**PROGRAM ON THE GLOBAL DEMOGRAPHY OF AGING**

## Working Paper Series

# Health systems and HIV treatment in sub-Saharan Africa: Matching intervention and program evaluation strategies

Till Bärnighausen, David E. Bloom, Salal Humair

January 2012

PGDA Working Paper No. 86
http://www.hsph.harvard.edu/pgda/working.htm

 The Program on the Global Demography of Aging receives funding from the National Institute on Aging, Grant No. 1 P30 AG024409-06.

# Health systems and HIV treatment in sub-Saharan Africa: Matching intervention and program evaluation strategies

Till Bärnighausen[1,2], David E. Bloom[1], Salal Humair[1,3]

[1]Harvard School of Public Health, Department of Global Health and Population, Boston, USA

[2]Africa Centre for Health and Population Studies, University of KwaZulu-Natal, South Africa, Mtubatuba, South Africa

[3]School of Science and Engineering, Lahore University of Management Sciences, Lahore, Pakistan

Corresponding author:

Till Bärnighausen

655 Huntington Avenue

02115 Boston, MA

USA

Tel: +1 617 379 0372

Fax: +1 617 432 6733

Key words: Impact evaluation, health systems, HIV, antiretroviral treatment, Africa

Word count: 4,852

**Abstract**

*Objectives*

International donors financing the delivery of antiretroviral treatment (ART) in developing countries have recently emphasized their commitment to rigorous evaluation of ART impact on population health. In the same time frame but different contexts, they have announced that they will shift funding from vertically-structured (i.e., disease-specific) interventions to horizontally-structured interventions (i.e., staff, systems and infrastructure that can deliver care for many diseases). We analyze likely effects of the latter shift on the feasibility of impact evaluation.

*Methods*

We examine the effect of the shift in intervention strategy on (i) outcome measurement, (ii) cost measurement, (iii) study-design options, and the (iv) technical and (v) political feasibility of program evaluation.

*Results*

As intervention structure changes from vertical to horizontal, outcome and cost measurement are likely to become more difficult (because the number of relevant outcomes and costs increases and the sources holding data on these measures become more diverse); study design options become more limited (because it is often impossible to identify a rigorously defined counterfactual in horizontal interventions); the technical feasibility of interventions is reduced (because lag times between intervention and impact increase in length and effect mediating and modifying factors increase in number); and political feasibility of evaluation is decreased (because national policymakers may be reluctant to support the evaluation).

*Conclusions*

In the choice of intervention strategy, policymakers need to consider the effect of intervention strategy on impact evaluation. Methodological studies are needed to identify the best approaches to evaluate the population health impact of horizontal interventions.

**Key messages**

- Many recent donor initiatives in global health have been vertically structured, but prominent donors have announced that they will increasingly fund horizontal interventions.
- At the same time, donors have recently declared that they are committed to rigorous evaluations of the impact of their investments on population health.
- The shift towards more horizontally structured interventions, however, will increase the difficulty in conducting impact evaluation.
- The shift will likely lead to additional problems in measuring exposure, costs and outcomes, and reduce the number of options for evaluation study design.

## Background

The last decade has seen a dramatic scale-up in the delivery of antiretroviral treatment (ART) to HIV-infected people in sub-Saharan Africa (SSA). At the end of 2009, almost four million people in the region were receiving ART, while five years earlier the figure was less than one million.[1] This achievement is especially notable because it occurred predominantly in countries with low income and relatively rudimentary health system infrastructures. Much credit goes to international donors that provided massive financial and programmatic support. Many of the donor initiatives, such as the U.S. President's Emergency Plan for AIDS Relief (PEPFAR) and the Global Fund to Fight AIDS, Tuberculosis, and Malaria (GFATM) have been vertically-structured,[2-4] i.e., they are focused on treating HIV infection and its accompanying diseases. The disease specificity of these programs has usually implied that they are organized separately from the general health system, employing their own front-line health and management staff, delivering treatment in their own facilities, operating their own supply chains, and using their own monitoring and evaluation (M&E) systems.

