Among the most striking features of the global spread of HIV is its heterogeneity across geographically and socially defined populations. According to recent estimates of the Joint United Nations Programme on HIV/AIDS (UNAIDS 2008), HIV prevalence among adults varies at the regional level by a factor of 50, from 0.1 percent in East Asia to 5.0 percent in sub-Saharan Africa. Considerable variation can be found also within regions. Adult HIV prevalence exceeds 20 percent in the Southern African countries of Botswana, Lesotho, and Swaziland, for example, but remains below 5 percent in most West African countries (UNAIDS 2008). These geographical variations cannot be explained by chronological differences in when HIV was first introduced into regional or national populations (Caraël and Holmes 2001).

The spread of HIV has been characterized also by sharp variations between socially defined groups living in close proximity. In the United States, HIV spread aggressively among injection-drug users and men who have sex with men in several cities between 1978 and 1983, with prevalence surpassing 80 percent in some cohorts (Jaffe et al. 1985; Hessol et al. 1989; Des Jarlais et al. 1991; Hoover et al. 1991; van Griensven et al. 1993), yet it scarcely gained a foothold among other residents of the same cities during that period. More recently, HIV infections and AIDS diagnoses in the United States have become concentrated among racial and ethnic minority groups (Smith et al. 2000; Karon et al. 2001; McQuillan et al. 2006; Morris et al. 2006). According to surveillance estimates (CDC 2006), the rate of AIDS diagnosis among African Americans now exceeds ten times that among non-Hispanic whites. Meanwhile, a survey conducted in several American cities during the late 1990s revealed an HIV-prevalence level of 29 percent among African American men who have sex with men (Catania et al. 2001).

Understanding the processes that generate these wide variations is of great interest to scientists and policymakers. In this study, we consider the changing role of behavioral mechanisms in epidemiological explanations for variations in the spread of HIV across geographically and socially defined populations and in strategies for controlling HIV epidemics. The discussion below is organized in part as a historical narrative.
In the 1980s, behavioral variations within and between populations were the central focus of AIDS research, and behavior change was seen as the primary if not the sole means of controlling HIV epidemics. Today, biological mechanisms—especially other sexually transmitted infections, antiretroviral therapy, and male circumcision—predominate in HIV epidemiology and prevention efforts. We consider reasons for this shifting emphasis, but argue for a sustained focus on behavioral mechanisms in HIV research. We also advocate for a revised formulation of HIV epidemiology’s central conceptual model to recognize explicitly the dynamism of human behavior and the interdependence between behavioral and biological determinants of risk.

HIV Epidemiology’s Conceptual Framework

For almost three decades, the conceptual framework that has guided epidemiological thinking and research about the spread of HIV has contributed to a division in focus between behavior and biology. It has posited three categories of proximate determinants: (1) the behavioral mechanisms through which exposures occur, (2) the biological mechanisms that govern the probability that an exposure results in an infection, and (3) the factors (demographic and biomedical) that shape the duration of infectiousness. The collective determinants in these three categories are sometimes denoted as $c$, $\beta$, and $D$, respectively. The notion that they operate in concert to determine the spread of HIV is embodied in mathematical models and simplified in the expression $R_o = c\beta D$, where $c$ is defined as the number of new sexual partnerships per person per unit of time, $\beta$ is the probability that HIV is transmitted during sex between a susceptible and an infected partner, and $D$ is the length of time a newly infected person is expected to remain alive, sexually active, and infectious. If infected people die at rates $k$ and $\mu$ from HIV infection and all other causes, respectively, but remain sexually active and infectious until they die, then $R_o = c\beta / (\mu + k)$. Specified in this way, $R_o$ can be interpreted as the expected number of secondary infections caused by introducing one infected person into a large susceptible population, and HIV prevalence is predicted to converge over time toward an equilibrium level of $P^* = 1 - 1 / R_o$.

In this conceptual framework, however, the function of the equation $R_o = c\beta D$ is heuristic rather than predictive. The parameters $c$, $\beta$, and $D$ refer loosely to collections of influences, and $R_o$ refers in a broad way to the severity of the HIV epidemic in a given population. The majority of HIV epidemiologists never work directly with mathematical models, but instead rely upon the formula $R_o = c\beta D$ mainly as a form of shorthand for the largely intuitive conceptual framework.

The conceptual framework provides the central organizing principle for HIV epidemiology and guides research and prevention efforts alike (Brunham 2005). By and large it encourages HIV epidemiologists to work backward from outcomes at the individual and population levels. It also promotes a compartmentalization of effort between “behavior” and “biology” camps. Some investigators focus on identifying routes of transmission and characterizing patterns in the corresponding behavioral mechanisms, represented by $c$. Others turn their attention to the search for biological mechanisms that influence susceptibility and infectiousness, summarized by $\beta$. In theory, the duration of infectiousness could also play a role, but mechanisms in this category have received little attention from HIV epidemiologists (except insofar as mechanisms such as antiretroviral therapy that influence $\beta$ are thought also to have countervailing influences on $D$).

HIV epidemiologists take for granted that variations in these proximate mechanisms may, in turn, be attributable to social factors. According to this perspective, “upstream” variables may contribute to the spread of HIV, but they do so only indirectly, through their effects on the proximate determinants (Aral and Holmes 1999; Boerma and Weir 2005; Hayes and Weiss 2006). That $c$, $\beta$, and $D$ (or their more complex generalizations) are seldom written explicitly as functions of social variables therefore reflects uncertainty about which social variables influence which proximate mechanisms in which populations, as well as epidemiologists’ greater comfort in dealing with biological than with social factors. This uncertainty reinforces the strategy of working backward. The assumption is implicit that research on more remote influences will be fruitful only if it is adequately informed by a detailed understanding of the intervening proximate mechanisms, be they behavioral, biological, or demographic.

