A Review of Methods to Evaluate HIV Prevention Interventions for People Who Inject Drugs in Low- and Middle-income Countries

Methods, Gaps, Challenges and Opportunities

Theodore M. Hammett, Ph.D.
Abt Associates

Danielle S. Parsons-Smith, M.P.H.
Consultant, Washington, DC

July 29, 2014
ABSTRACT

Background
The inter-related epidemics of HIV and injection drug use are global in geography but concentrated in low- and middle-income countries (LMIC). There has been substantial evaluation of HIV prevention programs for people who inject drugs but persistent debate about, and generally low uptake and coverage of, interventions in LMIC.

Methods
This paper reviews the major designs, methods, and types of data used to evaluate HIV prevention programs for people who inject drugs in LMIC, their strengths and weaknesses, and the gaps in the literature and evaluation practice. It is based on non-exhaustive search of peer-reviewed and “gray” literature designed to develop estimates of the frequency and informative examples of different types of evaluations.

Results: The most common evaluation designs found in this review are those based on routine process data (actually, monitoring rather than evaluation) and outcome evaluations comparing pre-post observations of risk behaviors in a single site. There have also been some ecological studies of coincident interventions and epidemiologic and behavioral trends. The major gaps in evaluation occur in rigorous outcome evaluations employing biological measures and appropriate comparisons, such as quasi-experimental designs and true experiments (randomized controlled trials) as well as in attempts to disentangle the effects of individual elements of combination prevention. There have been a limited number of cost-effectiveness and cost-benefit studies but no assessments that take into consideration the contributions by, or effects of interventions on, larger health systems.

Recommendations
Recommendations from the review include addressing the gaps in rigorous evaluation, as well as designing evaluations based on political and ideological context (e.g. receptivity to evaluation and specific types of data most likely to produce desired policy and program improvements), providing data to inform feedback loops and midcourse corrections in interventions, considering natural epidemic progression and programmatic factors in the interpretation of evaluation data, incorporating assessment of the sustainability of interventions, making better use of evaluation results to advocate for improvements in laws, policies, and programs, and devising better methods for evaluation of combination prevention, economic analysis, and understanding the relationships between interventions and larger health systems.
INTRODUCTION

The inter-related epidemics of HIV and injection drug use are global in geography but concentrated in low- and middle-income countries (LMIC) (Mathers, Degenhardt, Phillips, Wiessing, Hickman, et al., 2008; Ball, Rana & Dehne, 1998). There has been substantial evaluation of HIV prevention programs for people who inject drugs (PWID) but persistent opposition to harm reduction interventions in some countries, some politically and ideologically motivated, as well as resource constraints have resulted in generally low uptake and coverage of these interventions in LMIC. There remain significant gaps in evaluation methods and data and in the effective use of evaluation results to advocate for and inform policies and programs.

Data on size of and HIV prevalence among PWID populations are uneven in scope and quality, leading to wide ranges of estimates: for example, an estimated 16 million PWID worldwide but with a range of 11-21 million (Beyrer, Malinowska-Sempruch et al., 2010). About 25% of PWID live in Asia (Bergenstrom and Abdul-Qadar, 2010). As of 2008, 148 countries reported injection drug use and 120 reported HIV related to drug injection. Globally, 10% of all HIV infections are injection-related but the figure is 30% outside of Africa (Mathers et al., 2008).

HIV epidemics driven by drug injection are prevalent in Asia (Commission on AIDS in Asia, 2008; Dokubo et al., 2013), Eastern Europe and the former Soviet Republics, Central Asia, and Latin America, and more recently emerging in East and South Africa (Neburg & Carty, 2011). HIV prevalence among PWID in East and Southeast Asia is 15%-20% and higher than 20% in Eastern Europe (Dutta, Wirtz, Stanciole, et al., 2013). Injection-related HIV infection is disproportionately found among ethnic minorities who tend to be of lower socioeconomic status within countries (Des Jarlais, Bramson, Wong, Gostnell, Cepeda et al. 2012).

While heroin injection has been the primary drug-related risk factor for HIV, emerging use of amphetamine-type stimulants (ATS) and other injected and non-injected drugs in LMIC has also been associated with HIV, primarily by increasing risk of sexual transmission. To date, however, evidence of such links is inconclusive.

In addition, there have been few if any HIV prevention interventions specifically targeting ATS users (Degenhardt, Mathers, Guarnieri, Panda, Phillips, et al., 2010). Therefore, this paper will focus on HIV prevention among PWID and primarily among heroin injectors.

AIDS prevention efforts among PWID began almost simultaneously with the emergence of AIDS in this group. Des Jarlais and colleagues showed that in New York City, communities of PWID changed their injection behaviors without any formal interventions (Des Jarlais, Friedman & Hopkins, 1985; Des Jarlais & Friedman, 1988). By the middle- to late-1980s, peer education and needle exchange programs had emerged in a number of cities with substantial populations of PWID in the developed world, especially Europe, the U.S., and Australia (Des Jarlais & Semaan, 2008). Dual epidemics of HIV and injection drug use were not recognized in LMIC until the 1990s although they may actually have begun earlier. HIV prevention interventions were slower to appear and often met resistance as they did in the U.S. and some developed countries.

In recent years, it has been asserted that a comprehensive package of interventions is needed for effective control of HIV among PWID and their partners. The World Health Organization, UNAIDS and UNODC (2009) published a list of 9 prevention, care, treatment, and support interventions comprising a comprehensive package: needle/syringe programs (NSP); opioid substitution treatment (OST) and other drug dependence treatment; HIV testing and counseling (HTC); antiretroviral therapy (ART) for HIV disease; prevention and treatment...
for sexually transmitted infections (STI); condom programs for PWID and their sexual partners; targeted information, education and communication (IEC) for PWID and their partners; vaccination and treatment for hepatitis; and prevention and treatment of tuberculosis. A revised version, issued in late 2012 (WHO, UNODC & UNAIDS, 2012), reaffirmed the importance of these interventions, with a strengthened emphasis on quality to achieve increased impact.

Other organizations and authors have echoed the call for combining treatment and care with primary prevention in order to have the maximum interactive effect to prevent and control HIV among PWID (National Institute on Drug Abuse & International AIDS Society, 2010; Marshall and Wood, 2010; Degenhardt, Mathers, Vickerman, Rhodes, Latkin, et al., 2010a). It may be more feasible to implement comprehensive or combination prevention in the developed world but the concept is relevant to LMICs where limited resources may suggest leaner and more carefully chosen groups of interventions (Vlahov, Robertson & Strathdee, 2010).

Within each element of the UN package there are subcategories and elaboration is needed. There are various types of NSPs including strict one-for-one exchange, broad distribution with no limits or requirements to return used needles/syringes, and programs that employ pharmacies and other outlets in addition to fixed sites and mobile distribution by outreach workers. OST has predominantly employed methadone, but buprenorphine, suboxone (combination buprenorphine and naloxone), and naltrexone have also been used in various settings, principally in the developed world.