While the scale-up of ART delivery in SSA is an extraordinary achievement, the number of people now receiving ART constitute only about 40% of those who need ART based on WHO treatment eligibility criteria.[1] At roughly the same time as the United Nations (UN) has called on member states to work towards achieving universal ART coverage by 2015,[5] international donors have announced two major shifts in their involvement in ART delivery. First, several prominent donors have announced that they will increasingly fund horizontal interventions,[6] i.e., shifting resources from vertical programs into staff, systems, and infrastructure that can deliver care for many diseases. Reasons for this shift include the belief that ART could be delivered more efficiently in the general health system if the system's capacity were improved; a recognition that HIV patients commonly suffer from other diseases that are usually treated in the general health care system; and worries that vertical interventions for some priority diseases divert human and financial resources away from the treatment of other diseases that are also major causes of death and disability in developing countries.[7]

The second shift involves key donors declaring a commitment to rigorous scientific evaluations of the impact of their investments on population health[8] – a shift from assessing programs in terms of process and output indicators such as the number of people treated, to making

assessments based on outcome indicators such as life expectancy among HIV-infected individuals. Reasons for this second shift include the recognition that large investments in health interventions need to be justified by demonstrable improvements in population health and that "well-designed and empirically grounded research and evaluation of PEPFAR should promote improved performance, accountability, informed decision-making, and lessons from experience."[9]

Below, we describe the two shifts in more detail. We then analyze the likely effects of the first shift (from more vertically- to more horizontally-structured health interventions) on the second shift, in particular, on (i) outcome measurement, (ii) cost measurement, (iii) options for the design of impact evaluation studies, and the (iv) technical and (v) political feasibility of program evaluation.

## Two shifts in the delivery of HIV treatment and care

*From more vertically structured to more horizontally structured HIV treatment and care delivery*

Although PEPFAR and GFATM have achieved their success in bringing HIV treatment and care to scale in SSA through vertically structured programs, the two donor organizations have recently shifted funding towards more horizontally-structured interventions. As PEPFAR declares, it "has not had a strategic vision or plan to incorporate a health systems lens into its programming. In the first phase of PEPFAR [2004-2009], health systems activities were largely ad hoc, varied across countries, and did not always factor in an intervention's impact on the country's broader health system."[10] In contrast, in the current phase [2010-2014], PEPFAR "emphasizes the incorporation of health systems strengthening goals into its prevention, care and treatment portfolios."[11] For instance, PEPFAR has committed to train 140,000 "health care workers, managers, administrators, health economists, and other civil service employees critical to all functions of the health system."[12] Other examples of PEPFAR investment in horizontal interventions include improving "the general supply chain, procurement, and forecasting systems", "supervisory and planning skills" of healthcare management staff, and "referral systems".[13]

The movement toward horizontal structures is also articulated in the mission of the U.S. Global Health Initiative (GHI), the new umbrella organization for U.S. global health engagements, which includes "sustainability through health systems strengthening" along with interventions for specific diseases as one of its "core objectives".[7] The GHI also prioritizes not only HIV, malaria and tuberculosis interventions but also support for healthcare that is traditionally considered to be part of general primary care, such as family planning and reproductive, maternal, and child healthcare.[14] Similarly, GFATM states that an "effectively performing health system is key to improving the population's health status, providing protection against health-related financial risks and enhancing the health sector's responsiveness to customers' needs". For its latest funding round, GFATM encourages countries to apply for "cross-cutting HSS [health systems strengthening] interventions, which affect more than one of the three diseases (HIV, tuberculosis, malaria)" and "may address broader health MDGs [millennium development goals], by for example contributing to maternal and child health".[15] As examples of such "cross-cutting" interventions, GFATM lists "upgrading primary health care facilities, strengthening planning and policymaking capacity of the Ministry of Health, [and] improving the national health management information system".[15]

*From M&E to impact evaluation*

PEPFAR aims "to help save the lives of those suffering from HIV/AIDS around the world;"[16] GFATM states that its mission is "[i]nvesting the world's money to save lives."[17] Despite these declarations, the two organizations did not emphasize evaluations of intervention impact on population health outcomes in the first years of their operations. Rather, they focused primarily on process and output indicators. For instance, PEPFAR regularly reported the number of persons it "directly supported with life-saving antiretroviral treatment"[18] based on its M&E system, which requires recipients to record these numbers and to report them to PEPFAR country offices. The US Institute of Medicine's 2007 report, *PEPFAR Implementation: Progress and Promise,* did not estimate PEPFAR's effect on population health outcomes, but assessed PEPFAR strategy, resources, management, and the "number of people receiving ART [antiretroviral treatment] supported by PEPFAR."[19]