Early Prominence of Behavioral Mechanisms

Epidemiological research on the condition now known as AIDS began in 1981 when the first cases were reported to the Centers for Disease Control (CDC). Because the retrovirus now known as HIV had not yet been identified, epidemiological investigations were directed at narrowing the search for the cause(s) of the unusual cancers and infectious diseases appearing among gay and bisexual men and injection-drug users in several American cities. Clinical case studies provided valuable insights, particu-
larly regarding the underlying commonality of cellular immune dysfunction. Equally important, however, were individual- and group-level behavioral comparisons.

All of the early surveillance reports (CDC 1981a, 1981b, and 1981c) and clinical case studies (Gottlieb et al. 1981; Masur et al. 1981; Siegal et al. 1981) noted that the overwhelming majority of identified AIDS cases had occurred among men who characterized themselves as homosexuals. This observation provided the primary basis for speculation about etiological agents and behavioral mechanisms (Durack 1981; Fauci 1982; Haverkos and Curran 1982). The leading hypotheses involved behaviors presumed to differentiate male homosexuals in New York and California from the rest of the population of the United States at that time—specifically, recreational use of nitrite inhalants and sexual practices (especially penile-anal intercourse with multiple partners) through which exposure to the unknown cause(s) of this immune deficiency was presumed to occur.

Over the next several months, identification of AIDS cases among heterosexuals, injection-drug users, and hemophiliacs and other transfusion recipients pointed strongly toward a new, blood-borne, sexually transmissible agent (Jaffe et al. 1983a). That agent, the retrovirus now known as HIV, was first described in May 1983 (Barré-Sinoussi et al. 1983; Gallo et al. 1983), and its role as the cause of AIDS was established the following year (Gallo et al. 1984; Sarngadharan et al. 1984).

Following this achievement, the focus of epidemiological research shifted to clarifying the routes of transmission. This focus did not put an end to behavioral research, but resulted in a shift in the types of behavioral contrasts investigated. Whereas prior research was focused on presumed behavioral differences between homosexual males and the rest of the population (that is, group-level differences), the availability of an HIV-antibody test facilitated cross-sectional and prospective studies of prevalent and incident HIV infection and its association with specific behaviors among homosexual men (that is, individual-level differences) (for examples, see Goedert et al. 1984; Melbye et al. 1984; Nicholson et al. 1985; Boyko et al. 1986; Evans et al. 1986; Mayer et al. 1986; Schechter et al. 1986; Kingsley et al. 1987; Moss et al. 1987; Darrow et al. 1987; Winkelstein et al. 1987). These studies identified receptive anal intercourse and multiple sexual partners as the strongest and most consistent predictors of infection among men who have sex with men (Caceres and van Griensven 1994). Similar behavioral comparisons in studies of injection-drug users pointed to the sharing of needles as the primary mode of transmission in those populations (Chaisson et al. 1987 and 1989; Schoenbaum et al. 1989; Vlahov et al. 1990). Likewise, studies of heterosexuals, especially in sub-Saharan Africa, indicated that having multiple heterosexual partners was the key behavioral risk factor there, and suggested that male-to-female transmission was more efficient than female-to-male transmission during penile-vaginal intercourse (Quinn et al. 1986; Padian 1987; Johnson 1988; Johnson and Laga 1988).

Thus, during the first several years after the recognition of AIDS, between-group behavioral comparisons guided the search for the etiologic agent, and within-group behavioral variations helped to identify the primary routes of transmission. As the 1980s drew to a close, behavior continued to occupy a central position in HIV-related research, but its role shifted from independent to dependent variable. The rationale for this shift was clear. In the absence of a cure or vaccine, behavior change was the only obvious means available for controlling the spread of HIV. This sentiment was articulated as early as 1984 (see William 1984), and became a widely paraphrased dictum among AIDS researchers and health officials during the next dozen years (Curran 1985; Chin and Mann 1990; World Bank 1997). Accordingly, some investigators set out to identify the individual-level psychosocial correlates of behaviors by which HIV is transmitted (Hayes et al. 1990; Penkower et al. 1991; Basen-Engquist 1992; Peterson et al. 1992; Taylor et al. 1992; Perkins et al. 1993; Simpson et al. 1993; Catania et al. 1994; Stokes and Peterson 1998). Others developed generalized theoretical frameworks for the design of HIV-transmission-prevention programs, such as the AIDS Risk-Reduction Model (Catania et al. 1990 and 1994) and the Information-Motivation-Behavioral Skills Model (Fisher and Fisher 1992; Fisher et al. 1994). Still others developed and evaluated behavioral interventions targeting these psychosocial antecedents (Rickert et al. 1990; Kelly et al. 1991; Lemmott et al. 1992; Kipke et al. 1993; St. Lawrence et al. 1995). Several books on behavioral approaches to HIV prevention were published (for example, Valdiserri 1989; DiClemente and Peterson 1994). A few commentators (for example, Fineberg 1988) expressed doubt that large reductions in behavioral risk could be achieved and maintained in all populations, but even the skeptics conceded that a concerted effort must be made.

**Growing Interest in Biological Mechanisms**

Although enthusiasm for behavioral explanations and interventions persisted well into the 1990s, that decade was characterized also by a growing interest in biological mechanisms that influence per-contact transmission...
probabilities. The first such mechanism to capture the attention of HIV epidemiologists was the presence of other sexually transmitted infections (STIs). Statistical associations at the individual level between HIV/AIDS and other STIs were documented as early as 1983 (for example, Jaffe et al. 1983b; Piot et al. 1984). During the next several years, investigators identified numerous pathways by which other STIs might increase not only susceptibility to HIV (Weber et al. 1986; Quinn et al. 1987) but also the infectiousness of individuals already carrying the virus. Wasserheit’s (1992) seminal study provides an excellent summary of this research and its implications for synergistic interactions between HIV and other STIs at the individual and population levels (for recent reaffirmations, see Fleming and Wasserheit 1999; Corbett et al. 2002).

