0.033	(0.006)
Age_t^2	−0.0002	(0.00004)	−0.0002	(0.0001)
Black	−0.240	(0.013)	−0.227	(0.021)
Hispanic	−0.119	(0.015)	−0.073	(0.024)
Some college	0.169	(0.016)	0.170	(0.026)
College	0.318	(0.018)	0.309	(0.033)
More than college	0.336	(0.019)	0.331	(0.030)
$CommercialTreatment_t$	0.429	(0.015)	0.541	(0.064)
$CommercialTreatment_t * Insured_t$			−0.136	(0.058)
$ExperimentalTreatment_t$	0.313	(0.018)	0.131	(0.125)
$ExperimentalTreatment_t * Insured_t$			0.172	(0.131)
$LaborParticipation_t$, y_{2t}	0.105	(0.009)	0.087	(0.016)
$NoAilments_t$, y_{1t}	−0.122	(0.008)	−0.120	(0.017)
$Insured_t$			0.373	(0.050)
Constant	−1.459	(0.098)	−1.671	(0.197)
σ^o	0.862	(0.026)	0.741	(0.113)

NOTES: In the table, "est." stands for estimate; "se" stands for standard errors, in parenthesis. The *Baseline* columns are identical to the ones in Table A.13 reporting the estimated out-of-pocket expenditures equation we use in our model. The *Interacted* columns are from a model that adds interactions with insurance coverage. $CommercialTreatment_t = d_{J+2,t} + \sum_{k=1}^{J} d_{kt}$. $Insured_t$ is an indicator for whether the individual is covered by insurance. Out-of-pocket prescription expenditures y_{4t} are measured in real $U.S. indexed to 2000. Health, h_t, is given by the CD4 count measured in hundreds of cells per microliter. To use all observations, the regression with insurance includes indicators for missing insurance data, which affects only 330 observations out of 16,851.

TABLE A.9
DISTRIBUTION OF USERS PER CLUSTER

	Number of Clusters		
Percentile	2	3	4
1*st*	10	3	1
5*th*	15	5	3
10*th*	17	10	5
25*th*	27.5	15	12
50*th*	41.5	27	19
75*th*	61.5	44	33
90*th*	146	82	65
Cluster-Periods	68	102	136

NOTES: This table shows various percentiles of the distribution of number of users per cluster, conditional on a given number of clusters per period. The last row displays the number of cluster-periods obtained for a given number of clusters per period.

observations in the sample have insurance coverage, we think that these concerns do not represent a critical issue for us.

A.2.3 *The modified value function.* At the beginning of t, the realization of treatment assignment for those who selected a cluster in the previous period is drawn using the within-clusters probabilities q_{kjt}, the realization of last period's experimental treatment

TABLE A.10
HEALTH EFFECTS ON FUTURE HEALTH AND AILMENTS

	Ailments, $\gamma^{(ail)}$		Health, $\gamma^{(health)}$	
variable	est.	se	est.	se
h_t	0.008	(0.0004)	1.152	(0.015)
$h_t^2/10^3$	−0.013	(0.001)	−0.519	(0.048)
$h_t^3/10^7$	0.109	(0.020)	4.375	(0.564)
$h_t^4/10^{10}$	−0.040	(0.012)	−2.016	(0.281)
$h_t^5/10^{14}$	0.054	(0.026)	2.803	(0.494)
Constant	−0.929	(0.041)	−5.874	(1.455)

Notes: Parameters estimated using (A.12) and (A.13). In the table, "est." stands for estimate; "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. Health, h_t, is given by the CD4 count measured in hundreds of cells per microliter.

TABLE A.11
LABOR SUPPLY, y_{2t}

variable	$\gamma^{(labor)}$	se
h_t	0.009	(0.0003)
$h_t^2/10^3$	−0.013	(0.001)
$h_t^3/10^7$	0.075	(0.006)
$h_t^4/10^{10}$	−0.013	(0.002)
Age_t	0.102	(0.008)
Age_t^2	−0.001	(0.0001)
Black	−0.168	(0.026)
Hispanic	−0.040	(0.040)
Some college	0.312	(0.034)
College	0.537	(0.033)
More than college	0.613	(0.035)
$LaborParticipation_{t-1},\ y_{2t-1}$	4.458	(0.026)
Constant	−5.914	(0.182)

NOTES: Estimates of the Logit model in (25). In the table, "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. Health, h_t, is given by the CD4 count measured in hundreds of cells per microliter.

characteristics is drawn; health h_{it}, ailments y_{1it-1}, income y_{3it-1}, out-of-pocket payments y_{4it-1}, survival b_{it}, and current labor supply y_{2it} are realized. The number of new treatments is drawn as well as their characteristics; commercial treatments finish their life cycle following the $\{\underline{s}, \bar{s}\}$ rule. Clusters of treatments are formed according to the clustering rule c. Under the empirical specification in Section 4, the aggregate state, z_t^a, includes the treatments remaining on the market $\mathbf{K}_t$, the centroid for innovation ω_t, the magnitude of previous innovations κ_t, the previous share of the experimental treatment s_{et-1}, and the joint distribution of consumer demographics (including previous consumption) H_t. The individual state is formed by idiosyncratic preference shocks ϵ_{it}, and $\bar{z}_{it}$; the latter consists of the aggregate state, denoted as z_t^a, together with a collection of individual-specific variables: health h_{it}, labor supply y_{2it}, recent usage $\theta_{J+2,it-1}$, demographics a_{it}, and productivity η_i. Individuals have rational expectations and zero measure in the population. They observe their current state and choose $j \in \{0, 1, \ldots, J+1+r_{it}\}$. Aggregate choices at t determine market shares. The individual's ex ante value function in the modified decentralized problem is:

$$V(\bar{z}_{it}) \equiv E\left\{ \sum_{\tau=t}^{\infty} \sum_{j=0}^{J+1+r_{it}} \beta^{\tau-t} d^e_{ji\tau} b_{i\tau} [u_j(h_{i\tau}, y_{i\tau}) + \epsilon_{ji\tau}] \middle| \bar{z}_{it} \right\}, \tag{A.11}$$

TABLE A.12
GROSS INCOME, y_{3t}

Variable	$\gamma^{(inc)}$	se
h_t	0.018	(0.001)
$h_t^2/10^3$	−0.064	(0.007)
$h_t^3/10^7$	1.138	(0.169)
$h_t^4/10^{10}$	−1.030	(0.204)
$h_t^5/10^{14}$	4.854	(1.304)
$h_t^6/10^{18}$	−11.270	(4.182)
$h_t^7/10^{20}$	0.101	(0.053)
Age_t	0.482	(0.033)
Age_t^2	−0.006	(0.0004)
Black	−5.534	(0.117)
Hispanic	−4.167	(0.202)
Some college	2.497	(0.137)
College	5.812	(0.150)
More than college	8.203	(0.150)
$LaborParticipation_t$, y_{2t}	5.738	(0.070)
$NoAilments_t$, y_{1t}	0.207	(0.025)
Constant	−2.095	(0.790)

NOTES: Estimates of (26). In the table, "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. Random effects regression of gross income on covariates. y_{3t} is measured in real $U.S. indexed to 2000. Health, h_t, is given by the CD4 count measured in hundreds of cells per microliter.

where $u_j(h_{i\tau}, y_{i\tau}) + \epsilon_{ji\tau} \equiv U_{j\tau}$ in Equation (28), and $d^e_{ji\tau}$ denotes the optimal choices at τ. Because individuals in the decentralized economy do not take into account the consequences of their actions (e.g., their consumption of experimental treatments or their adoption of treatments with certain characteristics) on treatment development and hence on other individuals' future payoffs, the aggregate process generates an externality.

A.3 *Estimation Appendix.*

A.3.1 *Treatment characteristics.* We estimate treatment characteristics using the larger sample (visits 6–49) thereby using all data available on previous health, individual treatment usage, and subsequent health and ailments. Estimation equations follow from (22) and (23):

$$h_{t+1} = \sum_{s=0}^{5} \gamma_s^{(health)} h_t^s + \sum_{k \in \mathbf{K}_t} \tilde{d}_{kt} \theta_k^{(health)} + d_{J+1,t} \theta_{et}^{(health)} + \epsilon_{ht}, \tag{A.12}$$

$$\Pr[y_{1t} = 0|\cdot] = \left(1 + \exp\left(\sum_{s=0}^{5} \gamma_s^{(ail)} h_t^s + \sum_{k \in \mathbf{K}_t} \tilde{d}_{kt} \theta_k^{(ail)} + d_{J+1,t} \theta_{et}^{(ail)}\right)\right)^{-1}, \tag{A.13}$$

where $\tilde{d}_{kt}$ is an indicator for the individual using treatment $k \in \mathbf{K}_t$ at period t. Along with estimates of treatment characteristics, (A.12) and (A.13) provide parameter vectors $\gamma^{(health)}$ and $\gamma^{(ail)}$ that describe the health transition in (22) and the process for physical ailments in (23).