At the national level PEPFAR results are still largely assessed in terms of process and output indicators. But in the last few years both organizations have attempted to estimate the impact of ART in terms of life-years saved at the global level, for example, in the 2009 PEPFAR *Report to Congress*[18] and in the 2010 GFATM report *Innovation and Impact*.[20] For instance, the GFATM estimated that the programs it supported had saved a total of 4.9 million lives through the end of 2009. Most recently, GFATM has raised the total estimate to 6.5 million lives through the end of 2010. [21] The methodology for arriving at these estimates has not been documented extensively, but a brief outline of the approach taken appeared in the Global Fund's 2007 *Partners in Impact* report.[22] For estimates of HIV-related mortality averted through ART, two scenarios are compared, one with a survival curve assumed for patients on ART, and the other with an "untreated" survival curve of patients needing but not receiving ART. The survival assumptions follow those used by UNAIDS in their routine estimates on the global HIV epidemic.[23] In addition to the GFATM annual reports, there has also been an impact assessment published as a journal article by Komatsu and colleagues.[24] The methods used in the article differed only slightly from those used in the official GFATM reports.

While the causal link between the measured output (individuals receiving ART) and the modelled outcome (life-years saved) is relatively direct, these models require a number of assumptions that may not hold, leading to biased results. First, the models assume an ART effect on the survival of each patient receiving the treatment, which is derived from a comparison of AIDS mortality observed in persons needing but not receiving ART in some cohort studies with the mortality among persons needing and receiving ART in other studies. However, the mortality in the cohort studies of persons needing but not receiving ART may not be a good counterfactual for the mortality the persons needing and receiving ART would have experienced had they not received the treatment. Access to ART is not equal for all people eligible for ART[25] and thus actual ART patients are likely to differ from potential patients in many characteristics that affect mortality, such as sex, age, income, nutritional status, and retention and adherence behaviours. Second, the models assume that none of the life-years saved with PEPFAR or GFATM support would have been realized in the absence of the support. However, this counterfactual assumption will not be met if PEPFAR and GFATM caused changes – either increases or decreases – in national spending on ART, which would not have occurred if the organizations had not invested in ART.[26-29] Third, the models assume that ART delivery has no

effect on the capacity of the general health system to provide HIV-unrelated health services. However, it is plausible that the substantial investment in ART has either positively or negatively affected general health systems.[26-31] For instance, ART provision requires human and physical resources, which may have been drawn away from the delivery of other health care,[29,32] leading to loss of life-years, which would not have occurred without the investment in ART. On the other hand, it is also possible that infrastructure that has been built to provide ART, such as supply chains for medicines, has improved the delivery of other health care, and that the HIV-related initiatives have led to increased political commitment of governments to improve public health in general, improving health outcomes of patients who do not suffer from HIV.[29]

For these reasons, existing model-based impact estimates may not reflect the true causal effect of PEPFAR and GFATM – a fact that lends support to the need for an increased focus on empirical impact evaluations that base causal effect estimation on "a rigorously defined counterfactual to control for factors other than the intervention that might account for the observed change".[8]

The initial shift from output monitoring to mathematical models of intervention impact has now been followed by a shift from modeling to measurement of impact in empirical evaluation studies with rigorously defined counterfactuals. This new shift has been attributed to a change in the conception of large-scale donor support for ART from an "emergency response" to a foundation for long-term provision of the life-saving treatment to millions of people in sub-Saharan Africa for their entire lives.[6] As Padian and colleagues write:[33]

> "In the first phase of PEPFAR, these activities were appropriately carried out in an emergency fashion with the goal of using available interventions to reduce mortality and alleviate suffering from HIV disease as quickly and effectively as possible. … Commensurate with the emergency response, however, state-of-the-art monitoring, evaluation, and research methodologies were not fully integrated or systematically performed. In the second phase of PEPFAR, characterized by an increased emphasis on sustainability, programs must demonstrate value and impact to be prioritized within complex and resource-constrained environments. In this context, there is a greater demand to causally attribute outcomes to programs."

**Impact evaluation under alternative intervention strategies**

The shift toward rigorous, empirical evaluations of the impact of PEPFAR and GFATM interventions is critical for generating understanding of the true causal effects of the two initiatives. The feasibility of such studies, however, will differ depending on whether the intervention is structured vertically or horizontally. Figure 1 depicts in the abstract how different aspects of the shift in intervention structure from vertical to horizontal affect the feasibility of impact evaluation; Figure 2 provides an example. Below, we explain how intervention structure affects health outcomes measurement, cost measurement, study design options, and the technical and political feasibility of evaluations.