“The inclusion of condom programs on the WHO list of interventions highlights the need to address sex-related risks among PWID and their partners and the persistent inattention to sexual risk reduction in most HIV prevention programs for PWID. IEC programs may include peer delivery of information and counseling as well as network interventions, and media-based programs. Under the headings of HTC and ART, the concept of treatment as prevention (for PWID, often presented as “seek, test, and treat”) has received substantial attention in recent years as have pre- and post-exposure prophylaxis, although these are probably not feasible for large-scale implementation in LMIC. The inclusion of hepatitis highlights the extremely high prevalence of hepatitis C virus (HCV) infection among PWID worldwide and the high rates of HCV-HIV coinfection, as well as the facts that HCV is similarly but more easily transmitted than HIV and thus interventions that should address both are in fact more effective in controlling HIV than HCV (Wright & Tompkins, 2006; Palmateer, Kimber, Hickman, Hutchinson, Rhodes & Goldberg, 2009).

The scale and targeting of interventions are also topics of interest. Public-health scale programs that achieve very high coverage of target populations are probably more likely to be effective. However, “structural interventions” that rely on changing social norms or the physical or social context of risk taking, removing barriers, and promoting an enabling legal/policy environment at population or community level without directly reaching every member of the target group may be more cost-efficient (Degenhardt et al., 2010a).

Some of the most widely implemented and effective HIV prevention interventions for PWID, such as peer outreach and education, NSP, and OST, are included under the category of ‘harm reduction’.”
opposition on just these grounds — that they are based on acceptance of drug use rather than its prohibition or punishment and may be seen as giving permission and even encouragement to the use of drugs. In some countries, such as China, Vietnam, Iran, and Malaysia, practical evidence of harm reduction’s effectiveness has at least partially overcome ideologies that label drug addiction a “social evil” attributable to moral failure. In the U.S., Russia, Sweden, and some other countries, by contrast, persistent misinformation that such interventions will increase drug use even in the face of evidence to the contrary (e.g., Institute of Medicine, 2007) have perpetuated legal prohibitions on or policy-based discouragement of NSP, OST and other harm reduction interventions.

Alex Wodak, a leading advocate for harm reduction, has declared repeatedly that the scientific debate on the effectiveness of harm reduction is over (Wodak, 2008; Wodak, 2006) and perhaps it should be. In many countries including LMIC, however, debate and opposition persist, as do shortages of resources and political will and continuing stigmatization of people who use drugs (Des Jarlais, Pinkerton, Hagan, et al., 2013). Thus, there remains a need for rigorous evaluation and the marshaling of evaluation data and other evidence to inform policy and program decision-making. To improve the quality of evidence, the focus of evaluation should shift from process measures to outcomes and impacts and from demonstration of effectiveness alone to cost-effectiveness and cost-benefit analysis (Institute of Medicine, 2013).

This paper reviews the major designs, methods, and types of data used to evaluate HIV prevention programs for PWID in LMIC and the strengths, weaknesses and gaps in the literature and evaluation practice, and concludes with some recommendations (for a summary, see Hammett & Parsons, 2014).

METHODS

We carried out searches of peer-reviewed published literature on PubMed and PsychInfo, and supplemented the search with Google/Google Scholar, which identified items in the “gray” literature. Key words searched were combinations of “HIV prevention, drug users, people who use drugs (or PWID), AND evaluation.” We did not conduct an exhaustive search or systematic review of the literature, as numerous reviews of various aspects of this field have already been done. Instead, we developed estimates (but not precise quantitative tabulations) of the frequency of methods used to evaluate HIV prevention programs for PWID in LMIC. We also sought informative examples of these designs and identified key gaps, barriers, challenges and opportunities for the future. While we focus on designs and examples from LMIC, we also cite some from developed countries to illustrate important points when no or only very limited LMIC examples could be found.

RESULTS

Large reviews of the evaluation literature on key components of a comprehensive package of HIV prevention for PWID have already been done and most of these suggest the effectiveness of specific interventions such as NSP, peer outreach and education, and OST. In 2007, the Institute of Medicine (IOM) in the U.S. conducted an extensive review of the evidence regarding peer education, NSP, and OST in countries with injection-driven HIV epidemics and concluded that peer education and NSP were effective in reducing HIV risk behaviors in the context of broader multi-component programs but that the evidence linking these interventions to reductions in HIV incidence was inconclusive. The IOM report found only OST to have a stand-alone positive effect (Institute of Medicine, 2007).

Several other reviews concluded that OST was effective in reducing addiction severity and improving quality of life (Feelemyer, Des Jarlais, Arasteh, Phillips, and Hagan, 2014) and reducing HIV risk behaviors and/or HIV prevalence among PWID (WHO, 2004; Gowing,

Reviews of NSP back to the mid-1990s have consistently concluded that such programs are effective in reducing HIV risk behaviors but with limited and less consistent evidence from trends in prevalence and incidence (Institute of Medicine, 1995; WHO, 2004a; Wodak & Cooney, 2006; Palmateer, et al., 2009). Several of these employed systematic review criteria such as those of Bradford Hill, which include consistency of findings across settings as a criterion for attributing causal connections between interventions and outcomes (Wodak & Cooney, 2005). Recent reviews continued to find mixed and somewhat inconsistent evidence on the relationship between NSP and reduced HIV incidence (Rhodes and Hedrich, 2010) but one concluded that, with high coverage, NSPs can reduce HIV and HCV infection as much in LMICs as they have in developed countries (Feelemyer, et al., 2013a).

The evidence on community-based interventions, such as peer outreach and education, has also been positively reviewed (Needle, Burrows, Friedman, Dorabjee, Touze et al., 2005; WHO, 2004b; Medley, Kennedy, O’Reilly & Sweat, 2009). Psychosocial interventions received somewhat positive reviews, albeit acknowledging that substantial downward trends in risk behaviors were typically identified in both intervention and comparison groups (Meader, Li, Des Jarlais & Pilling, 2010). By contrast, Des Jarlais and Semaan (2005) concluded that interventions designed to reduce sexual risk behaviors among PWID were not demonstrably effective, a finding that further highlights a common weakness in HIV prevention programs for PWID.

A major gap either stated or implied in these reviews was the inability of extant studies to disentangle the separate effects of elements of multi-component HIV prevention programs (Degenhardt et al., 2010a; Institute of Medicine, 2007). However, a recent World Bank project began to address this gap by modeling outcomes based on different scenarios of scale-up and various combinations of interventions (Dutta, et al., 2013).

### Review of Evaluation Methods

Table 1 summarizes our estimates of the frequency of major evaluation designs and data types used in evaluations of HIV prevention programs for PWID in LMIC. (It is not intended to summarize all types of evaluation designs but only those used for these types of interventions.) Each of these utilized designs is discussed in more detail below. Notably, the list includes both monitoring and evaluation and the evaluation designs range from the strongest (randomized controlled trials) to the weakest (single-site pre-post comparisons) designs in terms of ability to attribute causation. Moreover, the designs are not mutually exclusive but are sometimes used in combination with multiple data types as in mixed methods bringing together quantitative and qualitative data. It is also important to emphasize that all evaluation designs, no matter how rigorous or powerful, are subject to measurement issues such as differential loss to follow up and bias in self-reported data. Assessment of the strength of evaluation findings must take into account both design and measurement issues.