The first scientific test of the usefulness of biomedical treatment of other STIs as a strategy for reducing the spread of HIV at the population level was undertaken in the Mwanza District of Tanzania in the early 1990s. In this cluster randomized trial (Hayes et al. 1995), intensive syndromic management of bacterial STIs resulted in a 40 percent reduction in HIV incidence in treatment relative to control communities (Grosskurth et al. 1995). The initial optimism among HIV epidemiologists at this result gave way to disappointment and consternation when STI-treatment campaigns apparently failed to reduce HIV incidence in two subsequent community randomized trials conducted in Uganda (Wawer et al. 1999; Kamali et al. 2003). As a result, HIV epidemiologists have now turned their attention from the treatment of bacterial STIs to the management of herpes simplex virus type 2 infection in hopes that suppressing reactivation of this chronic infection may help reduce both susceptibility to and the infectiousness of HIV (Celum et al. 2004; Corey et al. 2004; Wald 2004). Early results are not encouraging (Celum et al. 2008).

A second biological determinant of the per-contact transmission probability of HIV infection is male circumcision. Epidemiological interest in this mechanism dates to the late 1980s, when a prospective study of men in Nairobi, Kenya, showed that HIV incidence was more than eight times higher among uncircumcised men than among their circumcised counterparts (Cameron et al. 1989). Shortly thereafter, two ecologic studies (Bongaarts et al. 1989; Moses et al. 1990) showed a strong inverse association between the proportion of males who were circumcised and the level of adult HIV prevalence among geographically defined sub-Saharan African populations. Since that time, numerous cross-sectional, case-control, and prospective studies of men in various parts of Africa (Malamba et al. 1994; Seed et al. 1995; Quigley et al. 1997; Urassa et al. 1997; Lavreys et al. 1999) and the United States (Kreiss and Hopkins 1993) have provided further support for this association (see O’Farrell and Egger 2000 for a review).

Beginning in the mid-1990s, these findings prompted some epidemiologists to suggest the use of circumcision services for newborn and adult males in sub-Saharan Africa as a strategy for slowing the spread of HIV (Caldwell and Caldwell 1996; Moses et al. 1998; Halperin and Bailey 1999; Weiss et al. 2000). This position has been bolstered greatly by the results of three separate randomized controlled trials of male circumcision (Auvert et al. 2005; Bailey et al. 2007; Gray et al. 2007), all revealing substantial protective effects of male circumcision. Accordingly, UNAIDS (2007) now recommends that male circumcision be recognized as an important HIV-transmission-prevention strategy, and resources from international public health organizations are now being mobilized to offer male circumcision services in sub-Saharan Africa and beyond.

Epidemiologists have also studied possible genetic influences on per-contact transmission probabilities of HIV infection (Kaslow et al. 2005), but we do not consider them here. Rather, we discuss treatment with highly active antiretroviral therapy (HAART) because of its potential policy relevance. The primary purpose of HAART is to combat the pathogenic effects of HIV replication among people who are already infected. Yet because HAART dramatically suppresses the concentration of HIV in the serum and genital secretions of treated people (Gulick et al. 1997; Gupta et al. 1997; Hammer et al. 1997; Vernazza et al. 1997), it is also believed to reduce the per-contact probability of transmitting HIV to uninfected partners (Quinn et al. 2000; Gray et al. 2001). These epidemiological consequences remain uncertain because any reduction in the per-contact transmission probability could be offset by increases in the duration of infectiousness and, perhaps, by changes in behavior. These issues have been considered in several simulation studies (Blower et al. 2000; Law et al. 2001; Gray et al. 2003; Abbas et al. 2006; Baggaley et al. 2006), as we discuss in greater detail below. In spite of this uncertainty, the potential epidemiological impacts have been an important impetus to recent efforts to make low-cost antiretroviral therapy available to HIV-infected people throughout the developing world (Muntaner et al. 2006).

Skepticism about Behavioral Mechanisms and Interventions

No necessary mutual exclusivity exists between explanations focused on biological mechanisms and those focused on behaviors. In fact, policy statements frequently
emphasize the need to integrate biomedical and behavioral components into comprehensive HIV-control strategies (UNAIDS 2007; Coates et al. 2008; Merson et al. 2008; Padian et al. 2008; Piot et al. 2008). Nevertheless, the growing focus on biological mechanisms during the 1990s was accompanied by a mounting skepticism about the contribution of behavioral variations to inequalities in the spread of HIV, as well as a rising pessimism about the potential effectiveness of behavioral interventions.

This skepticism had several causes. One was the lack of consistent evidence of population-level variations in behavior corresponding in the expected way to variations in the severity of HIV epidemics. By the 1990s, HIV epidemiologists shifted their attention from identifying contrasts between gay men and injection-drug users on the one hand and the so-called general population on the other, to seeking explanations for the severe epidemics of heterosexually transmitted HIV infection underway in parts of sub-Saharan Africa (Clumeck et al. 1984; Piot et al. 1984; Van de Perre et al. 1984). These epidemics were perplexing because no large-scale, self-sustaining epidemics of heterosexually transmitted HIV had emerged in the United States or Europe (Quinn et al. 1986; Padian 1987; Piot et al. 1988), and because studies of HIV-infected hemophiliacs and their wives in the United States suggested that HIV transmission during penile-vaginal intercourse was extremely inefficient (Kreiss et al. 1985). What, then, could explain these African epidemics?