A.3.2 *Clusters.* In our empirical implementation, we assume that there are J clusters every period. We implement the following version of the k-means algorithm. At every period t:

1. Select the treatments that are still available for new consumers at t. Denote this set of treatments $\mathbf{A}_t$.

TABLE A.13
OUT-OF-POCKET EXPENDITURES, y_{4t}

Variable	$\gamma^{(spend)}$	se
h_t	−0.002	(0.0004)
$h_t^2/10^3$	0.009	(0.002)
$h_t^3/10^7$	−0.133	(0.033)
$h_t^4/10^{10}$	0.090	(0.031)
$h_t^5/10^{14}$	−0.266	(0.134)
$h_t^6/10^{18}$	0.279	(0.221)
Age_t	0.037	(0.004)
Age_t^2	−0.0002	(0.00004)
Black	−0.240	(0.013)
Hispanic	−0.119	(0.015)
Some college	0.169	(0.016)
College	0.318	(0.018)
More than college	0.336	(0.019)
$CommercialTreatment_t$	0.429	(0.015)
$ExperimentalTreatment_t$	0.313	(0.018)
$LaborParticipation_t,\ y_{2t}$	0.105	(0.009)
$NoAilments_t,\ y_{1t}$	−0.122	(0.008)
Constant	−1.459	(0.098)
σ^o	0.862	(0.026)

NOTES: Estimates of (27) using a Tobit Model for data censored at 0. In the table, "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. $CommercialTreatment_t = d_{J+2,t} + \sum_{k=1}^{J} d_{kt}$. Out-of-pocket prescription expenditures y_{4t} are measured in real $U.S. indexed to 2000. Health, h_t, is given by the CD4 count measured in hundreds of cells per microliter.

TABLE A.14
DEATH, $1 - b_t$

Variable	$\gamma^{(live)}$	se
h_t	−0.028	(0.001)
$h_t^2/10^3$	0.079	(0.006)
$h_t^3/10^7$	−1.104	(0.144)
$h_t^4/10^{10}$	0.704	(0.175)
$h_t^5/10^{14}$	−1.610	(0.811)
Age_t	−0.116	(0.020)
Age_t^2	0.002	(0.0002)
Black	−0.509	(0.065)
Hispanic	0.034	(0.074)
Some college	0.060	(0.062)
College	−0.353	(0.062)
More than college	−0.512	(0.065)
$NoAilments_t,\ y_{1t}$	−1.140	(0.051)
Constant	1.682	(0.465)

NOTES: Estimates of the Logit model in (24). In the table, "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. Health, h_t, is given by the CD4 count measured in hundreds of cells per microliter.

2. To keep comparability, rescale the characteristics of all treatments available for clustering at t by computing:

$$\tilde{\theta}_k^{(r)} = \frac{\theta_k^{(r)}}{\max_{k \in \mathbf{A}_t} \left| \theta_k^{(r)} \right|}, \text{ for } r = health,\ ail. \tag{A.14}$$

Thus, by construction, $\tilde{\theta} \in [-1, 1] \times [-1, 1]$.

TABLE A.15
TREATMENT CHARACTERISTICS

	Ailments, $\theta^{(ail)}$		*Health*, $\theta^{(health)}$			*Ailments*, $\theta^{(ail)}$		*Health*, $\theta^{(health)}$	
Treatment	est.	se	est.	se	Treatment	est.	se	est.	se
AZT	−0.500	(0.020)	−12.004	(0.796)	DDI/D4T/NVP	0.753	(0.190)	44.240	(3.699)
IFNs (α/β)/AZT	−0.600	(0.068)	−55.796	(2.969)	DDI/3TC/NFV	−0.810	(0.096)	47.816	(7.454)
AL-721	−0.433	(0.096)	−19.655	(3.592)	DDI/D4T/EFV	−0.626	(0.077)	41.280	(2.806)
AZT/ACV	−0.539	(0.047)	−12.752	(1.882)	3TC/ABC/EFV	0.108	(0.047)	53.341	(1.671)
ACV	−0.783	(0.047)	−0.017	(2.931)	AZT/NVP/3TC/ABC	0.038	(0.138)	39.379	(4.528)
AZT/ACV/DDI	−0.851	(0.039)	−16.474	(1.508)	AZT/3TC/ABC/EFV	0.348	(0.089)	78.914	(3.862)
ACV/DDI	−0.348	(0.049)	−4.159	(2.286)	AZT/3TC/EFV	0.342	(0.081)	43.526	(3.395)
AZT/DDC	−0.439	(0.032)	−5.155	(1.314)	AZT/3TC/ABC	−0.442	(0.073)	54.824	(3.098)
AZT/DDI	−0.571	(0.062)	−16.615	(2.363)	D4T/3TC/EFV	−0.346	(0.065)	47.978	(3.917)
DDI	−0.375	(0.062)	15.263	(2.708)	NVP/3TC/ABC	−0.470	(0.114)	17.866	(13.435)
AZT/DDC/ACV/DDI	−0.789	(0.108)	−13.351	(8.310)	D4T/3TC/LPV/RTV	−0.310	(0.127)	35.611	(5.194)
AZT/DDC/ACV	−0.514	(0.093)	−13.186	(2.277)	3TC/LPV/RTV/ABC	−0.934	(0.131)	51.570	(5.068)
AZT/DDC/DDI	−1.440	(0.052)	−32.700	(1.771)	AZT/3TC/LPV/RTV	−0.655	(0.139)	49.838	(3.906)
DDC/ACV	−0.310	(0.098)	2.415	(4.441)	AZT/3TC/LPV/RTV/ABC	0.298	(0.267)	9.855	(10.111)
DDC	−0.358	(0.097)	−18.630	(3.836)	3TC/ABC/EFV/TDF	−0.308	(0.084)	31.845	(3.820)
D4T	−0.717	(0.054)	39.776	(2.244)	AZT/3TC/ABC/TDF	−0.652	(0.071)	19.273	(5.574)
AZT/ACV/3TC	−0.527	(0.089)	42.267	(3.433)	AZT/3TC/LPV/RTV/TDF	−0.552	(0.084)	32.227	(2.904)
AZT/3TC	0.064	(0.054)	34.398	(1.878)	NVP/3TC/TDF	−0.258	(0.157)	27.246	(4.657)
ACV/D4T/3TC	−0.509	(0.092)	33.792	(4.283)	3TC/LPV/RTV/TDF	−0.092	(0.087)	51.672	(2.603)
AZT/3TC/SQV	−0.271	(0.058)	38.283	(1.893)	LPV/RTV/EFV/TDF	−0.966	(0.100)	47.617	(2.690)
D4T/3TC	−0.104	(0.115)	37.173	(4.302)	3TC/EFV/TDF	−0.011	(0.120)	47.790	(6.753)

(*Continued*)

TABLE A.15
CONTINUED

Treatment	*Ailments*, $\theta^{(ail)}$ est.	se	*Health*, $\theta^{(health)}$ est.	se	Treatment	*Ailments*, $\theta^{(ail)}$ est.	se	*Health*, $\theta^{(health)}$ est.	se
AZT/3TC/SQV/RTV	−0.591	(0.083)	57.776	(10.098)	AZT/3TC/LPV/RTV/ABC/TDF	−0.738	(0.125)	19.980	(4.010)
AZT/ACV/3TC/IDV	−0.479	(0.051)	63.734	(2.169)	DDI/LPV/RTV/TDF	−0.276	(0.123)	18.396	(4.109)
ACV/D4T/3TC/IDV	−0.295	(0.094)	78.559	(4.143)	DDI/EFV/TDF	−0.420	(0.117)	2.381	(2.548)
AZT/3TC/RTV/IDV	−0.567	(0.095)	35.032	(6.514)	ABC/EFV/TDF	−0.762	(0.150)	39.457	(3.520)
D4T/3TC/RTV/IDV	−0.767	(0.053)	33.510	(3.204)	LPV/RTV/ABC/TDF	−0.820	(0.174)	14.891	(2.591)
D4T/3TC/SQV/RTV	−0.444	(0.086)	42.631	(5.286)	3TC/RTV/ABC/ATV	−0.061	(0.036)	26.850	(1.219)
DDI/D4T/IDV	−0.048	(0.131)	32.286	(3.915)	EFV/TDF/FTC	0.118	(0.082)	54.798	(2.735)
D4T/3TC/IDV	−0.395	(0.091)	53.128	(4.197)	RTV/EFV/TDF/FTC/ATV	0.306	(0.051)	83.823	(1.654)
AZT/3TC/IDV	−0.075	(0.074)	65.041	(2.645)	3TC/RTV/ABC/TDF/ATV	−0.403	(0.143)	38.313	(9.330)
D4T/NVP/3TC	−0.386	(0.058)	46.846	(2.896)	DDI/RTV/TDF/ATV	0.049	(0.108)	47.800	(2.887)
AZT/NVP/3TC	0.109	(0.086)	46.275	(3.665)	RTV/TDF/FTC/ATV	0.138	(0.113)	53.028	(4.026)
AZT/3TC/NFV	−0.432	(0.073)	50.776	(4.319)	NVP/TDF/FTC	−0.205	(0.079)	37.227	(2.471)
DDI/D4T/NFV	−1.049	(0.058)	57.227	(3.512)	LPV/RTV/TDF/FTC	−0.183	(0.093)	46.723	(5.288)
D4T/3TC/NFV	−0.881	(0.142)	48.018	(9.771)	RTV/TDF/FTC/FPV	−0.372	(0.110)	30.226	(3.582)
Fringe Mixes									
ISO/RTCA/IFNs (α/β)	−1.017	(0.099)	−21.950	(6.340)	NVP/3TC/RTV/LPV/RTV/TDF	−1.265	(0.113)	45.683	(4.971)
IFNs (α/β)/3TC/SQV/IDV/EFV	−0.054	(0.238)	65.353	(5.181)	3TC/RTV/LPV/RTV/ABC/TDF/ATV	−0.465	(0.080)	28.440	(2.455)
NVP/3TC/SQV/RTV/IDV	0.068	(0.144)	6.457	(6.403)	RTV/TDF/FTC/ATV/FPV	−0.612	(0.137)	42.050	(3.689)
NVP/3TC/SQV/RTV/NFV	−0.689	(0.161)	30.293	(7.861)	SQV/RTV/TDF/FTC/ATV	−0.665	(0.124)	31.824	(4.243)
NVP/SQV/RTV/ABC/EFV	−1.121	(0.170)	19.278	(4.435)	3TC/RTV/ABC/TDF/ATV/FPV	−0.210	(0.083)	26.678	(5.011)
NVP/RTV/NFV/ABC/EFV	−0.697	(0.120)	31.044	(4.225)	SQV/RTV/ABC/TDF/FTC	0.072	(0.149)	32.865	(5.005)
NVP/RTV/LPV/RTV/ABC/EFV	−0.410	(0.167)	43.495	(5.557)	3TC/RTV/TDF/FTC/RAL	0.032	(0.107)	33.352	(2.832)
NVP/3TC/NFV/ABC/TDF	−0.467	(0.102)	27.893	(3.182)	RTV/TDF/FTC/DRV/RAL	−0.221	(0.071)	47.736	(3.348)

NOTES: Treatment characteristics are estimated as indicators for treatment usage in (A.12) and (A.13). In the table, "est." stands for estimate; "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. For *fringe mixes*, we only include the most used components in the mix up to six components.