<table>
<thead>
<tr>
<th>Evaluation Design</th>
<th>Strength of Evidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>Randomized Controlled Trials</td>
<td>Strong</td>
</tr>
<tr>
<td>Single-site Pre-Post Comparisons</td>
<td>Weak</td>
</tr>
<tr>
<td>Mixed Methods</td>
<td>Medium</td>
</tr>
</tbody>
</table>

“... all evaluation designs, no matter how rigorous or powerful, are subject to measurement issues such as differential loss to follow up and bias in self-reported data. Assessment of the strength of evaluation findings must take into account both design and measurement issues.”
Table 1: Designs, Frequency, and Data Used for Evaluations

<table>
<thead>
<tr>
<th>DESIGNS</th>
<th>FREQUENCY</th>
<th>DATA ELEMENTS/SOURCES</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>PROCESS/IMPLEMENTATION (SINGLE/MULTI-SITE)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Services/coverage (monitoring)</td>
<td>Common</td>
<td>Clients, services, commodities, population size</td>
</tr>
<tr>
<td>Structures/barriers</td>
<td>Uncommon</td>
<td>Laws, policies, relations with police/other stakeholders; qualitative data</td>
</tr>
<tr>
<td><strong>OUTCOME/IMPACT</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Observational</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Within-site, pre-post comparison</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Behavioral outcomes</td>
<td>Common</td>
<td>Cross-sectional/cohort behavioral surveys; surveillance data; qualitative data (mixed methods)</td>
</tr>
<tr>
<td>Biological outcomes</td>
<td>Very uncommon</td>
<td>HIV testing; surveillance data</td>
</tr>
<tr>
<td>Ecological (single/multi-site)</td>
<td>Uncommon</td>
<td>Secondary /surveillance data</td>
</tr>
<tr>
<td>Quasi-experimental</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Multi-site (e.g., concurrent, crossover)</td>
<td>Uncommon</td>
<td>Cross-sectional/cohort behavioral surveys, HIV testing; qualitative data (mixed methods)</td>
</tr>
<tr>
<td>Single site (e.g. stepped wedge)</td>
<td>Very uncommon</td>
<td>Cross-sectional/cohort behavioral surveys, HIV testing; qualitative data (mixed methods)</td>
</tr>
<tr>
<td>Experimental (RCT) (single/multi-site)</td>
<td>Very uncommon</td>
<td>Cross-sectional/cohort behavioral surveys, HIV testing</td>
</tr>
<tr>
<td>Economic and systems evaluations</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cost-effectiveness/benefit</td>
<td>Uncommon</td>
<td>Costs, outcome, behavioral surveys, surveillance data</td>
</tr>
<tr>
<td>Health Systems analysis</td>
<td>Very uncommon</td>
<td>Services, cost data, systems performance indicators</td>
</tr>
</tbody>
</table>

**Process/Implementation**

The most common method for “evaluating” HIV prevention programs for PWID in LMIC is collection and analysis of process or implementation data from single or multiple sites, which really constitutes monitoring rather than evaluation. Process monitoring cannot determine whether an intervention was successful in achieving its objectives but only whether it was successfully implemented. This is an important preliminary question but by no means the only question of interest. The frequency of purely process monitoring is in part the result of the fact that major donors such as PEPFAR, Global Fund, and World Bank require primarily the collection and reporting of process or implementation data rather than outcome or impact data. This may result from a desire to define individual inputs and outputs attributable to the support provided by the specific donor, rather than impact, which is more difficult to attribute to a particular donor or program. It may also stem from the donor's conviction, correct or incorrect, that the intervention works and so tracking outcomes is unnecessary.

The recent emergence of implementation science signals development of more sophisticated and outcome-oriented analysis employing process data to understand the barriers and facilitators of achieving the desired outcomes of interventions (Schackman, 2010).
Imported and common forms of process monitoring tabulate the number of individuals reached, the amount of services provided, and/or the quantity of commodities distributed in relation to the size and needs of the target populations. Coverage of programs may be measured in terms of geography (percentage of provinces or districts with certain interventions) or population (percentage of the target population reached). As noted, such analyses may be based primarily on routine project monitoring required by donors.

Accurate calculation of population coverage requires measures of both numerator (number reached) and denominator (target population size). There are challenges to both. Many HIV prevention interventions for PWID involve multiple contacts with clients by peer outreach workers or fixed site service providers but an accurate numerator must be the number of different individuals reached. Some programs have employed unique identifier codes (UIC) for this purpose. Use of UICs can provide needed monitoring, management, and planning data, such as analysis of patterns of service receipt, while maintaining the anonymity of clients, invariably a serious concern of drug users. UICs have been successfully implemented with HIV prevention and drug dependence treatment programs in Central Asia, Vietnam, and other LMIC (Zheluk & Burrows, 2007). However, establishment of a universal UIC for use throughout a country may be challenging, which limits the ability of UIC-based data to inform national-level rather than project-specific coverage estimates.

Estimation of denominators or target population sizes is also challenging. Several methods have been used, including multiplier, capture-recapture and respondent-driven sampling (RDS). The multiplier method employs the total number of individuals accessing a service and the estimated percentage of the population represented in that access figure (Stimson, Hickman, Rhodes, Bastos & Saidel, 2005). Capture-recapture is another indirect method in which counts of individuals appearing in several independently generated but overlapping datasets are used to estimate the number of population members that do not appear in any of them, a figure that can then be used to estimate the total population size. Capture-recapture has been used to estimate the size of PWID populations in Estonia (Uuskula, Rajaleid, Talu, Abello & Des Jarlais, 2013) and Iran (Khazei, Poorolajal, Mahjub, Esmailnasab & Mirzaei, 2012).

Respondent-driven sampling is a form of “snowball” or chain referral sampling based on distribution of coupons and rewarded but quota-limited referrals from initial “seeds” in the target population through expanding networks or chains, which can be used in models to estimate the demographic, racial/ethnic, and socioeconomic characteristics of the population, and total population size, along with standard errors of those estimates. RDS is an appealing, but not fool-proof, method for sampling hidden or hard-to-reach populations for which frames for standard probability sampling are not available (Heckathorn, 2002). To be effective, RDS must still rest on sound sampling logic. In the Indian Ocean island of Mauritius, data on RDS (recovery during survey recruitment of unique cards that had been previously distributed to PWID) were combined with data on PWID attendance at key service providers to generate estimates of the size and characteristics of the IDU population (Johnston, Saumtally, Corceal, Mahadoo, & Oodally, 2011).