*Health outcomes measurement*

Vertical interventions are intended to affect narrow sets of health outcomes (e.g., mortality and morbidity in HIV-infected individuals), while horizontal interventions are by definition intended to affect a wide range of health outcomes. Horizontal interventions that are financed by organizations with disease-specific mandates, such as PEPFAR or GFATM, thus raise an issue about the outcomes to observe for impact evaluation. Full evaluation of horizontal interventions is more difficult than full evaluation of vertical interventions because wider populations with a larger set of morbidities and causes of mortality have to be observed; but narrow evaluation of horizontal interventions for only HIV-infected persons is not useful for many decision-making purposes because potentially large components of the total effect are neglected. For instance, two health worker interventions may have comparable effects on morbidity and mortality in HIV-infected individuals, but one may save many more life-years in HIV-uninfected populations than the other. We expect policymakers to prefer the intervention that saves more life-years in HIV-uninfected populations and to thus require a comprehensive evaluation of intervention outcomes.

*Cost measurement*

A further complication in the evaluation of horizontal interventions arises when costs are considered in addition to outcomes. A common metric used to present the impact of the large-scale global health interventions in recent years has been "value for money", i.e., the size of a benefit or effect per financial investment in an intervention.[34 35] For instance, "[t]he Global Fund is committed to measuring and demonstrating the relationship between its financial investments

and their outcomes. Value for money measurement will also provide countries with tools to improve program efficiency, which they can use in resource mobilization."[35]

A vertical program is commonly financed exclusively by one agency. Where multiple funders contribute to a vertical program, the contributions of the different agencies are often clearly visible to all funders. For instance, in South Africa, the government and PEPFAR both contribute to the funding of the public-sector ART program.[36] This joint effort is well coordinated and both parties know the types and quantities of the resources the other party contributes and can easily obtain information on the financial outlays for these resources. In contrast, in some types of horizontal programs, it may be much more difficult for the primary funder to obtain realistic estimates of the financial contributions of other agencies, because these programs will likely require more diverse sets of inputs and because these inputs will not be utilized exclusively by the horizontal program. For instance, programs improving the supply chains of medicines to primary care clinics will likely require support by health workers in central pharmacies and by the health workers in the primary care clinics receiving the medicines. However, these health workers will only spend some portion of their time supporting the supply chain intervention. This portion is unlikely to be known without additional research effort, such as time-motion studies or health worker interviews. The shift from vertical to horizontal interventions is thus likely to imply substantially increased difficulty in measuring an intervention's cost-effectiveness or "value for money".

*Study designs*

As we note above, the shift to more rigorous evaluations of intervention impact requires studies to be based on "a rigorously defined counterfactual",[8] i.e., a study design that minimizes threats to validity in the estimation of the outcome that would have occurred had those who received the intervention not received it. Randomized controlled experiments are often viewed as the gold-standard of impact evaluation, because the random assignment of individuals or other units of observation to an intervention group and to a control group not receiving the intervention ensures that the two groups are as similar as possible regarding all relevant factors that can affect the outcome, other than the intervention itself. Given sufficiently large samples, randomization will adequately control confounding and allow unbiased effect size estimation. Where randomized controlled experiments are not possible, observational approaches can be employed to estimate

effect size using rigorously defined counterfactuals.[26] These approaches include difference-in-difference estimation (where the counterfactual is the difference in an outcome before and after the time of an intervention in individuals living in communities that did not receive the intervention),[37] interrupted time series (where the counterfactual is the evolution of an outcome over time before the intervention took place),[38] regression discontinuity (where the counterfactual is the outcome in individuals whose score on a certain variable falls below the threshold above which individuals receive the intervention),[39] or instrumental-variable approaches (where the counterfactual resembles that of a randomized experiment because "the forces of nature or government policy" have randomly determined treatment assignment or intensity).[40]

In general, vertical interventions are more likely than horizontal interventions to allow researchers control over the assignment of individuals or communities to intervention and control groups. For instance, randomized controlled experiments of vertical ART programs would be feasible, if researchers could randomize the time when a particular geographically bound community or cluster receives an intervention. An example is stepped-wedge cluster-randomized experiments[41] where those locations that receive the intervention later serve as counterfactuals for those that received them earlier. Stepped implementation of large-scale ART programs is common,[36] and it is technically feasible to randomize the implementation scale-up across space and time (although this approach has been uncommon, possibly because of political resistance to randomization of intervention implementation or simple failure to consider such an approach before starting large-scale ART programs). In contrast, many horizontal interventions cannot be randomized because they will always affect the entire health system of a country. For instance, changes in the legislation regulating a country's health workforce, the establishment of national institutions to systematically assess evidence on healthcare interventions, or market interventions to reduce drug prices, cannot be randomly assigned to individuals or sub-national areas. Such horizontal interventions that function through central mechanisms and affect an entire country equally are also difficult to evaluate using observational approaches, because they require cross-national data for a rigorous counterfactual.