TABLE A.16
DISTRIBUTION OF NUMBER OF NEW TREATMENTS, F_N

$\ln \mu$ Variable	coef.	est.	se	$\ln \alpha$ Variable	coef.	est.	se
κ_{t-1}	ϕ_1^N	0.432	(0.072)	*Constant*	ϕ_3^N	−0.206	(0.046)
s_{et-1}	ϕ_2^N	6.177	(0.492)	κ_{t-1}	ϕ_4^N	−1.019	(0.134)

NOTES: Model is specified in (A.2). In the table, "coef." stands for coefficient and "est." stands for estimate; "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. κ_{t-1} measures the magnitude of previous innovations. $E[N_t] = \mu_{t-1} \equiv \exp(\phi_1^N \kappa_{t-1} + \phi_2^N s_{et-1})$ and $\text{Var}[N_t] = \mu_{t-1}(1 + \alpha_{t-1}^N \mu_{t-1})$.

TABLE A.17
WITHIN CLUSTER SHARE FUNCTION

Variable	$\gamma^{(weight)}$	se
AilmentsRank	−0.427	(0.122)
AilmentsRank × *HealthRank*	0.074	(0.018)
$HealthRank^2$	−0.029	(0.007)
$AilmentsRank^2$	−0.019	(0.007)
ClusterSize	−0.509	(0.046)
HealthRank × *ClusterSize*	0.046	(0.009)
AilmentsRank × *ClusterSize*	0.063	(0.009)
AilmentsRank × *HealthRank* × *ClusterSize*	−0.007	(0.001)
New	−0.352	(0.751)
New × *ClusterSize*	0.027	(0.678)
Constant	0.786	(0.114)

NOTES: Parameters estimates from (A.9) and (A.10). In the table, "est." stands for estimate; "se" stands for standard errors, in parenthesis, computed using subsampling with 500 subsamples. *AilmentsRank* stands for the rank of the ailments characteristic as compared to other treatments within the cluster; *HealthRank* is defined similarly. *ClusterSize* is the number of treatments in the cluster. *New* indicates whether the treatment just entered the market.

3. Select the first J centroids using the scaled characteristics $\tilde{\theta}$ of J randomly selected treatments from $\mathbf{A}_t$.
4. Allocate all remaining treatments $k \in \mathbf{A}_t$ to clusters sequentially. At each step, select for allocation the treatment whose scaled characteristics $\tilde{\theta}_k$ are closest to one of the existing clusters. Assign treatment k to the closest cluster and update the centroid of the cluster. Repeat this process until all treatments in $\mathbf{A}_t$ are assigned to a cluster.
5. Taken the centroids as given, reallocate all treatments to their closest centroid.
6. Calculate the value of the clustering rule $c(\mathbf{K}_t)$ in (A.8) for the current allocation.
7. Repeat 200 times steps 3–6 using the scaled characteristics $\tilde{\theta}$ of different groups of J randomly selected treatments in $\mathbf{A}_t$ as initial centroids. The allocation with the lowest value of $c(\mathbf{K}_t)$ is chosen.[36]

In the empirical application, we set the number of clusters at $J = 3$. Table A.9 presents the distribution of users (in our sample) per cluster, for various values of J.

A.3.3 *Innovation.* According to (17), the characteristics of new and experimental treatments are displaced innovations about the centroid (current or previous), and depend on the previous share of the experimental treatment and a draw from the distribution of innovation shocks $f_\nu(\nu)$. To estimate (17) and $f_\nu(\nu)$, we use all periods in the MACS data with relevant information on treatment consumed, health, and ailments, from 1986 to 2008. We observe 86 realized innovations from newly introduced commercial treatments and 22 periods where individuals used experimental treatments. However, consistent with our definition of commercial treatments, we only estimate the characteristics of experimental treatments that have at

[36] When simulating clusters for a given parameter vector in estimation, we only repeat the process 50 times.

least 40 users. We do not impose that innovation vectors cannot be strictly negative; relative to the centroid, inferior treatments with lower quality in both dimensions (health and ailments) may be introduced.[37]

A.3.4 *Utility parameters.* We estimate the utility parameters in (28) using a GMM estimator and moment conditions that equate the log odds ratio of current CCPs with a representation of the differences in conditional value functions in terms of utility parameters and future CCPs, states, and choices (Hotz et al., 1994). Below we explain this part of the estimation process in more detail.

Moment conditions. Our moment conditions appeal to well-known results following from our assumption that the taste shocks ε_{jit} are i.i.d. extreme value type I distributed (Hotz and Miller, 1993). They rely on differences between the log odds ratio and an alternative representation of differences in conditional value functions $(v_j(\bar{z}_{it}) - v_0(\bar{z}_{it}))$ in terms of future CCPs, choices, states, and utility parameters. Recalling the definition of $V(\bar{z}_{it})$ in (A.11), the conditional value function of choosing alternative j at period t is:

$$v_j(\bar{z}_{it}) = E\left\{u_j(h_{it}, y_{it}) + \beta V(\bar{z}_{it+1}) \middle| \bar{z}_{it}, d_{jit} = 1\right\}. \tag{A.15}$$

Let $p_{jit}(\bar{z}_{it})$ be the probability that individual i chooses option j at time t conditional on his state $\bar{z}_{it}$. Let $\psi_{jit}(\bar{z}_{it})$ be the expected value of the jth taste shock conditional on alternative j being optimal, and let γ be the Euler constant. Since the joint distribution of ε_t is extreme value type I:

$$\psi_j(\bar{z}_{it}) \equiv E_\varepsilon\left[\varepsilon_{jit} | \bar{z}_{it}, d^e_{jit} = 1\right] = \gamma - \ln(p_{jit}(\bar{z}_{it})). \tag{A.16}$$

Define $E_j\{\cdot\}$ as the expectation conditional on $d_{jit} = 1$. Dropping the individual subindex i for simplicity, using (A.16), we can write the conditional value function in (A.15) in terms of future utility flows induced by all available alternatives, weighted by the future probabilities of those alternatives being chosen, and corrected by the fact that the alternative may not be optimal. Notably, the weighted average of corrected flow payoffs of a given period must be discounted by the probability of survival up to that period conditional on today's state and choice. Letting T^* be an arbitrary period with $t < T^* \leq T$, the alternative representation of the conditional value function is given by:

$$\begin{aligned}
v_{jt}(\bar{z}_t) &= E_j\left\{u_j(h_t, y_t) | \bar{z}_t\right\} + \beta E_j\left\{V(\bar{z}_{t+1}, \varepsilon_{t+1}) | \bar{z}_t\right\} \\
&= E_j\left\{u_j(h_t, y_t) | \bar{z}_t\right\} + \beta E_j\left\{ b_{t+1} E_\varepsilon\left\{ \sum_{j'=0}^{J+1+r_{t+1}} d^e_{j't+1}\left[u_{j'}(h_{t+1}, y_{t+1}) + \psi_{j'}(\bar{z}_{t+1})\right]\right\} \middle| \bar{z}_t \right\} \\
&\quad + \beta^2 E_j\left\{b_{t+2} V(\bar{z}_{t+2}, \varepsilon_{t+2}) | \bar{z}_t\right\} \\
&= E_j\left\{u_j(h_t, y_t) | \bar{z}_t\right\} + \beta E_j\left\{ b_{t+1} \sum_{j'=0}^{J+1+r_{t+1}} p_{j't+1}(\bar{z}_{t+1}), \left[u_{j'}(h_{t+1}, y_{t+1}) + \psi_{j'}(\bar{z}_{t+1})\right] \middle| \bar{z}_t \right\} \\
&\quad + \beta^2 E_j\left\{b_{t+2} V(\bar{z}_{t+2}, \varepsilon_{t+2}) | \bar{z}_t\right\} \\
&= E_j\left\{u_j(h_t, y_t) | \bar{z}_t\right\} + \beta E_j\left\{ b_{t+1} \sum_{j'=0}^{J+1+r_{t+1}} p_{j't+1}(\bar{z}_{t+1}), \left[u_{j'}(h_{t+1}, y_{t+1}) + \psi_{j'}(\bar{z}_{t+1})\right] \middle| \bar{z}_t \right\}
\end{aligned}$$

[37] This is consistent with what we observe in the data, and theoretical reasons why this may happen have been provided in the literature (Miller, 1988).