Estimates of intervention coverage may be used for planning and evaluation. There are varying standards of coverage depending on program objectives. Des Jarlais...
and colleagues argue that to achieve most desirable outcomes, interventions for PWID such as NSP must be implemented at public health scale as opposed to pilot scale (Des Jarlais & Semaan, 2008). This refers to both geographic and population coverage and elsewhere Des Jarlais defines 50% as high coverage for PWID interventions in LMIC (Des Jarlais, Feelmyer, Modi, Abdul-Qadar & Hagan, 2013a). Burrows presents case studies of high-coverage interventions for PWID (Burrows, 2006) and WHO has published a technical guide for achieving universal access to HIV prevention for people who use drugs, with a focus on comprehensive and integrated services (WHO, UNODC & UNAIDS, 2012). However, a 2010 review concluded that actual coverage of HIV prevention programs for PWID remains extremely low worldwide, and especially so in LMIC (Mathers, Degenhardt, Ali, Weissing, Hickman, et al., 2010). An exception is that the Chinese government’s national methadone treatment program is estimated to reach almost one-half of the country’s PWID (Bergenstrom & Abdul-Qadar, 2010).

Structures/Barriers and Facilitators

Another important, but often under-emphasized, dimension of process assessment is the analysis of structures and factors comprising the enabling environment for interventions. These include ideology, politics, laws and policies, and relationships with and collaboration among multiple government sectors such as police and non-governmental stakeholders such as civil society organizations. Such analyses do not appear commonly in the peer-reviewed literature on HIV prevention for PWID in LMIC. Examples include studies of police attitudes toward and collaboration with harm reduction programs in Vietnam (Jardine, Crofts, Monaghan & Morrow, 2012; Khuat, Nguyen, Jardine, Moore, Bui, & Crofts, 2012; Hammett, Bartlett, Chen, Ngu, Cuong et al., 2005) and other Southeast Asian countries receiving funding from an AusAID (whose name has since changed to Department of Foreign Affairs and Trade [DFAT]) regional HIV/AIDS project (Sharma & Chatterjee, 2012). A paper on Ukraine details how police inhibit PWID access to harm reduction programs (Arkin, 2011).

A paper from Malaysia highlights the role of civil society organizations in winning the adoption of harm reduction policies and programs and encouraging drug users to attend these programs (Narayanan, Vicknasingam, & Robson, 2011) while a study of Eastern and Central Europe emphasizes the lack of government funding to well-qualified civil society organizations to implement interventions for PWID (Amirkhanian, Kelly, Benotsch, Somlai, Brown et al., 2004). More accounts of the role of civil society in promoting and delivering harm reduction services for drug users appear in the “gray” literature and on organizational websites than in peer-reviewed literature.

Several studies of Asian countries have examined more broadly the effects of prevailing ideologies and “grammars” of harm reduction and “social evils” on the pace and patterns of implementing harm reduction interventions such as NSP and OST (Hammett, Wu, Duc, Stephens, Sullivan et al., 2008; Hammett, Des Jarlais, Johnston, Kling, Ngu et al., 2007; Edington & Bayer, 2012; Smith, Bartlett, & Wang, 2012) and evolution of the legal and policy framework on drug use in relation to HIV/AIDS (Vuong, Ali, Baldwin, & Mills, 2011; Sharma & Chatterjee, 2012a). A positive instance of evidence driving policy change comes from Malaysia, which moved to a system of voluntary “cure and care” substance abuse treatment centers based on evidence of the failure of its longstanding program of compulsory detention of drug users (Phaik, 2011).

Outcome/Impact

Full-scale evaluations of interventions must go beyond process measures to examine outcomes and impact (Institute of Medicine, 2013). There is some variation in the respective definitions of outcome and impact. Some use the terms almost interchangeably while others define outcome as the short-term effectiveness of an
intervention’s outputs in achieving its objectives and impact as longer term sustainable changes produced by an intervention whether directly or indirectly, intended or unintended. In any event, for a variety of reasons including cost, ethical concerns, technical capacity, and political considerations, outcome and impact evaluations of HIV prevention interventions for PWID are quite uncommon. In this section, we consider three basic types of outcome or impact evaluations: 1) observational studies; 2) true experiments (randomized controlled trials); and 3) economic and systems evaluations. Under each category, we consider both design and measurement issues. Issues of statistical power are also relevant to many of the types of evaluations discussed and must be addressed in the design phase so that sufficient sample sizes are established to detect the differences of interest and answer other identified evaluation questions.

Observational Studies

Observational studies, which include within-site, pre-post comparisons, ecological studies, and quasi-experimental designs, are by far the most common form of evaluation found in this review. These evaluations may examine single or multiple sites with the same or different interventions but do not involve any randomization. Observational studies raise a number of issues and challenges. As discussed in more detail below, they range from studies that pose major to minor validity concerns and thus limit inferences of causation between interventions and observed outcomes.

The most common form of observational study is based on pre-post comparison within sites and may employ cohort or cross-sectional data from project-specific surveys or existing behavioral and/or biological surveillance. “Dose-response” approaches may also be used, which assess association between extent of change in outcome measures and “dose” of intervention or services received.

Most evaluations assess changes in behavioral outcomes such as risk behaviors but much less commonly employ biological outcomes such as HIV prevalence and incidence trends. Biological outcome measures are generally more robust since they are objective (assuming the accuracy of test results) whereas behavioral measures are normally self-reported and thus potentially subject to biases.

Prevalence trends may also be affected by factors exogenous to the intervention such as death rates, mobility, instability or other changes in the target population forming the denominator. Still, major downward trends in prevalence probably reflect at least some reduction in incidence.

HIV incidence may also be affected by or attributable to external factors. For example, an incident HIV infection in a PWID may have been caused by sexual contact rather than the sharing of injection equipment that the intervention being evaluated is intended to prevent. Nevertheless, incidence offers the most direct and robust outcome measure for HIV prevention programs. The earliest evaluation using HIV incidence as an outcome was reported by Weibel (Weibel, Jimenez, Johnson, Ouellet, Jovanovic et al., 1996) based on results from a cohort of PWID in Chicago.

An evaluation of combined harm reduction interventions for PWID in two states of India offers a good example of a design employing pre-post cross-sectional surveys and dose-response analysis of behavioral outcomes (Armstrong, Humtsoe, & Kermode, 2011). Nguyen and colleagues used a cohort design with behavioral outcomes to evaluate a pilot methadone program in Haiphong, Vietnam (Nguyen, Nguyen, Pham, Vu, & Mulvey, 2012). Several papers on a Cross-Border HIV prevention intervention for PWID in Northern Vietnam and Southern China provide examples of evaluations.
based on serial cross-sectional surveys and employing both behavioral and biological outcomes. Twenty-four month evaluation results found significant reductions in drug-related risk behaviors and HIV prevalence among PWID in project sites (Hammett, Kling, Johnston, Liu, Ngu, Friedmann et al., 2007a). Later reports on the Cross-Border evaluation by Des Jarlais and colleagues (Des Jarlais, Kling, Hammett, Ngu, Liu et al., 2006) and an evaluation of an expanded NSP in Estonia by Uuskula and collaborators (Uuskula, Des Jarlais, Kals, Ruutel, Abel-Ollo et al., 2011) both employed prevalence data from cross-sectional surveys to estimate HIV incidence among new heroin injectors. The Cross-Border Project’s evaluation was the first in an LMIC to use HIV incidence as an outcome. A report on eight years of the Cross-Border project reported positive trends in behaviors, HIV prevalence, and HIV incidence following implementation of peer-based NSP, based on new injector analysis and testing for recent infection (Hammett, Des Jarlais, Kling, Kieu, McNicholl, Wasinrapee et al., 2012). A prospective cohort study in Xichang City, Sichuan Province, China found a significant drop in HIV seroincidence following implementation of a multi-faceted harm reduction program (Ruan, Liang, Zhu, Li, Pan, et al., 2013). By contrast, a study in Yunnan Province, China found no association between presence of harm reduction interventions and trends in HIV incidence estimated from cross-sectional data (Duan, Shen, Bultery, Jia, Yang et al., 2010).