Two primary hypotheses emerged (Biggar 1986) centered on (1) some biological cofactor (for example, other STIs) or (2) some difference in sexual behavior. Evidence for the former hypothesis was available in the form of high STI rates (Osoba 1981) and corresponding infertility (Belsey 1976) in parts of sub-Saharan Africa. Evidence that rates of sex-partner change (or other relevant behavioral parameters) were higher in sub-Saharan Africa than in the United States or Europe, however, was sketchy. No effort to collect comparable data on sexual practices in United States, European, and African populations using representative sampling and standardized interview methods had ever been undertaken. In the absence of such systematic data, some investigators drew upon information from case reports to suggest that heterosexual “promiscuity” played an important role in Africa’s HIV epidemics (Clumeck et al. 1985; Hrdy 1987). Others, citing lack of evidence, cautioned against racist stereotypes (Miller and Rockwell 1988).

Caldwell and his colleagues (1989) and Larson (1989) sought to overcome the lack of comparative survey data by compiling information from a large number of ethnographic studies conducted over many decades by anthropologists in different parts of sub-Saharan Africa. Both studies concluded that arrangements involving multiple sexual partners were common throughout the region and drew inferences about rates of partner change and patterns of sexual contact, linking these practices to kinship systems, changing economic circumstances, and cultural values in sub-Saharan Africa. Some of their conclusions were highly controversial (see Le Blank et al. 1991; Ahlberg 1994), and other investigators characterized this work as “heir to a deeply rooted legacy of Western racism” (Schoepf 1993:66), which may have had a chilling effect on this line of inquiry.

The first systematic comparative data on sexual practices in developing countries were published by the World Health Organization in 1995 (Cleland and Ferry 1995). That report provided evidence of high rates of premarital and extramarital sexual contact and multiple partnerships (Caraël 1995; Caraël et al. 1995) and low levels of condom use (Mehryar 1995) in several sub-Saharan African countries. For instance, whereas only 3 percent and 1 percent of unmarried males in Singapore and Sri Lanka, respectively, reported having sexual intercourse in the prior 12 months, the corresponding figures for Kenya, Lesotho, and Tanzania were 54 percent, 37 percent, and 33 percent (Caraël 1995). Moreover, whereas 66 percent and 52 percent of sexually experienced men in Singapore and Thailand reported ever using condoms, the corresponding figures for Kenya, Lesotho and Tanzania were 16 percent, 14 percent, and 16 percent (Mehryar 1995).

Since that time, several investigators have sought to establish the relative contributions of behavioral and biological mechanisms to variations among parts of sub-Saharan Africa in the severity of local HIV epidemics. The most ambitious of these projects was undertaken by the Study Group on Heterogeneity of HIV Epidemics in African Cities (Buvé et al. 2001). This team collected extensive biological and behavioral data from representative samples of adults in four sub-Saharan African cities with widely varying levels of HIV prevalence. They found that behavioral variables strongly associated with the risk of HIV at the individual level within each city did not vary widely between cities. In contrast, they found substantial cross-city differences in rates of male circumcision and of ulcerative STIs. The team concluded, therefore, that “differences in risky sexual behavior are outweighed by differences in factors influencing HIV transmission probability” (page 30). A similar study focusing on rural areas reached the same conclusion (Boerma et al. 2003).

In the United States, HIV epidemiologists were increasingly focused on the growing concentration of the HIV epidemic among racial and ethnic minority groups, especially African Americans (Karon et al. 2001). As in sub-Saharan Africa, efforts to explain this pattern tended
to pit the contributions of behavioral mechanisms against those of biological determinants (for reviews, see Smith et al. 2000; Millett et al. 2006). Overall, male circumcision in the United States appears to be more common among whites (81 percent) than among African Americans (65 percent), but this gap is narrower (and statistically insignificant) in younger cohorts (Laumann et al. 1997). Ulcerative STIs are markedly more common among African Americans than whites according to both passive surveillance data (Nakashima et al. 1996) and seroprevalence surveys (Fleming et al. 1997), suggesting that these STIs may well play a role. The contribution of behavioral mechanisms remains unclear, due in part to the countervailing influences of different behaviors. Data from nationally representative surveys show that African American adolescents and adults are more likely than whites to report having large numbers of sexual partners (Billy et al. 1993; Leigh et al. 1993; Ericksen and Trocki 1994; Laumann et al. 1994; Ku et al. 1998 and 2002; Holtzman et al. 2001). Yet most surveys also show that African Americans are more likely than whites to report using condoms (Sonenstein et al. 1989 and 1998; Tanfer et al. 1993; Holtzman et al. 2001; Anderson 2003).

A New Focus on Structural Constraints

During the 1990s, as the inequalities of greatest interest to HIV epidemiologists changed, the contribution of behavioral mechanisms became less clear. Interest in behavioral mechanisms was also quelled by the growing recognition that behavioral risk for HIV is often deeply intertwined with people’s social and economic circumstances, which in turn depend upon powerful macrosocial forces far beyond the reach of standard health-promotion programs. Scholarly publications on this theme began to emerge in the late 1980s, just as epidemiologists were beginning to grapple seriously with the severity of the HIV epidemics in sub-Saharan Africa. Schoepf, for example, wrote in 1988 that “AIDS has emerged and spread in Central Africa as a result of a set of historically specific economic and social relations” (page 626). One result of these circumstances was a “proliferation of multiple partner strategies” (page 628), which for women was largely motivated by a lack of alternative means of obtaining the basic necessities of life. The three factors identified by Schoepf—poverty, male labor migration, and the subordinate status of women—would be implicated again and again in the burgeoning literature on the social contexts of sexual behaviors and HIV infection in sub-Saharan Africa (for example, Ankrah 1991; Bassett and Mhloyi 1991; Caravano 1991; Jochelson et al. 1991; Sanders and Sambo 1991; de Bruyn 1992; Ulin 1992; Decosas et al. 1995; Lurie et al. 1995).