$$
\begin{aligned}
&+\beta^2 E_j\left\{b_{t+1}b_{t+2}\sum_{j'=0}^{J+1+r_{t+2}} p_{j't+2}\left(\bar{z}_{t+2}\right),\left[u_{j'}\left(h_{t+2},y_{t+2}\right)+\psi_{j'}\left(\bar{z}_{t+2}\right)\right]\middle|\,\bar{z}_t\right\}\\
&+\beta^3 E_j\left\{b_{t+1}b_{t+2}V\left(\bar{z}_{t+3},\varepsilon_{t+3}\right)|\bar{z}_t\right\}\\
&=E_j\left\{u_j\left(h_t,y_t\right)|\bar{z}_t\right\}\\
&+\sum_{\tau=1}^{T^*}\beta^\tau E_j\left\{\left(\prod_{r=1}^{\tau}\Pr\left(b_{t+r}=1|h_{t+r}\right)\right),\sum_{j'=0}^{J+1+r_{t+\tau}} p_{j't+\tau}\left(\bar{z}_{t+\tau}\right),\left[u_{j'}\left(h_{t+\tau},y_{t+\tau}\right)+\psi_{j'}\left(\bar{z}_{t+\tau}\right)\right]\middle|\,\bar{z}_t\right\}\\
&+\beta^{T^*+1}E_j\left\{\left(\prod_{r=1}^{T^*+1}\Pr\left(b_{t+r}=1|h_{t+r}\right)\right),V\left(\bar{z}_{t+T^*+1},\varepsilon_{t+T^*+1}\right)\middle|\,\bar{z}_t\right\}. \qquad \text{(A.17)}
\end{aligned}
$$

Let $w(\bar{z}_{it})$ be a vector of instruments orthogonal to the difference between the log odds ratio and the alternative representation. Hence, we can form the following moment conditions:

$$
\mathbb{E}\left\{w(\bar{z}_{it})\otimes\begin{bmatrix}\ln\left(\frac{p_{0it}(\bar{z}_{it})}{p_{1it}(\bar{z}_{it})}\right)+v_{1it}(\bar{z}_{it})-v_{0it}(\bar{z}_{it})\\ \vdots\\ \ln\left(\frac{p_{0it}(\bar{z}_{it})}{p_{J+1+r_{it},it}(\bar{z}_{it})}\right)+v_{J+1+r_{it},it}(\bar{z}_{it})-v_{0it}(\bar{z}_{it})\end{bmatrix}\right\}=0. \qquad \text{(A.18)}
$$

Conditional choice probabilities. The probability that an individual chooses one of the $J+1+r_{it}$ alternatives in his choice set depends on the individual and aggregate elements of his state. In estimation, we include ω_t, κ_t and s_{et-1} directly in the CCPs and characterize other components of the aggregate state as follows: The set of available treatments $\{\theta_k\}_{k\in\mathbf{K}_t}$ is characterized by the distribution of treatment characteristics of all clusters. We use the first two moments of these distributions in estimation. The distribution of consumer characteristics H_t is controlled for using a set of nonparametric moments denoted by $\tilde{H}_t$.[38] Let m_{jit} be the moments describing the distribution of characteristics induced by alternative j for individual i at period t, including the mean vector and the variance matrix. Effectively, m_{jit} is heterogeneous across individuals only when $j=J+2$, that is, when the individual decides to purchase the same treatment, he consumed last period. Let $m_{jit}m_{jit}$ denote a vector of interactions between the elements of m_{jit}. Let $\tilde{x}_{it}^1$ and $\tilde{x}_{it}^2$ be subsets of the individual-specific components of the state.[39] Let $\omega_t m_{jit}$ denote a vector of interactions between the centroid and the elements of m_{jit}. Similarly, let $m_{jit}\tilde{x}_{it}^2$ be a vector of interactions between the components of m_{jit} and individual-specific state components and let $\omega_t m_{jit}\tilde{x}_{it}^2$ be defined in a similar fashion. Our flexible CCPs are given by:

$$
p_{jit}=\frac{\exp\left(I_{jit}\right)}{\sum_{j'=0}^{J+1+r_{it}}\exp\left(I_{j'it}\right)}, \qquad \text{(A.19)}
$$

where

$$
I_{0it}\equiv 0, \qquad \text{(A.20)}
$$

$$
\begin{aligned}
I_{jit}\equiv{}&\gamma_J\tilde{x}_{it}^1+\beta_0 m_{jt}+\beta_1 m_{jt}m_{jt}+\beta_2\omega_t m_{jt}+\beta_3 m_{jt}\tilde{x}_{it}^2+\beta_4\omega_t m_{jt}\tilde{x}_{it}^2\\
&+\beta_5 m_{jt}\tilde{H}_t+\beta_6\kappa_t+\beta_7 s_{et-1}, \qquad j=1,\ldots,J, \qquad \text{(A.21)}
\end{aligned}
$$

[38] We specify these moments as shares of people with different sets of characteristics.

[39] $\tilde{x}_{it}^2$ includes h_{it-1}, a_{it-1}, $race_i$, y_{2it}, whereas $\tilde{x}_{it}^1$ includes a constant, a_{it-1}, $race_i$.

$$I_{J+1,it} \equiv \gamma_{J+1}\tilde{x}^1_{it} + \beta_0 m_{J+1,t} + \beta_1 m_{J+1,t} m_{J+1,t} + \beta_3 m_{J+1,t}\tilde{x}^2_{it} + \beta_5 m_{J+1,t}\tilde{H}_t + \beta_6\kappa_t + \beta_7 s_{et-1}, \tag{A.22}$$

$$I_{J+2,it} \equiv \gamma_{J+2}\tilde{x}^1_{it} + \beta_0 m_{J+2,it} + \beta_1 m_{J+2,it} m_{J+2,it} + \beta_2\omega_t m_{J+2,it} + \beta_3 m_{J+2,it}\tilde{x}^2_{it} + \beta_4\omega_t m_{J+2,it}\tilde{x}^2_{it} + \beta_5 m_{J+2,it}\tilde{H}_t + \beta_6\kappa_t + \beta_7 s_{et-1}. \tag{A.23}$$

Although the characteristics of the choice sets are nonstationary due to treatment entry and exit, by interacting our time-varying regressors $\tilde{x}^2_{it}$ with the characteristics of the choice set for individual i, m_{jit}, we are able to control for the state of the world in our CCPs.[40] This procedure yields CCPs for any simulated world as long as our observed worlds cover the space of possible worlds. Additionally, we include in the CCPs ancillary coefficients that are unrelated to the state of technology, denoted by γ in (A.21) to (A.23), which capture stationary taste differences between alternatives. Consistent with our assumptions regarding the flow utility in Equation (28), we impose $\gamma_j = \gamma_J$ for any $j = 1, \ldots, J$.

Figure A.2 displays the mean predicted CCPs using (A.19) over time against the corresponding share of the population who chose the alternative.[41]

Simulation. To form the sample analog of the moment condition in (A.18), we obtain a simulated version of the conditional value function in (A.17) truncated at T^* for every observation $\{i, t\}$ and alternative $j \in \{0, 1, \ldots, J+1+r_{it}\}$. We select $T^* = 10$ so that the product $\beta^{T^*+1}\prod_{r=1}^{T^*+1}\Pr(b_{it+r} = 1|h_{it+r})$ approaches 0, eliminating further differences in conditional value functions beyond T^*. Let S denote the number of simulated paths for each $\{j, i, t\}$ and let the superscript s indicate that a quantity is simulated. For individual i and alternative j at period t, we write the simulated counterpart of the truncated value function as

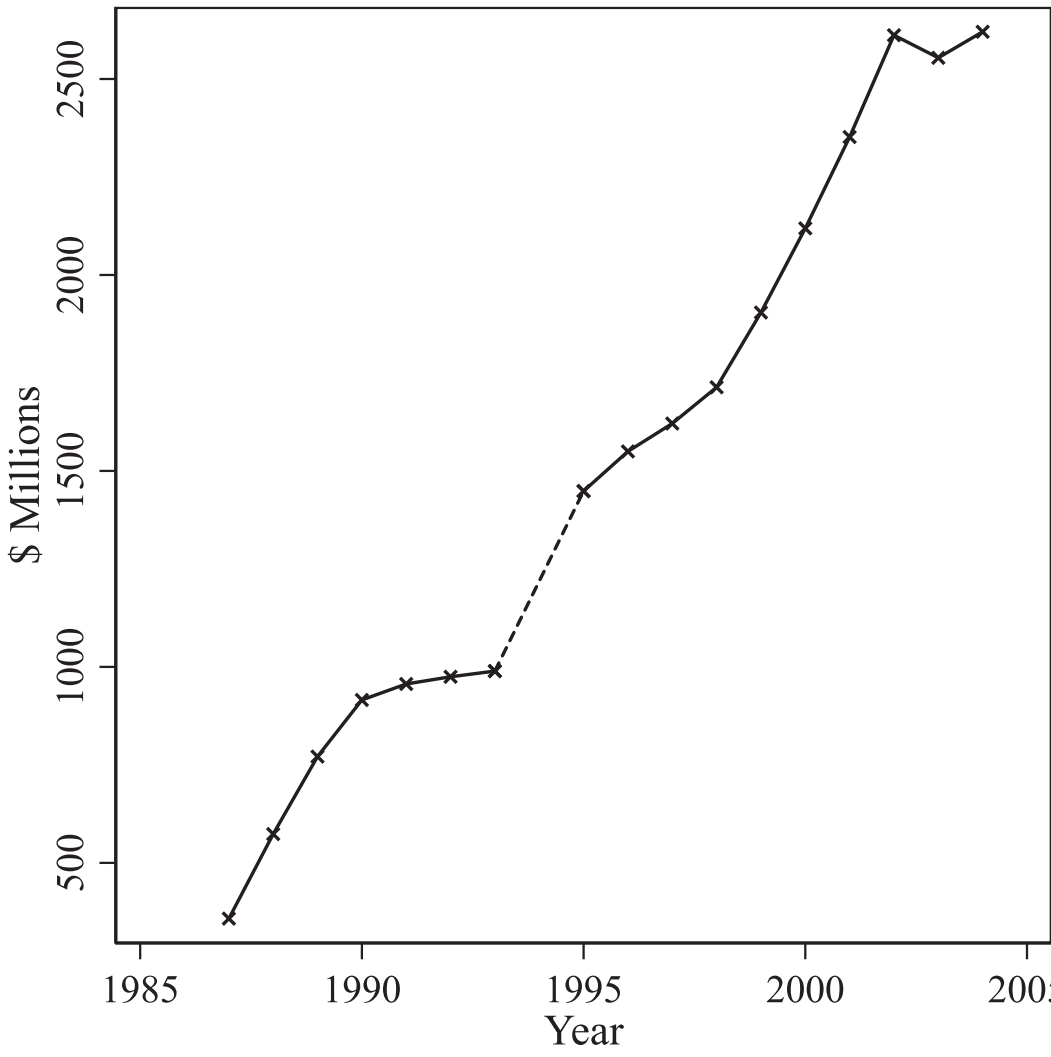
$$\bar{v}_{jit}(\bar{z}_{it}) \equiv \frac{1}{S}\sum_{s=1}^{S}\left\{u_j(h_{it}, y^s_{it}), + \sum_{\tau=1}^{T^*}\beta^\tau\left(\prod_{r=1}^{\tau}\Pr\left(b_{it+r} = 1|h^s_{it+r}\right)\right) \times \sum_{j'=0}^{J+1+r_{t+\tau}} d^s_{j'it+\tau}\left[u_{j'}\left(h^s_{it+\tau}, y^s_{it+\tau}\right) + \psi_{j'}\left(\bar{z}^s_{it+\tau}\right).\right]\right\} \tag{A.24}$$