Another form of observational evaluation using biological data assessed temporal trends in HIV-contaminated syringes discarded in New Haven, Connecticut before and after scale-up of an NSP. In this evaluation, reduction in percentage of discarded syringes that were contaminated with HIV was used to argue that the NSP reduced the sharing of injection equipment and associated HIV prevalence among PWID (Kaplan & Heimer, 1992; Heimer, Kaplan, Khoshnood, Jariwala & Cadman, 1993). This evaluation design has not been used in any LMICs.

Cohort studies are generally preferable to cross-sectional designs because they allow the direct observation of changes in individuals’ behavior, HIV status, or other measures over time, whereas cross-sectional survey data may mask patterns of change in individuals. However, cohort studies may suffer from loss to follow up that reduces the power of results. “Structural interventions” that reach extremely high coverage or change norms of behavior among members of a target population beyond those directly reached by the services may be appropriately evaluated at the community level using serial cross-sectional surveys (Hammett et al. 2012; Des Jarlais, 2000).

Survey data may be subject to recall and social desirability bias. Recall bias may be ameliorated by calendar-based survey questions. Social desirability may be mitigated to some extent by use of audio-computer assisted self-interview (A-CASI) technology but this is expensive and has been rarely used in LMICs. More broadly, participation in the evaluation itself and its attendant interviews, clinic visits, or other activities may influence outcomes (Hawthorne effects) (McCarney, Warner, Iliffe, van Hasalen, Griffin, et al., 2007).

There may be selection bias both into interventions (i.e. those most likely to change their behaviors may be those most likely to participate in the intervention, which may obscure the true cause of observed change) and/or into the evaluation (i.e. in the non-random sampling most commonly used in surveys of PWID and other “hidden” or stigmatized populations). There are several methods of adjusting for selection bias. Instrumental variables — those that do not themselves belong in the explanatory equation and are correlated with the endogenous explanatory variables, e.g. a policy change or clients’ distance from the intervention — may be used in regression models (McClellan, McNeill, & Newhouse, 1994). Propensity score analysis, which allows comparison of the effects of an intervention among groups with a similar likelihood of contacting that intervention, was used in the evaluation of an HIV

Most evaluations of any design have been of short duration. Therefore, they are unable to measure the durability of an intervention’s effects. Most long-duration evaluations have been among men who have sex with men rather than people who use drugs (Coleman & Ford, 1996). An exception is the Cross-Border project in Vietnam and China that was evaluated over an eight-year period (Hammett et al., 2012).

HIV incidence is the most robust outcome measure for evaluating HIV prevention interventions but it also poses challenges. Incidence estimation based on HIV prevalence among new injectors assumes that all observed new infections are the result of injection-related risk behavior and that the subject was uninfected at initiation of injection. In fact, some infections may be the result of sexual contact before initiation of injection, which artificially inflates the estimated HIV incidence (Hammett et al., 2012).

Tests for recent infection offer an alternative source for estimating incidence but there are also problems with these, especially the BED assay that may give false-recent results for individuals on ARV treatment or with long-term HIV infection (Welte, McWalter, Laeyendecker & Hallett, 2010; Marinda, Hargrove, Preiser, Slabbert, van Zyl et al., 2010; Hallett, Ghys, Barnighausen,Yan, & Garnett, 2009; Laeyendecker, Brookmeyer, Oliver, Mullis, Eaton, et al., 2012; Leyendecker, Brookmeyer, Mullis, Donnell, Lingappa, et al., 2012a; Karita, Prie, Hunter, Chomba, Allen, et al., 2007; Havashida, Gatanaga, Tanuma, & Oka, 2008), although a recent study in Thailand revealed very few false-recent results on the BED assay (McNicholl, McDougal, Wasinrapee, Branson, & Martin, 2011). Various location-specific and universal adjustment factors have been used to correct for false-recent BED test results in incidence estimation (Welte, McWalter, & Barnighausen, 2009). Newer tests for recent HIV infection, such as the LAg-avidity assay, show promise of reducing false-recent results but studies comparing incidence estimates based on LAg-avidity and BED tests have been inconsistent. In a Vietnam study LAg-based incidence was higher than BED-based incidence (Sexton, Costenbader, Vinh, Chen, & Hoang, 2011), while a study in Swaziland found that LAg-based incidence and incidence observed in a cohort were similar, but the BED-based estimate for the same group was five times higher (Parekh, Duong, Mavenger, Dobbs, Chang, et al., 2012).

There is work underway to develop new methods and technologies for identifying recent HIV infection (Mastro, 2013; Incidence Assay Critical Path Working Group, 2011), including multi-assay algorithms (Laeyendecker, Brookmeyer, Cousins, Mullis, Konikoff et al., 2012b; Laeyendecker, Piwowar-Manning, Fiamma, Kulich, Donnell et al., 2013) and a point-of-care method providing more accurate identification of recent infection based on automated readings of multiple analytes (Pilcher, Kassanjee, Motta, Facente, Keating, et al., 2012). These may be difficult to implement in LMIC, however, due to shortages of funding and equipment.

Ecological studies analyze coexisting aggregate phenomena and attempt to draw connections between them. They depend for their persuasiveness on building a body of consistent evidence that points overwhelmingly in the same direction. In general, sound ecological studies rely on evaluating the effectiveness of interventions based on consistently sharp differences across multiple sites with and without the interventions of interest.

Data for ecological studies are largely drawn from program data (e.g. presence and timing of interventions)
and data on drug- and sex-related risk behaviors or aggregate biological measures (e.g. HIV prevalence or incidence) from survey and/or surveillance data.

Large, multi-country ecological reviews by Hurley and colleagues (Hurley, Jolley, & Kaldor, 1997), and updated by McDonald et al. (McDonald, Law, Kaldor, Hales, & Dore, 2003), which covered both developed countries and LMIC, revealed much more pronounced downward trends in HIV prevalence among PWID in cities with NSPs than in those without such programs. A survey of 67 cities/areas in 30 countries found that those in which PWID tended to use low-dead space syringes (that virtually eliminate the possibility of drug residue or blood being retained and transmitted to another person if the syringe is shared) had lower HIV prevalence among PWID than those cities/areas in which use of high dead-space syringes was the norm (Zule, Cross, Stover, & Pretorius, 2013).