One byproduct of this structural view was a gnawing pessimism among HIV epidemiologists concerning whether people could or would change their sexual behaviors to an extent that would be sufficient to contain the spread of HIV. As early as 1993, for example, in a major review article on the epidemiology of HIV in sub-Saharan Africa, Hunter enumerated the socioeconomic barriers to behavior change and concluded that “these issues of socioeconomic development vastly complicate the search for effective strategies to influence sexual behavior in Africa” (page 69). Another review, published two years later, argued that the standard approaches to behavior change, rooted in clinical and social psychology, “have limited utility when considering social and environmental determinants of HIV infection” (Tawil et al. 1995: 1,299). The authors called for the development and evaluation of interventions that would change socioeconomic circumstances and thereby remove barriers to precautionary behavior, but offered few concrete suggestions regarding the form such interventions might take. Five years later, when the journal AIDS published a special issue devoted to structural factors in HIV prevention (Sumartojo 2000), there were still precious few examples of true structural interventions that the journal could suggest. The situation remains largely unchanged today (Gupta et al. 2008).

Meanwhile, the aggressive advance of HIV epidemics into Southern Africa, in spite of widespread knowledge of the modes of transmission and means of prevention, seemed to confirm the view that behavioral risk was too deeply embedded within powerful social forces to be changed substantially. The history of South Africa’s epidemic is perhaps the quintessential, if also the most extreme, example. As late as 1988, data from blood banks suggested that HIV prevalence in adult males and females, both African and white, remained well below one tenth of one percent in South Africa (Schoub et al. 1988). In 1990, although AIDS case reports were increasing, it still seemed possible to avert a severe HIV epidemic in that country if the appropriate steps were taken in time (Zwi and Bachmayer 1990). Some steps were taken, yet HIV prevalence increased inexorably throughout the 1990s, surpassing 20 percent nationally and 30 percent in the province of Kwa Zulu/Natal by the end of the decade (Williams and Gouws 2001). Today the South African epidemic stands as one of the most severe in the world, with adult prevalence estimated at 18 percent (UNAIDS 2008). That this epidemic, which in theory could have been averted, was allowed to unfold into such a devastating tragedy is, to many observers, a testimony to the insufficiency of sexual behaviors tied to entrenched social determinants (Campbell 2003; Allen 2004).
In this context, the appeal of relatively simple interventions targeting biological determinants of per-contact transmission probabilities became undeniable. If providing antibiotics to people with STI symptoms or providing circumcision services to uncircumcised men could substantially reduce susceptibility or infectiousness, potentially thousands of infections could be prevented and thousands of lives saved at minimal cost without requiring the amelioration of poverty, a restructuring of regional labor markets, or the elevation of women’s status.

Reconsidering the Importance of Behavior

Although gravitation toward biological mechanisms and away from behavioral determinants to reduce the spread of HIV is understandable given these circumstances, compelling arguments favor sustaining emphasis on behaviors in HIV epidemiology. We highlight several of these arguments below.

The first involves the nonlinear relationship between behavioral risk and HIV prevalence at the population level, which suggests that large reductions in HIV prevalence may be realized by small changes in behavior. Figure 1 presents a graph of the equilibrium prevalence of HIV infection as a function of the rate of partner change in a hypothetical population. The curve in this graph is derived from the equation \( P^* = 1 - 1/R_0 \) with the background mortality rate set to \( \mu = 0.02 \), the HIV-specific mortality rate set to \( k = 0.10 \), and the per-partnership transmission probability set to \( \beta = 0.06 \). In such a population, a rate of partner change of two partners per year is the threshold for epidemic take-off. At a partner-change rate of 2.1 per year, the equilibrium prevalence of HIV would be 5 percent, but a partner-change rate of 2.5 per year would lead to an equilibrium prevalence of 20 percent. Thus, a small difference in the rate of partner change leads to a fourfold difference in the equilibrium prevalence of HIV, a difference similar in magnitude to that between the countries of Southern Africa and those of West Africa. The underlying model is simple, but the basic conclusion is widely applicable: small variations in behavior—variations that would be extremely difficult to detect in behavioral surveys—can translate into substantial variations in the spread of HIV. The findings of the Four African Cities study (Buvé et al. 2001) must be interpreted with this possibility in mind. A further implication of this nonlinearity is that small reductions in behavioral risk may have dramatic population-level impacts on HIV prevalence.4

A second motivation for ongoing research on behavioral mechanisms in HIV epidemiology is the growing evidence that, across a wide range of settings, people can and do alter their sexual and drug-use practices in response to the spread of HIV. In the early 1980s, substantial declines in the incidence of gonorrhea provided the first indication of sharp behavioral changes among men who have sex with men in the United States (Judson 1983; CDC 1984) and elsewhere (Weller et al. 1984). Further evidence came from surveys of community samples of gay and bisexual men (McKusick et al. 1985a and 1985b; Martin 1987; Ekstrand and Coates 1990). Similar behavioral changes appear to have occurred among heterosexual teenagers and young adults in the United States (Sonenstein et al. 1998). The adoption of risk-reducing practices has also been documented among injection-drug users in the United States (Des Jarlais et al. 1986 and 2000) and elsewhere (Des Jarlais et al. 1994; Peters et al. 1994), including several Asian and sub-Saharan African countries (Nelson et al. 1996; Killian et al. 1999; Stoneburner and Low-Beer 2004; Gregson et al. 2006). These changes were not always rapid or dramatic, but they raise important questions about the individual and collective processes by which behavior change occurs and, in particular, about how HIV-transmission-prevention programs can help facilitate those processes. Recognizing this potential is especially important in light of the implication of the figure—that small behavioral shifts may reap large rewards in halting the spread of HIV.