Horizontal interventions that do not equally affect the entire health system of a country (at least not immediately) can theoretically be randomized[42] or evaluated, exploiting "naturally

occurring" differences in intervention expansion over time and across space.[43] However, such horizontal interventions will often be difficult to evaluate in practice because neither the exposure nor the outcome of interest are easy to assess. For instance, to evaluate the impact of a horizontal program to train health workers in a country, it would be necessary to keep track of all health workers trained by the program as they find – and leave – employment in different areas of the country and then to establish data collection infrastructures in those areas to measures the health outcomes of interest. By contrast, in the evaluation of vertical interventions, the exposure is usually the vertical intervention itself, such as ART delivery, and the outcome can be assessed directly through the vertical structures.

In addition, it may be more difficult to identify meaningful observational approaches to determine causal relationships for horizontal than for vertical interventions. For instance, discontinuities in treatment assignment, which are necessary for regression discontinuity estimation of intervention effects, will often exist for vertical interventions, because treatment indicators for particular diseases are commonly determined by applying a fixed threshold to a continuously measured biological variable. A case in point is ART, which is initiated when a patient's CD4 count falls below a certain threshold (e.g., 350 CD4 cells/µl under the latest WHO ART treatment guidelines, up from 200 CD4 cells/µl prior to mid-2010). This discontinuity (and changes in this discontinuity) in the relationship between CD4 count and ART allows estimation of the effect of ART on outcomes such as mortality, employment, or sexual behaviour, by comparing these outcomes in patients with CD4 counts just above and just below the threshold.

It is unlikely that such discontinuous treatment assignment rules can be identified for most horizontal interventions, such as health worker training, supply chain improvements, or the establishment of new procurement and forecasting systems, because the processes through which such interventions are placed and reach patients are influenced by the complex interplay of many factors that are poorly understood and cannot be summarized in a simple discontinuity on a single variable. Similarly, it may be more difficult to use instrumental-variable approaches to estimate the effectiveness of horizontal as compared to vertical interventions. One reason for this increased difficulty is that horizontal interventions are more likely to consist of multiple important components, the presence and intensity of which can vary across interventions. For

instance, an intervention aiming to reduce drug stock-outs in primary care clinics may consist of a range of separate components intervening in the forecasting system, the procurement system, and the supply chain from a central pharmacy to the clinic. To estimate the overall effectiveness of a horizontal intervention, researchers may thus face the difficult task of identifying and measuring separate instruments for each intervention component. Overall, it seems likely that the shift from vertical to horizontal intervention strategies will decrease the options for evaluation study designs that are appropriate and feasible to examine intervention impact.

*Technical feasibility of evaluation*

Another aspect of interventions – the time lag between intervention start and the first possible observation of intervention impact on outcomes – both affects the feasibility of evaluation and differs between vertical and horizontal interventions. Vertical interventions will commonly generate health impacts more quickly than horizontal ones. This occurs because horizontal interventions to benefit HIV-infected populations generally need to be in place before vertical ART delivery can begin. For instance, many types of health workers, such as doctors and nurses, need to be educated and trained for several years before they can deliver ART. The establishment of an electronic patient record system may require procurement of laptops, development of software, health worker training, and field testing before it can contribute to the quality or efficiency of ART delivery and improve health outcomes in patients. The longer the time lags between intervention and outcomes, the more complicated and costly it will be to evaluate the intervention. Evaluation planning will need to precede the implementation of the evaluation by longer times and the timing of baseline and post-intervention measurements of exposures and outcomes becomes more difficult.