Using a difference-in-difference approach, the China-Vietnam Cross-Border project evaluation found greater downward trends in HIV prevalence among PWID in the provinces with the interventions than in otherwise similar provinces without such interventions (Hammett, et al., 2012).

A study from Sichuan Province, China associated reductions in HIV risk behaviors among PWID over time with the concurrent increases in coverage of harm reduction interventions, thus taking ecological analysis to the level of individual behavior (Lau, Zhang, Zhang, Wang, Lau, et al., 2008). A paper on Dhaka, Bangladesh presented a counter-factual ecological analysis using mathematical modeling of surveillance data, which suggested that there would have been much higher HIV prevalence and incidence had an NSP not been implemented (Foss, Watts, Vickerman, Azim, Guiness, et al., 2006).

The conclusions of ecological studies are necessarily weaker if their data are limited in quantity or the identified differences are narrow. Such analyses may also fall victim to the “ecological fallacy” (Robinson, 1950) — for example, assuming that aggregate patterns, such as average declines in HIV prevalence in the presence of an intervention, reflect individual-level correlations between attendance at an NSP and averted HIV infection. This could be a problem for the comparisons of HIV prevalence in cities with and without NSP (Hurley, et al., 1997; McDonald et al., 2003) but the difficulty is eased in that these papers provide the range of prevalence reductions in cities with and without NSP so that the average reductions in prevalence should not mask lack of consistent correlation at the city level.

The distinguishing feature of quasi-experimental designs is that they involve non-randomized comparisons between sites or groups of individuals in different conditions. To be maximally powerful, the evaluation must compare sites or groups that are as similar as possible demographically, socio-economically and along other key dimensions except in the nature of the condition they experience.

Quasi-experimental evaluations may be conducted in a single site using “stepped-wedge” designs (in which all participants receive all conditions according to a staggered time sequence (Brown & Lilford, 2006), in multiple sites across which conditions are evaluated concurrently, or in crossover designs in which different sites experience different conditions at different times. Quasi-experimental studies may use behavioral and biological data on pre-post observations drawn from
cohort or cross-sectional surveys. All of these designs may be susceptible to problems of cross-contamination between sites.

This review found very few examples of quasi-experimental evaluations of HIV prevention interventions for PWID in LMIC and only one employing biological outcomes. A crossover evaluation in Ukraine compared a brief HTC intervention with a more intensive outreach intervention among PWID and found only small differences in behavioral outcomes (Booth, Lehman, Dvoryak, Brewster, & Sinitsyna, 2009). Another study in Ukraine compared network and individual interventions across five cites and found that the network approach was more effective in reducing the sharing of injection equipment among peer educators and network members (Booth, Lehman, Latkin, Dvoryak, Brewster, et al., 2011). One city in Sichuan Province, China with a comprehensive harm reduction intervention for PWID was compared to another city in the province with no interventions in a quasi-experimental evaluation that found significantly greater reductions in risk behaviors in the intervention city (Zhou, Zhang, Zhang, Kang, & Zhang, 2009). A study in Tehran compared a neighborhood with an NSP to one without an NSP and found that smaller proportions of PWID in the intervention neighborhood reported sharing injection equipment than in the comparison neighborhood (Zamani, Vazirian, Nassirimanesh, Razzaghi, Ono-Kihara, et al., 2010). A CDC-funded evaluation of HIV prevention programs for PWID in Northwest Vietnam compared trends in risk behaviors, HIV prevalence, and HIV incidence across three sites with differing levels of comprehensiveness in their HIV services for PWID and identified no clear associations between comprehensiveness and positive trends in outcome measures (Phan, Kieu, & Hammett, 2012).

In addition to issues of timing and duration of follow up periods and biases in recruitment and reporting, quasi-experimental designs may experience problems due to improper comparisons (Stimson & Power, 1992), cross-contamination between intervention and comparison sites, especially if populations are mobile and sites are close together, and changes in interventions over time that may be difficult to control for in data analyses. Oversight is needed to maximize the likelihood that interventions are delivered faithfully and consistently during the evaluation period, as long as this does not result in an unethical withholding of services for research purposes.

**True Experiments (Randomized Controlled Trials)**

Generally speaking, randomized controlled trials (RCT) are the strongest design for evaluation of interventions. Randomization normally occurs at the individual level but may also be done at the community level. The profound advantage of the RCT is that if random assignment is carried out properly it removes selection bias and counteracts unmeasured variation across intervention and control groups.

RCTs are virtually unknown in evaluations of HIV prevention interventions for PWID in LMIC. This review found only three. One is a multi-site (Thailand, China) trial (HPTN058) of sublingual suboxone for the treatment of heroin addiction in which the primary outcome is HIV incidence. Participants were randomized to treatment regimens of differing length and intensity. The study was discontinued because HIV incidence was too low to detect a difference between the arms but there was a significantly lower rate of heroin injection

“RCTs are virtually unknown in evaluations of HIV prevention interventions for PWID in LMIC. This review found only three...The primary barriers to implementing RCTs to evaluate interventions with people who use drugs and other marginalized groups are cost and ethical considerations.”
in the intervention arm over the short term (Metzger, Donnell, Jackson, Celentano, Aramrattana, et al., 2012). In a second RCT in Malaysia, comparing buprenorphine, naltrexone, and placebo, the buprenorphine and naltrexone groups both had significantly longer times to relapse than the placebo group but there was no difference between the two intervention arms. Moreover, there were reductions in HIV risk behaviors across the three groups but no significant differences among them (Schottenfeld, Chawarski, & Mazlan, 2008). Third, in northern Vietnam, Go and colleagues randomized PWID and their network members to HTC and a series of peer education sessions or to HTC only and found greater reductions in unprotected sex in the intervention arm and reductions in sharing of injection equipment across all arms (Go, et al., 2013).

The primary barriers to implementing RCTs to evaluate interventions with people who use drugs and other marginalized groups are cost and ethical considerations. RCTs are very expensive to implement properly and, particularly in LMIC, such expenditure, even if funding was available, might be considered inappropriate diversion of resources from underfunded but vitally necessary service provision. Moreover, in many instances it is considered ethically unacceptable to withhold services for research purposes. Very scrupulous human subject reviews are necessary to guard against ethical violations.

Opportunities for “natural experiments” may exist where differences in service provision across localities or individuals occur as if randomly assigned but in the absence of a formal experiment and in such cases comparisons may be more ethically acceptable.

In RCTs, using “intention to treat” analysis, in which all participants are included in the analysis of their assigned group whether they dropped out or not and regardless of how much actual treatment they received, may provide some remedy for the problems of attrition during the intervention, although it will not work properly in the absence of outcome data for all participants (Hollis & Campbell, 1999).