Another impetus for further research on behavioral mechanisms in HIV epidemiology is that our current picture of population similarities and differences with regard to sexual behaviors may be considerably distorted by differential reporting bias in the extant research. Surveys covering sexual practices have become commonplace in developed and developing countries alike, spurred to a large extent by the spread of HIV (Curtis and Sutherland 2004). Yet the question remains: What can accurately be
inferred from these self-reported data? Investigators have long been aware of the possibility of receiving inaccurate information in survey interviews, especially where sensitive behaviors are the topic (for example, see Bleek 1987). Indeed, in their 1989 paper, Caldwell and his colleagues cautioned that “international survey programs concerning sexual practices are likely to be hardly worth the paper on which they are printed” (page 226). HIV epidemiologists are acutely aware of the shortcomings of self-reported sexual behavior data (Aral and Peterman 1996; Fishbein and Pequegnat 2000; Cleland et al. 2004). Researchers comparing such data collected from the same people using multiple methods have found alarming levels of inconsistency and troubling signs of underreporting (Plummer et al. 2004; Gallo et al. 2006). Some researchers even refuse to study self-reported sexual behaviors, preferring instead to use perceived risk of HIV infection as a proxy variable (Smith and Watkins 2005). These observations point to the need for ongoing methodological work to improve our ability to obtain accurate information about sensitive behaviors in diverse populations (Mensch et al. 2003; Gregson et al. 2004).

Another important reason for continued behavioral research in HIV epidemiology involves the recognition that the spread of HIV depends upon properties of sexual networks that are not fully captured by standard measures of individual risk, such as numbers of partners and frequency of condom use. Perhaps the most important is concurrency. Classical models of STI transmission implicitly assume that partnerships occur sequentially and do not overlap in time. When multiple partnerships are maintained concurrently, however, a newly infected person can quickly pass the infection along to several other partners. Therefore, even holding the overall number of partners per person-period constant, sexual networks characterized by high levels of concurrency have greater potential for propagating HIV. This idea was proposed in the 1990s on the basis of mathematical models and simulation studies (Watts and May 1992; Morris and Kretzschmar 1995 and 1997; Kretzschmar and Morris 1996). Since then, epidemiologists in the United States have found, in both national and city-specific samples, that African Americans’ sexual networks are characterized by higher levels of concurrency than those of whites (Adimora et al. 2002; Manhart et al. 2002). Concurrency, then, may help to explain racial disparities in HIV prevalence in the United States. It may also help explain the aggressive spread of heterosexually transmitted HIV in sub-Saharan Africa (Halperin and Epstein 2004; Epstein 2007), although few systematic studies of levels of concurrency in sub-Saharan African sexual networks have been published to date (but see Lagarde et al. 2001). Further research is needed, therefore, not only to establish levels of concurrency in the sexual networks of various populations, but also to link concurrency and perhaps other network properties to other features of social organization.

A fifth reason for continued research on behavioral mechanisms is the likelihood that social structural or cultural forces may impede attempts to realize the theoretical promise of interventions targeting aspects of HIV risk, be they behavioral or biological. Although the difficulties of mounting successful behavior-change interventions are acknowledged, similar challenges face attempts to target biological risk. The success of biomedical interventions necessary to alter susceptibility or infectiousness is likely to be partial if efforts fail to recognize or account for contextual mediators. For example, a man’s willingness and ability to be circumcised or to access and adhere to medical therapies are likely to be culturally mediated by population-specific opportunities and constraints. A related point is that most so-called biological mechanisms have behavioral components. Male circumcision is not only a status but also an act; other sexually transmitted infections also involve behavioral components at the exposure and treatment stages.

Once their effectiveness is established, interventions targeting susceptibility or infectiousness will have a reduced epidemiological impact if they are undermined by behavioral risk compensation, suggesting another compelling reason for continued research on behavioral mechanisms. For example, when interventions such as male circumcision or STI treatment succeed in reducing susceptibility or infectiousness, some people may respond to these changes by increasing their rate of partner change, decreasing their use of condoms, or otherwise changing their behavior in ways that could partially or even fully counteract the positive epidemiological effects of these interventions. The idea of risk compensation has a long history in the injury-prevention and public health literature (for a review, see Hedlund 2000). Although this possibility remains controversial in the HIV epidemiology literature, one group recently opined that risk compensation may be the “Achilles’ heel” of interventions targeting the biological determinants of HIV transmission (Cassell et al. 2006).

The nature and magnitude of behavioral risk compensation that may occur in different populations in response to interventions targeting biological mechanisms remain empirical questions. Ignoring the possibility of risk compensation, however, may lead epidemiologists to overstate the potential impacts of those interventions. Examples of this overstatement may be found in the use of simulation models to predict the impact of alternative STI-control programs on future HIV incidence and prevalence. One research group that used this approach predicted that a 50 percent reduction in the duration of
all STI episodes would prevent 43 percent of new HIV infections over a ten-year period in a Ugandan population (Robinson et al. 1995; see also Korenromp et al. 2000; Nagelkerke et al. 2002). More recently, epidemiologists used a simulation model to predict that universal circumcision of sub-Saharan African males would avert 5.7 million HIV infections and three million deaths over a 20-year period (Williams et al. 2006).

In these studies, investigators assumed that dramatic changes in rates of STIs or male circumcision would have no impact on sexual behavior. Had they assumed instead that people would increase their rates of partner change or decrease their use of condoms in response to decreased susceptibility or infectiousness, they might have concluded that STI-control programs or male circumcision would have more modest or even adverse impacts on future HIV incidence and prevalence.