Each future path depends on the current individual state $\bar{z}_{it}$, and hence on the current aggregate state z^a_t, and the current choice j. We first simulate as many aggregate paths starting at t as there are individuals at period t. Letting I be the number of individuals in the sample, this yields IT paths of technological innovation. Then, for each observation $\{i, t\}$ and alternative j, we generate sequences of future choices and payoffs taking as given S artificial technological paths chosen at random from the set of I simulated technological paths that start at date t.[42] This simulation process maintains the assumption, needed for consistency of the estimator, that the sample draws from the moment conditions—the contribution from each observation—are independent from each other, and it prevents simulation errors in technology paths from propagating across all observations.

[40] Because some of the components of m_{J+1t} are linear functions of ω_{t-1} (see (17)), we avoid collinearity by not including terms $\omega_t m_{J+1,t}$ and $\omega_t m_{J+1,t}\tilde{x}^2_{it}$ in (A.22).

[41] We also explore the fit of our CCP estimates by comparing the relative shares that clusters received in reality against the predictions from our estimated CCPs. We ranked the three clusters at every period by the share they received and compare this ranking against the ranking obtained from our estimated CCPs. Predicted ranks match observed ranks in about 80% of the periods.

[42] Notice that we could rely on Hotz et al. (1994) and set $S = 1$ and obtain consistency of our estimator. However, we choose $S = 20$ after trying different values for robustness.



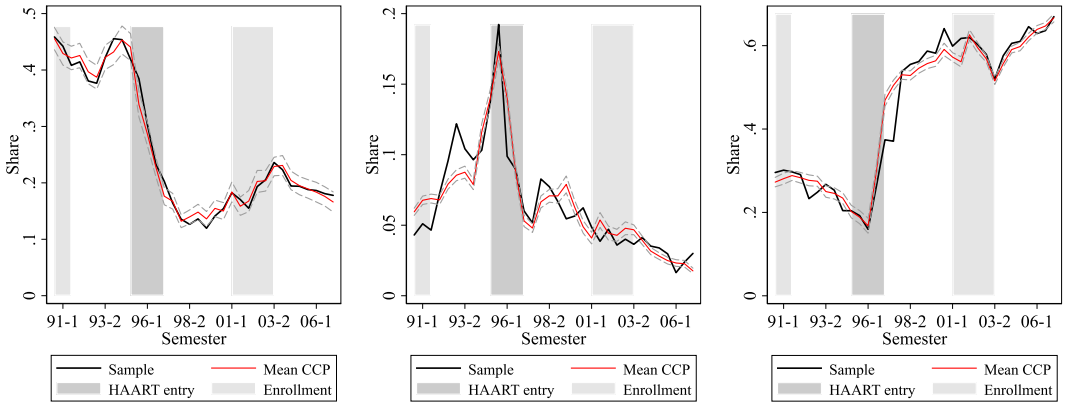
NOTES: Data from Institute of Medicine (1991), Gonsalves and Harrington (1992), and Summers and Kates (2004). Amounts in real $U.S. indexed to 2000.

FIGURE A.1

NIH HIV RESEARCH BUDGET

Simulation of the aggregate state. Taking as given the current aggregate state z_t^a, we create as many simulated aggregate state paths $\{z_{t+\tau}^{a,s}\}_{\tau=1}^{T^*}$ as there are individuals at every t. In other words, we repeat the algorithm below to generate I simulated aggregate paths for every period t:

1. Let $\tau = 1$.
2. *Entry and exit of treatments*. Simulate a number of new treatments at $t + \tau$, $N_{t+\tau}^s$, using the entry process in (A.2). If $N_{t+\tau}^s > 0$, for each simulated new treatment, draw simulated characteristics using (17). Simulate the characteristics of the experimental treatment using (17). Obtain $\kappa_{t+\tau}^s$ using (19) and (A.3). For all incumbent treatments, apply the exit rule $\{\underline{s}, \bar{s}\}$ taking into account the extent to which it binds according to (A.7). From the simulated set of treatments in $\mathbf{K}_{t+\tau}^s$ that have not yet satisfied the $\underline{s}$ exit rule, form clusters following the clustering rule in (A.8). Obtain the distribution of



(a) No Treatment, $j = 0$ (b) Experimental, $j = J{+}1$ (c) Repeat, $j = J{+}2$

NOTES: Average estimated conditional choice probability against the share of people choosing the alternative. Dashed lines represent 95% confidence intervals around the mean predicted CCPs. Three periods of special relevance are highlighted in the figure: two periods during which enrollment into the sample was undertaken and the period in which treatments belonging to the HAART class were introduced.

FIGURE A.2

AVERAGE CCPS
[COLOR FIGURE CAN BE VIEWED AT WILEYONLINELIBRARY.COM]

characteristics of each cluster using (A.9) and (A.10). For $\tau > 1$, compute the simulated centroid $\omega^s_{t+\tau}$ using (16).
Demand. For all individuals i' at t: If $\tau = 1$, define $h^s_{i't+1} \equiv h_{i't+1}$ and $d^s_{i't} \equiv d_{i't}$; otherwise, simulate $h^s_{i't+\tau}$ using (22). Draw a simulated labor state $y^s_{2i't+\tau}$ using (25). Compute deterministic transitions (e.g., age). Using $\bar{z}^s_{i't+\tau}$, and hence $z^{a,s}_{t+\tau}$, and (A.19) to (A.23) compute simulated CCPs $p^s_{ji't+\tau}(\bar{z}^s_{i't+\tau})$ for every alternative $j \in \{0, 1, \ldots, J+1+r^s_{it+\tau}\}$ and draw a decision $d^s_{i't+\tau}$. Obtain the simulated share of trial participation $s^s_{e,t+\tau}$ and the nonparametric representation of the simulated distribution of consumer characteristics $\tilde{H}^s_{t+\tau}$.

3. *Cycle back*. If $\tau = T^*$, end the loop. Otherwise, let $\tau = \tau + 1$ and go back to step 2.

Simulation of individual paths. For every observation $\{i, t\}$ and every alternative $j \in \{0, 1, \ldots, J+1+r_{it}\}$, we generate S sequences of future states, choices, and outcomes $\{\bar{z}^s_{it+\tau}, d^s_{it+\tau}, y^s_{it+\tau}\}_{\tau=1}^{T^*}$ taking as given a subset of S simulated aggregate paths—that start at t—chosen at random without replacement. We follow the steps below:

1. Let $\tau = 1$.
2. *Demand*. Same as above but only for individual i. When j is not equal to the observed choice for $\{i, t\}$, we also simulate health at the beginning of period $t+1$. For this, we back out the realized health residual using (A.12) and use (22) to simulate health h^s_{it+1} under counterfactual choice j. Additionally, we compute the simulated one-period-ahead survival probability $\Pr(b_{it+r} = 1 | h^s_{it+r})$.
3. *Outcomes*. Only for individual i: Simulate ailments using (23) and the relevant distribution of treatment characteristics implied by the simulated choice $d^s_{it+\tau}$. Simulate income using (26) and out-of-pocket prescription expenditures using (27).[43]
4. *Cycle back*. If $\tau = T^*$, end the loop. Otherwise, let $\tau = \tau + 1$ and go back to step 2.

[43] Even though individuals know their idiosyncratic income shocks ϵ^m_{it}, we do not need to simulate these shocks as they are i.i.d., have mean 0, and enter linearly in the flow utility, which results in them averaging out to 0 in the moment condition.

When simulating a path following an alternative j that is not the observed choice for $\{i,t\}$, we obtain current-period simulated payoffs $u_j(h_{it}^s, y_{it}^s)$ by simulating current income, out-of-pocket expenditures, and ailments conditional on the counterfactual choice j at t.