**Economic and systems evaluations**

*Cost-benefit and cost-effectiveness analyses* are relatively recent additions to the literature on evaluation of HIV prevention among PWID in LMIC, and they remain uncommon. Cost-effectiveness analyses (CEA) compare the costs of different interventions designed to prevent HIV infection and that have equivalent outcomes in terms of cases averted or reductions in incidence. Cost-benefit analyses (CBA), on the other hand, compare the costs of preventing an HIV infection to the measured or projected costs of care and treatment for an HIV infection or to a defined, non-monetary benefit that is assumed to be good for society at large (e.g. cases of HIV averted, quality-adjusted or disability-adjusted life years). This distinction in methodology is important particularly for the purposes of advocacy and policy making, since the lives and well-being of PWID may be undervalued by decision-makers relative to the lives of non-PWID.

The vast majority of evaluations employing CEA or CBA are from 2005 onwards. Vickerman et al. (Vickerman, Balakireva, Guinness, Artyukh, Semikop, et al., 2006) conducted one of the earliest CBAs of harm reduction in a LMIC, focusing on the one-year cost-benefit of NSP in Odessa, Ukraine. However, the methodology failed to cost in-kind donations (which appear to have been significant) to the single NGO-led intervention observed, and did not attempt to cost expenditures on HIV care and treatment services averted through prevention. Results from a single-city study in Svetlogorsk, Belarus (Kumaranayake, Vickerman, Walker, Samoshkin, Romantzov, et al., 2004) assessing the costs of HIV prevention interventions between 2000 and 2002
employed similar methodology, with similar results: a compelling case for the cost-benefit of harm reduction programming in terms of infections averted but the real costs averted in terms of care and treatment (which, as Vickerman notes, would have been extremely limited for PWID during this time [Vickerman, et al., 2006]) were not calculated.

In Bangladesh, Guinness et al. (Guinness, Vickerman, Quayyum, Foss, Watts, et al., 2009) provided a more thorough CEA of a single-NGO intervention using dynamic mathematical modeling. This methodology provided a detailed picture of costs averted in Dhaka over a three-year period through NSP and condom distribution, as well as treatment of abscesses and STI. Findings confirm that early intervention with harm reduction activities is more cost-effective than later introduction, but do not triangulate with alternative interventions in Dhaka or beyond to look at cost-effectiveness of differing levels of service.

Such comparisons were carried out by Khan and Khan (2011), who report on the costs of NSP in various cities in Pakistan, triangulated with bio-behavioral surveillance data. Notably, this evaluation employed real coverage data, not modeled or projected coverage. It is unique in the field in comparing intervention coverage and cost within different regions of the country. This methodology provides a rare glimpse into how well resources are being deployed in existing programming. In the Uighur Autonomous Region of China, Ni et al. (Ni, Fu, Chen, Hu, & Wheeler, 2012) provide another look at the effects of real coverage, assessing cost-effectiveness for 7 years of intervention, and modeling prospective lifetime cost-effectiveness.

Also from China, Li et al. (Li, Gilmour, Zhang, Koyanagi, & Shibuya, 2012) have taken a different approach, using compartmental mathematical modeling to propose prospective cost-effectiveness for four different interventions: ART, VCT, and harm reduction (considered here to be NSP and OST). Li’s analyses consider different combinations of these interventions, by both effectiveness in averting infections and cost-effectiveness, providing valuable data for decision-making in the age of combination prevention. Alistar et al. (Alistar, Owens, & Brandeau, 2011) present a similar compartmental modeling, assessing the cost-effectiveness of ART and OST (separate and in tandem) relative to health care costs and QALYs in Ukraine. In this evaluation, dual scenarios for OST coverage (“high OST coverage” and “low OST coverage”) were included to reflect realistic advances that might be made in scaling up OST under social and political pressure.

Wilson et al. (Wilson, Zhang, Kerr, Uuskula, Kwon, et al., 2012) present a new wave of CEA from 8 countries in Eastern Europe and Central Asia, considering the burden of both HIV and hepatitis C virus (HCV) that can be averted by reducing unsafe injecting practices. Utilizing a mathematical transmission model, this evaluation examines retrospectively (2005-2010) cases averted, as well as modeling future cases averted due to interventions during that time period. Similar methodology was employed in China by Zhang et al. (Zhang, Yap, Xun, Wu, & Wilson, 2011), taking advantage of highly centralized data in an environment with relatively low funding of programs by international donors. The result is a clear picture of return on investment for the Chinese government. However, these evaluations consider only the cost-effectiveness of NSP, which becomes increasingly confounded as OST becomes more widely available and ART treatment initiation standards are changed (Li, et al., 2012).

A recent World Bank study (Dutta et al., 2013) modeled outcomes based on different combination and scale-up scenarios for HCT, ART, NSP, and OST in Ukraine, Pakistan, Thailand, and Kenya and reached the general conclusion that expansion of these interventions would be cost-effective in averting HIV infections.

Regardless of increasingly sophisticated modeling for both retrospective and prospective CEA, evaluations can only be as good as the quality of their data. Compiling accurate costing data in many LMIC can be challenging, suffering from the same selection bias seen in observational studies — that is, the cost of...
reaching the current level of coverage among PWID may not accurately reflect the costs of scaling up to cover harderto-reach individuals. It may also be difficult to cost “critical enablers” — peripheral programming that supports behavior change outside of interventions such as NSP and OST (UNAIDS, 2012) encouraging greater use of or adherence to those interventions — particularly because core programming and critical enablers may be funded through different mechanisms (e.g. domestic funding versus international donor funding).

Nevertheless, cost-linked evaluations of these interventions become increasingly important as global donors shift towards an investment-minded framework for distribution of funds (Schwartlander, Stover, Hallett, Atun, Avila, et al., 2011). Multi-site evaluations, looking at the relative cost-effectiveness of different levels and delivery methods of interventions, such as presented by Li et al. (2012), are likely to become more critical for making investment cases.

The shortcomings of the current body of cost-effectiveness evaluations outlined above highlight the vertical nature of programming for PWID in many settings. Core PWID prevention interventions, including NSP, OST and behavior change communication are often donor-funded and implemented at the project level, rather than as part of ongoing programs integrated into the broader health system. As such, perspective on long-term sustainability and stability of PWID programming is limited.

A limited number of donor-driven interventions have attempted to strengthen HIV prevention programs for PWID using a health systems approach. In Indonesia, Australia’s DFAT has directed funds towards increasing leadership and governance on HIV issues related to PWID, and improved service delivery for PWID (Hind, Lowe, Kharisma, & Woodland, 2011). In the Central Asian Republics, USAID-funded efforts to strengthen human resources for health, service delivery, health information systems, and leadership and governance are ongoing (Quality Health Project in the Central Asian Republics, 2011). However, no comprehensive impact evaluations of these projects or their target systems have been conducted.