The latter outcome, indeed, corresponds exactly to what investigators found in their efforts to predict the impacts of HAART on HIV incidence and prevalence using simulation studies. One group, for example, sought to predict the impact of HAART on HIV incidence and mortality among gay men in San Francisco. After conducting simulations under a range of alternative behavior-change scenarios (Blower et al. 2000), the researchers concluded that widespread use of HAART could reduce future HIV incidence and mortality, but only if risky behavior did not increase substantially. They wrote, “An increase in risky behavior of only 10 percent was enough to counterbalance the benefits of ART. Greater increases in risk behavior resulted in the incidence rate increasing and hence effectiveness becoming negative” (page 653). Similar conclusions have emerged from simulation studies of the effects of HAART on HIV prevalence in other populations, including gay men in Australia (Law et al. 2001) and heterosexual women and men in Uganda and elsewhere in sub-Saharan Africa. These findings are clearly at odds with the equation $R_0 = cBD$, in which behaviors are summarized in a fixed parameter, $c$, and they suggest that our conceptual model should be altered to accommodate the possibility of behavior change. One simple way of accomplishing this would be to replace the parameter $c$ with a behavior-change function $c(P)$ that describes how people reduce their number of partners, increase their use of condoms, or take other precautionary measures in response to increases in HIV prevalence. Indeed, some economists have adopted this strategy (Kremer 1996; Kremer and Morcom 1998). In most cases, the assumption may be reasonable that the behavior-change function for a given population takes an initial value $c_0$ when HIV prevalence is zero and that it decreases smoothly with increasing prevalence thereafter.

In this revised formulation, the standard equation for equilibrium prevalence would no longer be accurate. Instead, it would be replaced by

$$P^* = 1 - \frac{D}{[c(P^*)]\beta},$$

where $c(P^*)$ is the level of behavioral risk that prevails when prevalence has reached its equilibrium level. This equation defines the equilibrium prevalence only implicitly. Depending upon the specification of $c(P)$, solving the equation to find an explicit expression for $P^*$ in terms of $\beta, D$, and the parameters of the behavior-change function may or may not be possible. In general, however, as
long as $c(P)$ is a decreasing function, the new equilibrium prevalence is less than that given by $P^* = 1 - 1 / R_0$.

Clearly, this formulation, too, is a simplification. It assumes that the population responds instantaneously to changes in HIV prevalence. Yet several investigators (for example, Macintyre et al. 2001) suggest that behavior change occurs not in response to changes in HIV prevalence per se, but rather to changes in HIV-related mortality. This possibility, combined with the long incubation period of HIV, suggests that behavior change may lag several years behind changes in HIV prevalence. Moreover, other investigators have emphasized that behavior change in response to the spread of HIV is only partially a matter of individual choice and should also be regarded as a collective process by which people evaluate the threat of infection and consider alternative behavioral responses (Watkins 2004; Smith and Watkins 2005). These collective processes take time, generating a further lag between increasing prevalence and consequent behavioral change. Nevertheless, replacing the behavior parameter $c$ with a behavior-change function $c(P)$ in our conceptual model would go a long way toward introducing a useful element of dynamism into epidemiological thinking about the spread of HIV.

The standard formulation of $R_0 = c \beta D$ rests upon an additional simplifying assumption that we dispute: that behavior is independent of other model parameters. One way in which this assumption could be violated is if behavioral risk—or behavior change—depended systematically on susceptibility. Yet this dependency is exactly what numerous models of health behavior predict. The Health Belief Model (Strecher and Rosenstock 1997) holds that perceived susceptibility is a key determinant in the adoption of risk-reducing behavior for a broad range of health outcomes, and similar thinking is central to the AIDS Risk Reduction Model (Catania et al. 1990). The dependence of behavior upon perceived susceptibility is also apparent in our prior discussion of behavioral risk compensation. When exogenous factors create decreases in individual susceptibility to HIV infection, some affected people may respond by increasing their level of behavioral risk, since their incentive to take precautionary measures has been reduced.

This response is not the only way in which behavior could depend systematically upon other model components. Another possibility is that behavior—or behavior change—could be conditioned by the mortality parameters $\mu$ and $k$. As noted above, in simple models $\mu$ is the background mortality rate in the population and $k$ is the HIV-specific mortality rate among infected people, leading to $D = 1/(\mu + k)$. Behavior may depend upon $k$ for the obvious reason that people are more likely to take precautions against diseases that can kill them quickly. This idea is widespread in the health-behavior literature. In the Health Belief Model, for example, it takes the form of the perceived severity construct (Strecher and Rosenstock 1997). In the case of HIV epidemics, individuals might partially abandon precautionary measures if the disease became less often fatal (as a result, for example, of the widespread availability of HAART), or they might increase their use of precautionary behaviors if the disease became more often fatal (as a result, for example, of genetic mutation in the virus). Behavior and behavior change might also depend upon the background mortality rate, $\mu$. This notion has been expressed by several investigators. Writing about the problems of violence and HIV infection among urban minority populations in the United States, for example, Dow and his colleagues (1999) point out, “People who will die young from violence have little incentive to protect against HIV” (page 1,369). Other investigators (Zwi and Cabral 1991; Caldwell 2000) have articulated similar views about the role of mortality from competing risks in shaping behavioral responses to the spread of HIV in sub-Saharan African populations.