Estimator. Let $j=0$ be the base alternative, and let δ_{it} be an indicator of whether individual i is in the data at period t. The simulated sample analog of the moment condition in (A.18) is

$$\frac{1}{\sum_i \sum_t \delta_{it}} \sum_{i=1}^{I} \sum_{t=1}^{T} \delta_{it} w(\bar{z}_{it}) \otimes \begin{bmatrix} \ln\left(\frac{p_{0it}(\bar{z}_{it})}{p_{1it}(\bar{z}_{it})}\right) + \bar{v}_{1it}(\bar{z}_{it}) - \bar{v}_{0it}(\bar{z}_{it}) \\ \vdots \\ \ln\left(\frac{p_{0it}(\bar{z}_{it})}{p_{J+1+r_{it},it}(\bar{z}_{it})}\right) + \bar{v}_{J+1+r_{it},it}(\bar{z}_{it}) - \bar{v}_{0it}(\bar{z}_{it}) \end{bmatrix} = 0. \tag{A.25}$$

Denote Λ as the M-dimensional vector of parameters of the utility function. Following Hotz et al. (1994), we estimate Λ as the vector that minimizes the following objective function:

$$\left((IT)^{-1} \sum_{i=1}^{I} \sum_{t=1}^{T} \delta_{it} w(\bar{z}_{it}) \otimes A_{it}(\bar{z}_{it}, \Lambda)\right)' W_n \left((IT)^{-1} \sum_{i=1}^{I} \sum_{t=1}^{T} \delta_{it} w(\bar{z}_{it}) \otimes A_{it}(\bar{z}_{it}, \Lambda)\right), \tag{A.26}$$

where

$$A_{it}(\bar{z}_{it}, \Lambda) \equiv \begin{bmatrix} \ln\left(\frac{p_{0it}(\bar{z}_{it})}{p_{1it}(\bar{z}_{it})}\right) + \bar{v}_{1it}(\bar{z}_{it}) - \bar{v}_{0it}(\bar{z}_{it}) \\ \vdots \\ \ln\left(\frac{p_{0it}(\bar{z}_{it})}{p_{J+2it}(\bar{z}_{it})}\right) + \bar{v}_{J+2it}(\bar{z}_{it}) - \bar{v}_{0it}(\bar{z}_{it}) \end{bmatrix}. \tag{A.27}$$

and W_n is a square weighting matrix. Using the linear structure of the utility function in (28), we collect and factor terms to write the jth component of the vector $A_{it}(\bar{z}_{it}, \Lambda)$ as the linear form

$$\tilde{y}_{jit} - \tilde{x}'_{jit}\Lambda. \tag{A.28}$$

Define Y as a vector with $(J+2)IT$ rows that stacks all $\tilde{y}_{jit}$, and X as a $(J+2)IT \times M$ matrix that stacks all $\tilde{x}_{jit}$. Define Z as the $IT \times R$ matrix whose columns contain the R instruments orthogonal to the difference between the log odds ratio of current CCPs and the alternative representation of the differences in conditional value functions.[44] Thus,

$$Y = \begin{bmatrix} \tilde{y}_{1,1,1} \\ \tilde{y}_{1,1,2} \\ \vdots \\ \tilde{y}_{1,I,T-1} \\ \tilde{y}_{1,I,T} \\ \vdots \\ \tilde{y}_{J+2,1,1} \\ \tilde{y}_{J+2,1,2} \\ \vdots \\ \tilde{y}_{J+2,I,T-1} \\ \tilde{y}_{J+2,I,T} \end{bmatrix}, \quad X = \begin{bmatrix} \tilde{x}_{1,1,1,1} & \cdots & \tilde{x}_{1,1,1,M} \\ \tilde{x}_{1,1,2,1} & \cdots & \tilde{x}_{1,1,2,M} \\ \vdots & & \vdots \\ \tilde{x}_{1,I,T-1,1} & \cdots & \tilde{x}_{1,I,T-1,M} \\ \tilde{x}_{1,I,T,1} & \cdots & \tilde{x}_{1,I,T,M} \\ \vdots & & \vdots \\ \tilde{x}_{J+2,1,1,1} & \cdots & \tilde{x}_{J+2,1,1,M} \\ \tilde{x}_{J+2,1,2,1} & \cdots & \tilde{x}_{J+2,1,2,M} \\ \vdots & & \vdots \\ \tilde{x}_{J+2,I,T-1,1} & \cdots & \tilde{x}_{J+2,I,T-1,M} \\ \tilde{x}_{J+2,I,T,1} & \cdots & \tilde{x}_{J+2,I,T,M} \end{bmatrix}, \quad Z = \begin{bmatrix} w(\bar{z}_{11})_1 & \cdots & w(\bar{z}_{11})_R \\ w(\bar{z}_{12})_1 & \cdots & w(\bar{z}_{12})_R \\ \vdots & & \vdots \\ w(\bar{z}_{IT})_1 & \cdots & w(\bar{z}_{IT})_R \end{bmatrix}. \tag{A.29}$$

[44] Hence, W_n is a $(J+2)R$-dimensional square matrix.

Finally, let $\mathbf{I}_{[J+2]}$ be a $(J+2)$-dimensional identity matrix and define $\tilde{Z} \equiv \mathbf{I}_{[J+2]} \otimes Z$. Then we can write the objective function in (A.26) as

$$\Big((IT)^{-1}\tilde{Z}'(Y - X\Lambda)\Big)' W_n \Big((IT)^{-1}\tilde{Z}'(Y - X\Lambda)\Big). \tag{A.30}$$

Equation (A.30) is a linear arrangement, so we can obtain a closed-form solution for $\hat{\Lambda}$ as the optimal GMM estimator. It entails first and second stage estimators given by

$$\hat{\Lambda}^{1S} = \big(X'\tilde{Z}\tilde{Z}'X\big)^{-1}\big(X'\tilde{Z}\tilde{Z}'Y\big), \qquad \hat{\Lambda}^{2S} = \Big(X'\tilde{Z}\hat{S}^{-1}\tilde{Z}'X\Big)^{-1}\Big(X'\tilde{Z}\hat{S}^{-1}\tilde{Z}'Y\Big), \tag{A.31}$$

where

$$\hat{S} = \frac{1}{I^*}\tilde{Z}'D\tilde{Z}, \qquad I^* = IT(J+1) + \sum_{i=1}^{I}\sum_{t=1}^{T} r_{it} \tag{A.32}$$

accounts for the fact that some individuals cannot repeat their previous consumption (for instance, if the treatment was withdrawn), and D is the $I(J+2)$ square diagonal matrix with diagonal elements $\hat{u}^2_{jit} = (\tilde{y}_{jit} - \tilde{x}'_{jit}\hat{\Lambda}^{1S})^2$. As instruments, we use initial health h_{it}, lagged labor state y_{2it-1}, income fixed effect η_i, race/ethnicity, education, and age a_{it}, the centroid ω_t and the lagged share of trial participation s_{et-1}, as well as interactions between these variables. The variance–covariance matrix of the second-stage estimator is

$$\hat{V}^{2S} = I^*\Big(X'\tilde{Z}\hat{S}^{-1}\tilde{Z}'X\Big)^{-1}. \tag{A.33}$$

A.3.5 *Standard errors.* The uncorrected standard errors for our utility parameters yield from the variance-covariance matrix in (A.33). We obtain corrected standard errors using subsampling taking as given the following objects obtained from the full sample: the definition of treatments (i.e., what their components are, for instance, AZT or AZT + DDI) and the exit thresholds $\{\underline{s}, \bar{s}\}$ specified in Subsection A.8. We draw $R = 500$ subsamples containing a proportion $\tilde{p} = 0.9$ of the individuals in the sample drawn without replacement, and estimate all parameters in the model using each subsample. This includes estimating treatment characteristics, parameters governing transition and outcome processes, and simulating forward paths of technology to obtain utility parameters. For any parameter γ, the subsampling standard errors are obtained as

$$se(\hat{\gamma}) \approx se(\hat{\gamma}_r) \cdot \sqrt{\tilde{p}}, \tag{A.34}$$

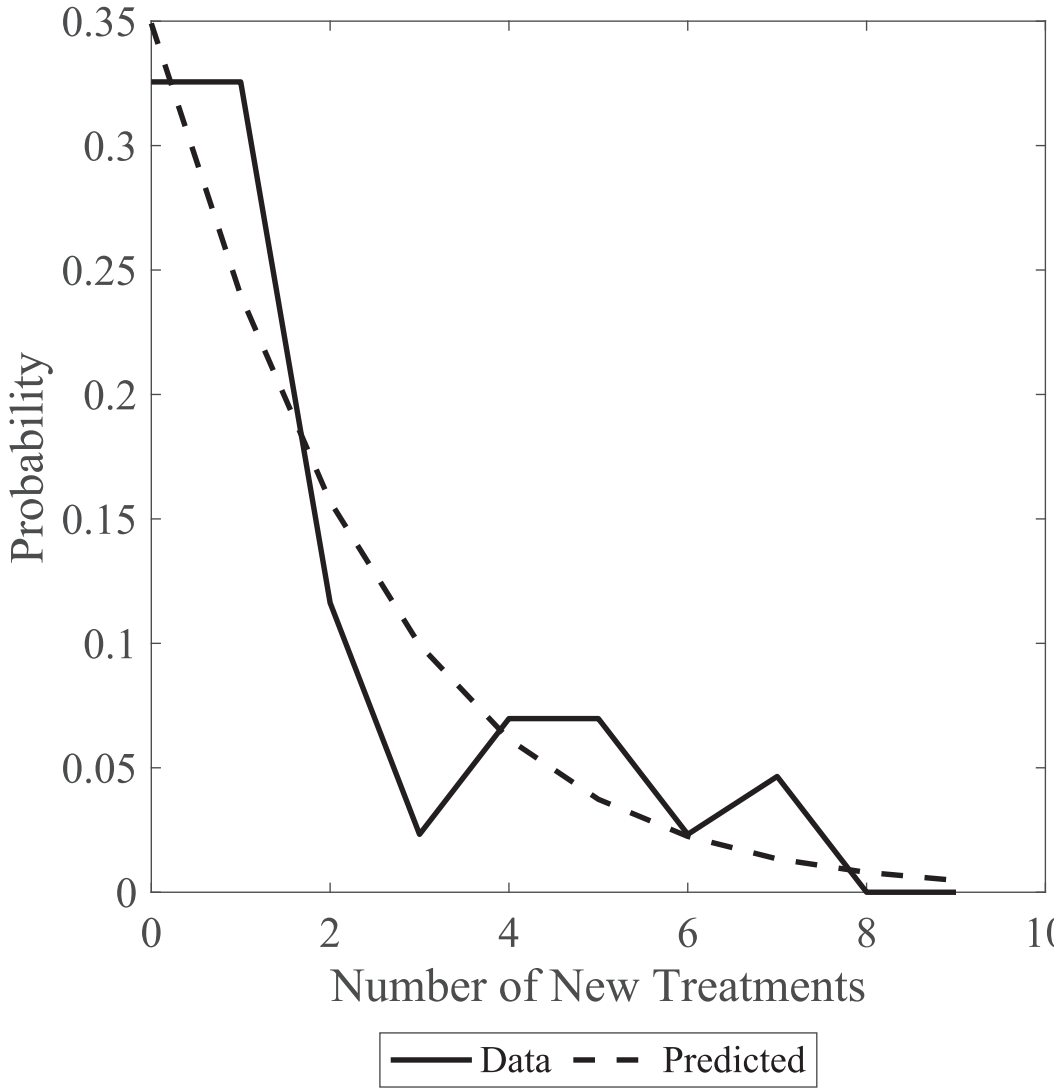
where $\hat{\gamma}_r$ is the estimated value from the rth subsample, and $se(\hat{\gamma}_r)$ is estimated as the standard deviation of the R estimates $\hat{\gamma}_r$.