Likewise, no comprehensive evaluations have assessed the effectiveness or cost-effectiveness of providing core services for PWID (NSP, OST, HTC, ART) inside the government health system versus through NGOs or stand-alone donor-funded projects. Further, no evaluations have examined how the quality or function of critical health systems elements affects prevention of HIV among PWID at the national level, or, conversely, how HIV prevalence and incidence among PWID burdens the health system outside of vertical HIV programs.

As governments and international donors alike assume a rhetoric of investment frameworks (Ni, et al., 2012) and increased focus on health systems strengthening (Biesma, Brughia, Marmer, Walsh, Spicer, et al., 2009), providing such comprehensive data on the role of health systems in HIV prevention for PWID will become increasingly important.

In summary, the most common designs for assessing HIV prevention interventions among PWID in LMIC found in this review are those based on routine process monitoring data (which does not really constitute evaluation) and outcome evaluations comparing pre-post observations of risk behaviors. The major gaps in evaluation occur in rigorous outcome and impact studies employing biological measures and appropriate comparisons, as well as in attempts to disentangle the effects of the elements of combination prevention.”
effects of the elements of combination prevention. It is noteworthy that these are essentially the same gaps identified six years ago by the Institute of Medicine review of the evidence (Institute of Medicine, 2007). Moreover, there have only been a limited number of cost-effectiveness and cost-benefit studies and no assessments of either interventions’ contributions to, or effects on, larger health systems.

RECOMMENDATIONS

To fill the gaps in rigorous evaluation, advance the state of the art, and improve the evidence available to inform policy and practice in LMIC, funders and evaluators should consider the following recommendations for the design of evaluations and the interpretation and use of evaluation data. Particularly in resource constrained settings, the costs of evaluation should be weighed against the possible benefits in terms of improved HIV prevention programs. Evaluations should use the least expensive data sources, the smallest sample sizes, the shortest follow up periods, and the fewest comparisons that can reveal the needed results with acceptable precision.

- Design evaluations based on the political and ideological context. Evaluation should be both policy-driven and method-driven. Evaluators and funders should consider the amount of further evaluation needed to achieve desired policy and programmatic goals. The availability of better surveillance data would facilitate use of more economical observational evaluation designs and development of better estimates of intervention coverage (Mathers, Degenhardt, Ali, Wiessing, Hickman et al., 2010; Mathers, Cook, & Degenhardt, 2010). The extent to which governments are willing to accept international as opposed to domestic data for decision-making should also be considered. Other factors include the availability of funding and the appropriateness of applying it to evaluation, the availability of and capability to produce the required data, the political and ideological factors affecting acceptance of HIV prevention interventions for PWID, and the effects of such factors on the level and rigor of evaluation needed (Stimson, et al., 2005).
- Although local contexts vary, in general, there is a need for more rigorous evaluation with the focus shifting from process measures to outcomes and from demonstration of effectiveness alone to cost-effectiveness and cost-benefit analysis (Institute of Medicine, 2013).
- To meet the needs of program planners and implementers, evaluation designs should provide data to inform feedback loops and midcourse corrections in interventions (Stimson & Power, 1992). However, such midcourse corrections may introduce complications for ongoing evaluation of programs.
- In investigating causation, evaluations should consider the effects of natural epidemic progression, demographic changes, and programmatic factors in the interventions under study. HIV epidemics driven by injection drug use may naturally plateau or decline regardless of the presence or absence of interventions. In Amsterdam, de Vos and colleagues (2013) concluded that observed reductions in HIV and HCV incidence were as much the result of naturally occurring changes in the PWID population as of the city’s long-standing harm reduction programs. However, a commentary by Vickerman and Hickman (2013) emphasized the complexities of distinguishing among the effects of epidemic progression, demographic changes, and implemented interventions, particularly when modeling is based on incomplete data. Several studies in Vancouver and Montreal in the 1990s suggested that attendance at NSPs was associated with higher risk of HIV infection among PWID, findings that fuelled opposition to NSP. However, subsequent investigation revealed that this association probably resulted from the extremely high-risk population drawn to the NSP and service deficiencies of the programs (Bastos & Strathdee, 2000).
• **Incorporate assessment of and recommendations for intervention sustainability, replicability, and scalability,** which will be particularly important in countries with resource constraints.

• Since HIV prevention programs for drug users rarely involve only one intervention, it is important to **devise better methods to evaluate combination prevention** in order to understand the relative effectiveness of components both singly and in combination (Degenhardt, Mathers, Vickerman, Rhodes, Latkin et al., 2010).

• **Conduct more rigorous, multi-component cost-effectiveness evaluations** to complement existing data on cost-benefit and cost-effectiveness of single interventions (e.g. NSP only) and to build the most compelling investment cases. A small number of cost-effectiveness investigations have used compartmental modeling for costing intervention components (Li, et al., 2012; Alistar, Owens & Brandeau, 2011) and combinations of interventions (Dutta, Wirtz, Stanciole, Oelrichs, Semini et al., 2013). Such information could inform choices required in resource-constrained settings.

• **Assess the relationships between HIV prevention interventions and health systems** to increase understanding of the effects of HIV infection and HIV prevention interventions on systems and, conversely, the effects of integrated services and other health system enhancements on the funding, implementation, and outcomes of interventions.

• Importantly, evaluations **should include plans to use the evidence to advocate for improvements in laws, policies, and programs.** Experience shows that even compelling evidence often cannot by itself achieve needed changes in policies or programs. This does not lessen the need to continue improving evaluations and evidence but it does underline the concurrent need for advocacy as well as sound science. Evaluators and advocates should design studies to produce the types of evidence most likely to influence decision makers. In some cases, this may be evidence of infections averted, cost effectiveness, and cost savings. Strong evidence that harm reduction interventions do not increase the prevalence or severity of drug use or other negative outcomes may also be influential. Wherever there is strong political or ideological opposition to harm reduction, careful assessment of the types of evidence most likely to counteract this opposition should drive the design of evaluations and analyses of data.

After three decades of accumulating evidence, harm reduction interventions for HIV prevention among PWID remain under-utilized, particularly in LMIC. “Getting to Zero” new infections among PWID will require not only better and more compelling evidence to end, finally and completely, the scientific debate on harm reduction but also the more effective deployment of that evidence to advocate for and achieve universal adoption of harm reduction interventions.

**ACKNOWLEDGEMENTS**

The authors gratefully acknowledge the funding support of Abt Associates’ International Health Division, and especially of Division Director Dr. Diana Silimperi and former Deputy Director Dr. John Palen, in the preparation of this paper. We also thank William Rhodes and Ayman Abdelmohsen of Abt Associates for the valuable comments they provided as reviewers of the manuscript for Abt’s Thought Leadership Paper Series.
REFERENCES


For more information, please contact:

Theodore M. Hammett, Ph.D.
Vice President/Principal Associate
International Health Division
Abt Associates
55 Wheeler Street
Cambridge, Massachusetts 02138 USA
+1-617-520-2584
ted_hammett@abtassoc.com

Danielle S. Parsons-Smith, M.P.H.
Consultant, Washington, D.C.
daniellesparsons@gmail.com

www.abtassociates.com

© 2014 Abt Associates