Even if horizontal interventions can be rigorously evaluated, we may learn less from the evaluation results than in the case of the evaluation of vertical interventions, because the horizontal interventions are commonly mediated through longer causal chains than vertical ones, and the number of factors that can modify intervention effects will likely increase. Take, for instance, a training program to increase the capacity of district health managers to plan the delivery of health programs. For this horizontal intervention to have an effect on population health, it will be necessary for district health managers to attend the training, successfully acquire

new skills in the training, and be willing to use their new skills. The impact on health outcomes will then further depend on the capacity of the manager to effect changes in the actual delivery of ART. It is this actual delivery, on the other hand, which is usually the starting point for the evaluation of vertical ART programs. Thus, the mediating steps from the district health worker intervention to a health impact are many more than those from the ART program to health impact, and contextual factors influencing district health managers' capacity to use newly acquired skills will likely increase the heterogeneity of effects across settings.

Differences in mediating factors will lead to heterogeneity in estimated impacts across settings. The larger the number of mediating factors between the intervention and the outcome, the more resources will be required to either observe or control for all mediating factors. As the number of mediating factors will commonly increase as the intervention structure changes from vertical to horizontal, it is likely that impact evaluation that can shed light on the effects of programs across settings or populations will be more complex and require more resources for horizontal than for vertical interventions.

*Political feasibility of evaluation*

In addition to the technical feasibility of evaluation, intervention structure is also likely to affect the political feasibility of evaluation. Horizontal interventions, such as improvements in supply chains, patient management systems, or health worker training, are likely to take place in close interaction with or to be integrated into the general public-sector health system. Hence, researchers will likely need the cooperation of the front-line staff working in the general health system to measure exposure and outcomes, or to assess possible mediating factors. Unlike in vertical ART programs, where front-line staff will often be employed or otherwise supported by initiatives such as PEPFAR and GFATM, which fund the program, front-line staff in the general health system are likely to be public-sector employees. As such, they may be less likely to participate in the evaluation of interventions that are funded by initiatives that are not part of the public sector.

Moreover, national and sub-national politicians may be more interested in the evaluation of horizontal interventions supporting the general health system than in the evaluation of vertical programs, because they are directly responsible for the functioning of the former but not of the

latter. This increased interest in the evaluation could imply that they more strongly support the evaluation process, because they see the value of the evaluation results for their own work. However, it could also imply that they will not be supportive of evaluations by outside agencies, because they will not be able to control the dissemination of evaluation results that could be embarrassing to them or because they feel that it is their responsibility to initiate and fund evaluations of the general health system. Since – unlike in many evaluations of vertical programs – the cooperation of policymakers will be needed to conduct evaluation of interventions that are related to the general health system, it is possible that such evaluations will not be feasible for political reasons, impeding the generation of evidence on their impact.

**Discussion**

We have described possible implications of the shift from vertically- to horizontally-structured health interventions for outcome measurement, cost measurement, options for impact evaluation study designs, and the technical and political feasibility of program evaluation. In general, we argue that the shift will increase the difficulty in conducting impact evaluation, because it will lead to larger sets of relevant outcomes, more diversity of sources of intervention co-funding and use of inputs whose value is difficult to establish, fewer options for evaluation study designs, increased difficulty in forecasting when and where exposure to the intervention and outcomes will occur, longer lag times between intervention start and impact on outcomes, larger numbers of mediating factors in the causal chain from intervention to outcome, and reduced motivation of political leaders and front-line staff to support impact evaluation. Donors' expressions of commitment to rigorous evaluation – and to the use of evaluation results in adjusting interventions – conflict with their actions shifting investments away from vertical HIV interventions to interventions benefitting the general health system. Of course, such a conflict does not necessarily imply that donors are not credibly committed to rigorous evaluation. The two shifts have plausibly occurred for independent reasons and may not have been considered jointly. The vertical HIV interventions of the past were not well evaluated because the commitment to research proving intervention impact in different contexts was lacking. As a result of the two conflicting shifts in priorities, we may yet again fail to learn about intervention impact, despite donors' commitment to rigorous impact evaluation, because of the difficulty in evaluating impact of interventions affecting the general health system.