Putting these ideas together suggests that the behavior parameter $c$ in the standard formulation of the conceptual model should be replaced not just by a behavior-change function, $c(P)$, but also by a mortality- and susceptibility-dependent behavior-change function, $c(P; \beta, \mu, k)$. In such a formulation, the mortality and susceptibility parameters would operate in two distinct ways to influence the spread of HIV. First, they would operate directly, as they do in the standard formulation, to influence the per-contact transmission probability and the duration of infectiousness. Second, they would operate indirectly through their influences on behavior and behavior change. In many circumstances, the direct and indirect effects are likely to operate in countervailing directions; in other cases, they could be mutually reinforcing.

A conceptual model based upon a mortality- and susceptibility-dependent behavior-change function may have limited empirical usefulness. Like the standard formulation, it treats background and HIV-specific mortality rates, as well as susceptibility and infectiousness, as exogenously determined population parameters. It also collapses diverse sources of influence on susceptibility and infectiousness into a single, exogenously-fixed population parameter. Although, unlike the standard formulation, it allows behavior to change in response to changes in prevalence, and it allows those changes to depend upon other model parameters, it provides little guidance on the specification of $c(P; \beta, \mu, k)$ beyond the aforementioned unrealistic assumption that behavior change occurs instantaneously in response to changes in prevalence.

The potential value of this revised formulation, therefore, lies not so much in how we model the spread of
HIV in populations, but rather in how we think about the processes by which HIV spreads and in what types of research questions we pose about those processes. The standard formulation promotes static thinking about the determinants of the spread of HIV in populations and contributes to a division in HIV epidemiology between those who think that behavioral factors are more important and those who think that biological determinants of susceptibility and infectiousness play more important roles. Epidemiological researchers and international policy-setting institutions alike increasingly recognize that the division between research and prevention efforts has become an impediment to effective action to limit the spread of HIV (UNAIDS 2007). In their contribution to a recent series of essays in The Lancet about the future of HIV prevention, for example, Merson and his colleagues called for “combination prevention—i.e., a combination of behavioural, structural, and biomedical prevention paradigms and approaches adapted and prioritized to specific contexts” (2008: 485). This call was echoed in other contributions to the series (Coates et al. 2008; Padian et al. 2008; Piot et al. 2008).

To be sure, part of the challenge to achieving the vision of combination prevention is institutional. Often, one set of organizations is specialized for behavioral interventions, whereas another set is specialized for biological or biomedical ones, and the two sides compete for scarce resources rather than collaborate in the formulation of a coherent, multipronged strategy. Solving this institutional problem is important, but may not be sufficient. Altering our core conceptual model in ways that make explicit the interdependency of behavioral and biological mechanisms may be equally necessary. The alternative formulation that we propose, although unrealistic in some ways, may help to promote more dynamic thinking among HIV epidemiologists and at the same time may create opportunities for a constructive reconciliation of the behavior and biology camps.

Notes

1 We acknowledge that our focus is mainly on sexual behaviors rather than on the sharing of injection-drug equipment or other behaviors by which HIV is transmitted. This focus reflects our greater familiarity with the research literature on sexual transmission.

2 The distinction between behavioral and biological mechanisms, employed widely in the epidemiological literature, is somewhat artificial, for behaviors are governed, at least in part, by neuroendocrine systems and thus are not wholly outside the realm of biology. Likewise, most of the factors that are conveniently placed in the biological category have substantial behavioral components. Thus, although the distinction between behavioral and biological mechanisms may be analytically convenient, it is also arbitrary.

3 Models of this type, typically specified as systems of differential equations, have a long history in infectious disease epidemiology (Kermack and McKendrick 1927; Hethcote 1976). They were first applied to sexually transmitted HIV in the mid-1980s (Anderson et al. 1986; May and Anderson 1987). In these models, the theoretical significance of $R_0$—the basic reproductive number—is then threefold (Anderson et al. 1986; May and Anderson 1987; Anderson 1999). First, its value determines whether ongoing transmission of HIV can be sustained in a population. When $R_0 < 1$, each infected person infects fewer than one susceptible person, on average, and therefore an epidemic cannot be maintained. Second, $R_0$ is related to the initial growth rate of the epidemic according to the doubling-time equation $t_d = 0.693 \times D / (R_0 – 1)$. Third, $R_0$ is related to the equilibrium prevalence toward which the epidemic converges over time, which for simple models (and for $R_0 \geq 1$) is expressed as $P^* = 1 – 1 / R_0$. Thus, $R_0$ provides an appealing summary measure of the severity of an HIV epidemic in a population. Technically, these formulas are derived from very simple—and therefore unrealistic—models. More complicated generalizations of the same types of models have proved useful in elucidating several aspects of HIV epidemiology. These include the role of heterogeneity in sexual activity and patterns of mixing between high and low sexual activity classes (Blythe and Anderson 1988; Jacquez et al. 1988; Koopman et al. 1988; Gupta et al. 1989; Garnett and Anderson 1993 and 1994) as well as the epidemiological consequences of the staged progression of AIDS and especially of elevated infectiousness during primary infection (Hethcote and Van Ark 1992; Jacquez et al. 1994; Koopman et al. 1997). These models lead to complex mathematical expressions for $R_0$.

4 This assertion contradicts the claim that “radical” behavioral change is required to bring about reductions in HIV transmission at the population level (Coates et al. 2008).

References


202 Studies in Family Planning


Acknowledgments

This research was supported by an NIA grant to the Michigan Center on the Demography of Aging, a grant from the Eunice Kennedy Shriver National Institute of Child Health and Human Development to the Population Research Institute at Pennsylvania State University, a fellowship to Arline T. Geronimus from the Center for Advanced Study in the Behavioral Sciences at Stanford University, and the Robert Wood Johnson Health and Society Scholars Program at the Harvard School of Public Health. We thank John Bound, Stephen Raudenbush, Carl Simon, and Rachael Snow for thoughtful feedback on earlier versions of this study.