A.4 *Results Appendix.*

A.4.1 *Estimates.*

A.4.2 *The likelihood of observed technological progress.*

A.4.3 *Eliminating the effect of repeat purchase.* The evolution of technology, and ultimately consumer welfare, is affected by demand externalities arising in the innovation process. We measure the importance of these externalities by describing how the market would evolve if consumers had less influence over the process of innovation, restricting the role of



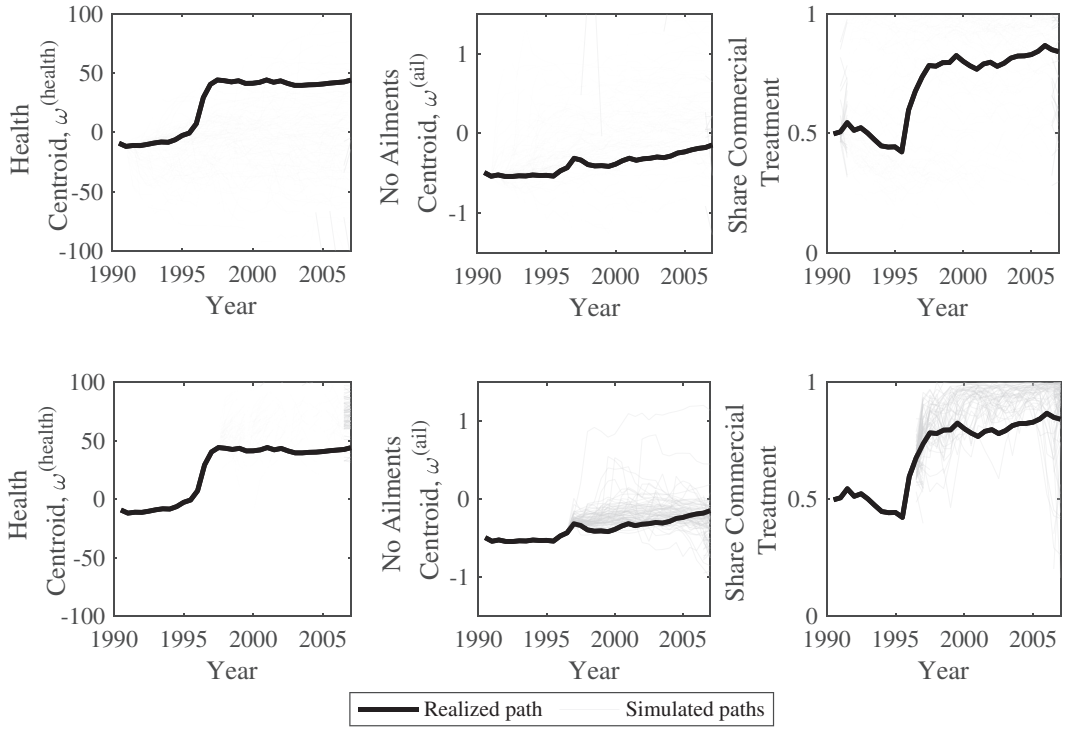
NOTES: Model is specified in (A.2). This figure shows the empirical distribution of the number of new treatments and the average of the predicted probabilities using the estimated parameters in Table A.16.

FIGURE A.3

DISTRIBUTION OF NUMBER OF NEW TREATMENTS

demand pull. In this counterfactual, we eliminate the effect of repeat purchase on innovation. Recall that in the model, individuals who repeat purchase have full information regarding treatment characteristics. We present results averaging over 500 simulated paths starting at the first semester of 1991.

Since consumers dislike changing treatment, they face a trade-off between old and new technologies, and are more likely to repeat purchase if prior treatment offers better characteristics than current clusters. In this counterfactual, we study the evolution of treatment quality when the process of innovation remains responsive to demand but demand by repeat consumers is not guided by their preferences, individual characteristics, or their knowledge of the quality of the treatment they are consuming. Concretely, we assign individuals to alternatives in the choice set in the same proportions as the baseline (including the experimental treatment and no treatment), but make repeat consumption of old technologies random. By



NOTES: One hundred simulated paths conditional on the state of the world in 1991 and 1996.

FIGURE A.4

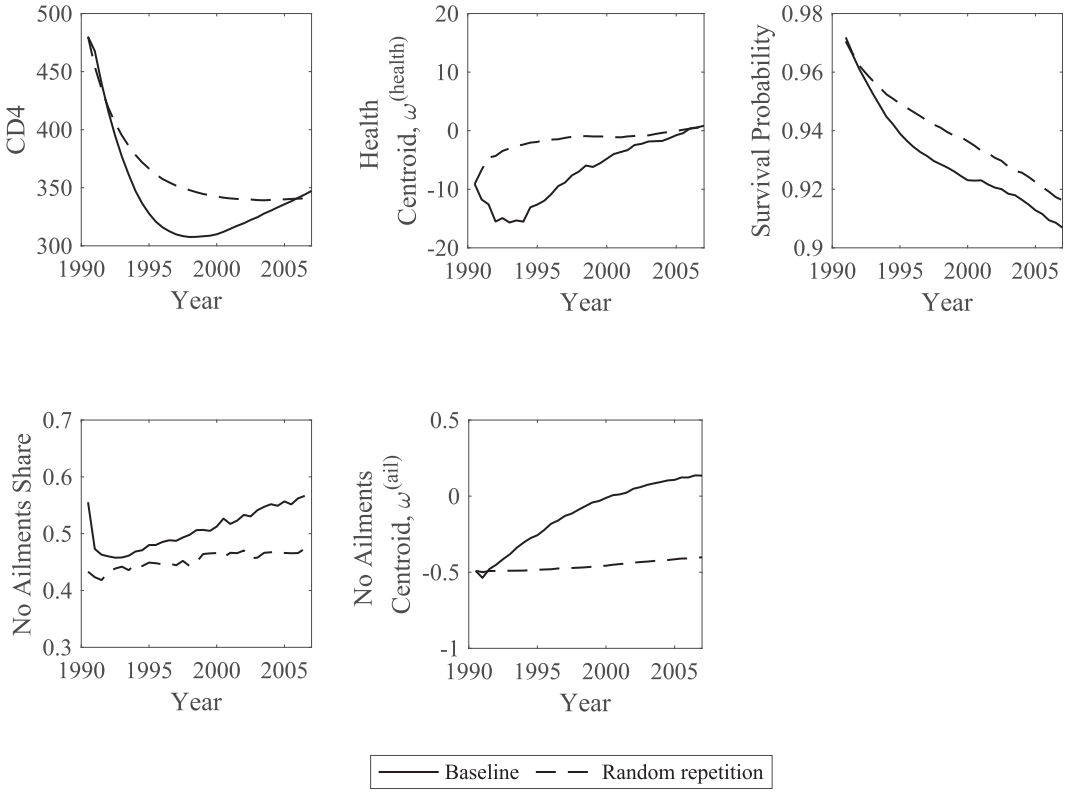
DISTRIBUTION OF TECHNOLOGY PATHS: TECHNOLOGY AND TREATMENT CONSUMPTION

matching the unconditional shares of this counterfactual regime to the unconditional shares in the baseline (the estimated model of demand-pull innovation in Section 4), we avoid spurious effects on the process of innovation yielding from arbitrary aggregate shares (e.g., $1/G$ for a choice set of size G). This regime neutralizes the dependence of the technological path on the preferences and characteristics of repeat consumers without changing the nature of the law of motion of available treatments.

Figure A.5 shows that in the counterfactual regime, the path of innovation is tilted toward more effective treatments with greater side effects. In other words, eliminating the effects of repeat consumption improves health and survival, but leads to more physical ailments. Informed repeat customers trade efficacy for fewer side effects despite the detrimental impact on their survival.

A.4.4 *Continuation values and smoothing.* We obtain continuation values for every subsidy value in Section 6 by implementing the following algorithm:

1. Create a collection, denoted by $\mathcal{V}$, of 500 continuation value vectors computed for all $t+1$ states. Each row in a value vector is an individual. Each value vector $v \in \mathcal{V}$ corresponds to a $t+1$ aggregate state $z_{t+1}^{a,v}$.
2. For each subsidy value n, we compute each individual's current payoff and their future state, as well as the implied $t+1$ aggregate state $z_{t+1}^{a,n}$.
3. We match the vector of current payoff under subsidy n to the continuation value vector $v^* \in \mathcal{V}$ corresponding to the $t+1$ aggregate state that is closest to the aggregate state induced by subsidy n. In other words, we match subsidy n to the continuation value vector



NOTES: Average paths computed over 500 simulations that are conditional on the state of the world at 1991. The *baseline* is the estimated model of demand-pull innovation in Section 4. The baseline solid lines in Figure A.5 are the averages of the gray lines in Figure 10 and Figure A.4 in Appendix A.15.2. Individuals in the alternative regime are assigned alternatives using the unconditional shares from the baseline model as assignment probabilities.