This is not to say that rigorous evaluations of horizontal interventions will be impossible. Examples of rigorous evaluations of horizontal health systems interventions exist.[42 44 45] However, it is likely that such evaluations will require substantially more resources than evaluations of vertical programs and that it will take longer before evaluations results will become available, implying that one of the purposes of evaluation – timely adjustments of interventions – will be harder to meet. Delays in results on impact will be risky because interventions that have little impact, or cause harm, will continue for longer before corrective action can be taken. Our accounts of effects of the shift in intervention structure on our ability to evaluate the impact of interventions are based on a highly stylized distinction between vertical and horizontal interventions and are meant as ideal-type descriptions rather than as occurring necessarily in all comparisons of particular vertical and horizontal interventions. In fact, it is not difficult to identify a few exceptions from many of the general comparisons. For instance, the provision of bus transport to improve patient access to care may be easier to evaluate than many vertical programs (such as vertical ART programs), because this particular horizontal intervention has few inputs and is likely to improve health outcomes relatively quickly.

Policymakers and funders thus need to consider carefully the feasibility of impact evaluation of each particular intervention. In addition, it is likely that new methods need to be developed and tested to allow timely, rigorous evaluation of horizontal interventions intended to benefit primarily HIV-infected populations. Our article describes a general framework that can help guide these considerations and serve as a starting point for developing new methods. None of the factors we present represents a decisive argument against a change from vertical to horizontal interventions, but they do merit consideration in the choice of broad intervention strategy.

**Acknowledgments**

Larry Rosenberg provided helpful comments.

**References**

1. WHO/UNAIDS/UNICEF. *Towards universal access: scaling up priority HIV/AIDS interventions in the health sector*. Geneva: WHO, 2010.

2. Stringer JS, Zulu I, Levy J, et al. Rapid scale-up of antiretroviral therapy at primary care sites in Zambia: feasibility and early outcomes. *JAMA* 2006;296(7):782-93.

3. Mutevedzi PC, Lessells RJ, Heller T, et al. Scale-up of a decentralized HIV treatment programme in rural KwaZulu-Natal, South Africa: does rapid expansion affect patient outcomes? *Bull World Health Organ* 2010;88(8):593-600.

4. Ferradini L, Jeannin A, Pinoges L, et al. Scaling up of highly active antiretroviral therapy in a rural district of Malawi: an effectiveness assessment. *Lancet* 2006;367(9519):1335-42.

5. UNAIDS. Bold new AIDS targets set by world leaders for 2015. 2011; http://www.unaids.org/en/resources/presscentre/pressreleaseandstatementarchive/2011/june/20110610psdeclaration/; (accessed 6 August 2011).

6. Bärnighausen T, Bloom DE, Humair S. Going horizontal--shifts in funding of global health interventions. *N Engl J Med* 2011;364(23):2181-3.

7. U.S. Global Health Initiative (GHI). *Implementation of the Global Health Initiative: consultation document*. Washington, D.C.,: GHI, 2009.

8. USAID. *USAID evaluation policy*. Washington, D.C., 2011.

9. The United States President's Emergency Plan for AIDS Relief (PEPFAR). Metrics, monitoring, research and innovation. 2009; http://www.pepfar.gov/strategy/ghi/134856.htm; (accessed 12 July 2011).

10. PEPFAR. Health Systems Strengthening (HSS). 2009; http://www.pepfar.gov/strategy/ghi/134854.htm; (accessed 1 March 2011).

11. PEPFAR. Executive summary of PEPFAR's strategy. 2009; http://www.pepfar.gov/strategy/document/133244.htm; (accessed 12 July 2011).

12. PEPFAR. Health systems strengthening. 2010; http://www.pepfar.gov/about/138338.htm; (accessed 1 March 2011).

13. PEPFAR. Annex V – health system strengthening priority-setting. 2011; http://www.pepfar.gov/guidance/framework/120741.htm; (accessed 15 October 2011).

14. The United States Government Global Health Initiative (GHI). *Strategy document*. Washington, DC: GHI, 2011.

15. The Global Fund to Fight AIDS TaMG. *The Global Fund's approach to health systems strengthening (HSS)*. Geneva: GFATM, 2011.

16. PEPFAR. About PEPFAR. 2011; http://www.pepfar.gov/about/index.htm; (accessed 27 January 2011).

17. The Global Fund to Fight AIDS TaMG. About the Global Fund. 2011; http://www.theglobalfund.org/en/about/secretariat/; (accessed 12 July 2011).

18. PEPFAR. 2009 Annual Report to Congress on PEPFAR Program Results. Washington, DC: PEPFAR, 2010.

19. Institute of Medicine. PEPFAR implementation: progress and promise. Washington, DC: The National Academies Press, 2007.

20. Global Fund to Fight AIDS, Tuberculosis and Malaria. Global Fund 2010 Innovation and Impact. Geneva: Global Fund to Fight AIDS, Tuberculosis and Malaria, 2010.