FIGURE A.5

ELIMINATING THE EFFECT OF REPEAT PURCHASE

v^* that solves:

$$v^* = \arg\min_{v \in \mathcal{V}} ||z_{t+1}^{a,n} - z_{t+1}^{a,v}||. \tag{A.35}$$

We use a measure of Euclidean distance that yields from discretizing the aggregate states $z_{t+1}^{a,n}$ and $z_{t+1}^{a,v}$ into vectors with 196 components. We scale each component of the discretized aggregate state vectors to be between 0 and 1 by dividing over its largest value.

4. We repeat 1,000 times steps 2 and 3 for every subsidy n and average over repetitions.

Our method of matching continuation values generates noise around the mapping from subsidy values into average consumer lifetime utility. (See the point-dash line in Figure 11.) We use local polynomials to smooth the mapping in an interval starting at the decentralized share s_{et}.

REFERENCES

ACEMOGLU, D., and J. LINN, "Market Size in Innovation: Theory and Evidence from the Pharmaceutical Industry," *Quarterly Journal of Economics* 119 (2004), 1049–90.

AGUIRREGABIRIA, V., and A. MAGESAN, "Solution and Estimation of Dynamic Discrete Choice Structural Models Using Euler Equations Markets," Working Paper, 2018.

Alsan, M., and M. Wanamaker, "Tuskegee and the Health of Black Men," *Quarterly Journal of Economics* 133 (2018), 407–55.
AltuAltuğ, S., and R. A. Miller, "The Effect of Work Experience on Female Wages and Labour Supply," *Review of Economic Studies* 65 (1998), 45–85.
Bhaskaran, K., O. Hamouda, M. Sannes, F. Boufassa, A. Johnson, L. P.C., K. Porter, and C. Collaboration, "Changes in the Risk of Death after HIV Seroconversion Compared with Mortality in the General Population," *JAMA* 300 (2008), 51–9.
Black, D., G. Gates, S. Sanders, and L. Taylor, "Demographics of the Gay and Lesbian Population in the United States: Evidence from Available Systematic Data Sources," *Demography* 37 (2000), 139–54.
———, "Why Do Gay Men Live in San Francisco?," *Journal of Urban Economics* 51 (2002), 54–76.
Bolton, P., and C. Harris, "Strategic Experimentation," *Econometrica* 67 (1999), 349–74.
Brown, D., J. Thorne, G. Foster, J. Duncan, L. Brune, A. Munana, C. L. Meinert, and D. Jabs, "Factors Affecting Attrition in a Longitudinal Study of Patients with AIDS," *AIDS Care* 18 (2006), 821–9.
Chan, T. Y., and B. H. Hamilton, "Learning, Private Information, and the Economic Evaluation of Randomized Experiments," *Journal of Political Economy* 114 (2006), 997–1040.
———, and N. W. Papageorge, "Health, Risky Behaviour and the Value of Medical Innovation for Infectious Disease," *Review of Economic Studies* 83 (2016), 1465–510.
Chen, J., S. Esteban, and M. Shum, "When Do Secondary Markets Harm Firms?," *American Economic Review* 103 (2013), 2911–34.
Crawford, G. S., and M. Shum, "Uncertainty and Learning in Pharmaceutical Demand," *Econometrica* 73 (2005), 1137–73.
Danzon, P. M., Y. R. Wang, and L. Wang, "The Impact of Price Regulation on the Launch Delay of New Drugs–Evidence from Twenty-Five Major Markets in the 1990s," *Health Economics* 14 (2005), 269–92.
Darden, M., "Smoking, Expectations, and Health: A Dynamic Stochastic Model of Lifetime Smoking Behavior," *Journal of Political Economy* 125 (2017), 1465–522.
Dickstein, M. J., "Efficient Provision of Experience Goods: Evidence from Antidepressant Choice," Working Paper, 2018.
Dranove, D., C. Garthwaite, and M. Hermosilla, "Pharmaceutical Profits and the Social Value of Innovation," NBER Working Paper 20212, 2014.
Dubois, P., O. de Mouzon, F. Scott-Morton, and P. Seabright, "Market Size and Pharmaceutical Innovation," *RAND Journal of Economics* 46 (2015), 844–71.
Duda, R. O., and P. E. Hart, *Pattern Classification and Scene Analysis* (New York: Wiley, 1973).
Fernandez, J. M., "An Empirical Model of Learning under Ambiguity: The Case of Clinical Trials," *International Economic Review* 54 (2013), 549–73.
Finkelstein, A., "Static and Dynamic Effects of Health Policy: Evidence from the Vaccine Industry," *Quarterly Journal of Economics* 119 (2004), 527–64.
Gable, C., J. Tierce, D. Simison, D. Ward, and K. Motte, "Costs of HIV+/AIDS at CD4+ Counts Disease Stages Based on Treatment Protocols," *Journal of Acquired Immune Deficiency Syndromes and Human Retrovirology* 12 (1996), 413–20.
Goettler, R. L., and B. Gordon, "Does AMD Spur Intel to innovate More?," *Journal of Political Economy* 119 (2011), 1141–200.
Goldman, D., and D. Lakdawalla, "The Global Burden of Medical Innovation," Technical Report, USC Leonard D. Schaeffer Center for Health Policy & Economics, 2018.
Gonsalves, G., and M. Harrington, "AIDS Research at the NIH: A Critical Review," Technical Report, VIII International Conference on AIDS, Amsterdam, 1992.
Gowrisankaran, G., and M. Rysman, "Dynamics of Consumer Demand for New Durable Goods," *Journal of Political Economy* 120 (2012), 1173–219.
Grossman, S. J., and J. E. Stiglitz, "On the Impossibility of Informationally Efficient Markets," *American Economic Review* 70 (1980), 393–408.
Harris, Y., P. B. Gorelick, P. Samuels, and I. Bempong, "Why African Americans May Not Be Participating in Clinical Trials," *Journal of the National Medical Association* 88 (1996), 630–4.
Hicks, J. R., *The Theory of Wages* (London: Macmillan, 1932).
Holmberg, S. D., "The Estimated Prevalence and Incidence of HIV in 96 Large US Metropolitan Areas," *American Journal of Public Health* 86 (1996), 642–54.
Hotz, V. J., and R. A. Miller, "Conditional Choice Probabilities and the Estimation of Dynamic Models," *Review of Economic Studies* 60 (1993), 497–529.
———, S. Sanders, and J. Smith, "A Simulation Estimator for Dynamic Models of Discrete Choice," *Review of Economic Studies* 61 (1994), 265–89.
Igami, M., "Estimating the Innovator's Dilemma: Structural Analysis of Creative Destruction in the Hard Disk Drive Industry, 1981-1998," *Journal of Political Economy* 125 (2017), 798–847.

Institute of Medicine, "The AIDS Research Program of the National Institutes of Health," Technical Report, National Academy Press, Washington, DC, 1991.

Jovanovic, B., and G. M. MacDonald, "Competitive Diffusion," *Journal of Political Economy* 102 (1994), 24–52.

Kyle, M. K., "Pharmaceutical Price Controls and Entry Strategies," *Review of Economics and Statistics* 89 (2007), 88–99.

Malani, A., and T. J. Philipson, "Can Medical Progress Be Sustained? Implications of the Link between Development and Output Markets," NBER Working Paper 17011, 2011.

Miller, R. A., "Innovation and Reputation," *Journal of Political Economy* 96 (1988), 741–65.

Mills, E., K. Wilson, B. Rachlis, L. Griffith, P. Wu, G. Guyatt, and C. Cooper, "Barriers to Participation in HIV Drug Trials: A Systematic Review," *The Lancet Infectious Diseases* 6 (2006), 32–8.

OECD, "Pharmaceutical Expenditure," in *Health at a Glance 2017: OECD Indicators* (Paris: OECD Publishing, 2017), 186–7.

Papageorge, N. W., "Why Medical Innovation is Valuable: Health, Human Capital, and the Labor Market," *Quantitative Economics* 7 (2016), 671–725.

Petrin, A., "Quantifying the Benefits of New Products: The Case of the Minivan," *Journal of Political Economy* 110 (2002), 705–29.

Radner, R., "Rational Expectations Equilibrium: Generic Existence and the Information Revealed by Prices," *Econometrica* 47 (1979), 655–78.

Santolaya, Perrín, R., and F. J. García López, "Incremental Drug Treatment Cost in HIV-Positive Patients in Industry-Sponsored Clinical Trials," *The Annals of Pharmacotherapy* 42 (2008), 1586–91.

Sarnak, D. O., D. Squires, and S. Bishop, "Paying for Prescription Drugs Around the World: Why Is the U.S. an Outlier?" Technical Report, The Commonwealth Fund, Washington, DC, 2017.

Scherer, F. M., "Demand-Pull and Technological Invention: Schmookler Revisted," *Journal of Industrial Economics* 30 (1982), 225–37.

Schmookler, J., *Invention and Economic Growth* (Cambridge: Harvard University Press, 1966).

Summers, T., and J. Kates, "Trends in U.S. Government Funding for HIV/AIDS, Fiscal Years 1981 to 2004," Technical Report, Kaiser Family Foundation, 2004.

Waldfogel, J., "Preference Externalities: An Empirical Study of Who Benefits Whom in Differentiated-Product Markets," *RAND Journal of Economics* 34 (2003), 557–68.