21. Global Fund to Fight AIDS, Tuberculosis and Malaria. Global Fund-supported programs deliver AIDS treatment for 3 million people. http://www.theglobalfund.org/en/pressreleases/?pr=pr_101201; (accessed 27 January 2011).

22. Global Fund to Fight AIDS, Tuberculosis and Malaria. Partners in Impact, Results Report. Geneva: Global Fund to Fight AIDS, Tuberculosis and Malaria, 2007.

23. Stover J, Johnson P, Zaba B, et al. The Spectrum projection package: improvements in estimating mortality, ART needs, PMTCT impact and uncertainty bounds. *Sex Transm Infect* 2008;84 Suppl 1:i24-i30.

24. Komatsu R, Korenromp EL, Low-Beer D, et al. Lives saved by Global Fund-supported HIV/AIDS, tuberculosis and malaria programs: estimation approach and results between 2003 and end-2007. *BMC Infect Dis* 2010;10:109.

25. Makwiza I, Nyirenda L, Bongololo G, et al. Who has access to counseling and testing and anti-retroviral therapy in Malawi - an equity analysis. *Int J Equity Health* 2009;8:13.

26. Bärnighausen T, Kyle M, Salomon J, et al. Assessing the population health impact of market interventions to improve access to antiretroviral treatment. *Health Policy and Planning (in press)* 2011.

27. Samb B, Evans T, Dybul M, et al. An assessment of interactions between global health initiatives and country health systems. *Lancet* 2009;373(9681):2137-69.

28. Biesma RG, Brugha R, Harmer A, et al. The effects of global health initiatives on country health systems: a review of the evidence from HIV/AIDS control. *Health Policy Plan* 2009;24(4):239-52.

29. Yu D, Souteyrand Y, Banda MA, et al. Investment in HIV/AIDS programs: Does it help strengthen health systems in developing countries? *Global Health* 2008;4:8.

30. Kruger AM, Bhagwanjee S. HIV/AIDS: impact on maternal mortality at the Johannesburg Hospital, South Africa, 1995-2001. *Int J Obstet Anesth* 2003;12(3):164-8.

31. Kruk ME. HIV and Health Systems: Research to Bridge the Divide. *J Acquir Immune Defic Syndr* 2011;57 Suppl 2:S120-3.

32. Barnighausen T, Bloom DE, Humair S. Universal antiretroviral treatment: the challenge of human resources. *Bull World Health Organ* 2010;88(12):951-2.

33. Padian NS, Holmes CB, McCoy SI, et al. Implementation science for the US President's Emergency Plan for AIDS Relief (PEPFAR). *J Acquir Immune Defic Syndr* 2011;56(3):199-203.

34. PEPFAR. About PEPFAR. 2010; http://www.pepfar.gov/about/index.htm; (accessed 10 August 2011).

35. The Global Fund to Fight AIDS TaMG. Value for money. 2011; http://www.theglobalfund.org/en/performance/effectiveness/value/; (accessed 10 August 2011).

36. Houlihan CF, Bland RM, Mutevedzi PC, et al. Cohort profile: Hlabisa HIV treatment and care programme. *Int J Epidemiol* 2011;40(2):318-26.

37. Donald SG, Lang K. Inference with difference-in-difference and other panel data. *Review of Economics and Statistics* 2007;89:221-233.

38. Shadish WR, Cook TD, Campbell DT. *Experimental and quasi-experimental designs for generalized causal inference*. Boston: Houghton Mifflin, 2001.

39. Lee DS, Lemieux T. Regression discontinuity designs in economics. *Journal of Economic Literature* 2010;48(June):281-355.

40. Angrist JD, Krueger A. Instrumental variables and the search for identification: from supply and demand to natural experiments. *The Journal of Economic Perspectives* 2001;15(4):69-85.

41. Mdege ND, Man MS, Taylor Nee Brown CA, et al. Systematic review of stepped wedge cluster randomized trials shows that design is particularly used to evaluate interventions during routine implementation. *J Clin Epidemiol* 2011;64(9):936-48.

42. De Allegri M, Pokhrel S, Becher H, et al. Step-wedge cluster-randomised community-based trials: an application to the study of the impact of community health insurance. *Health Res Policy Syst* 2008;6:10.

43. Miller G. Contraceptions as development. *Economic Journal* 2010;120(545):709-736.

44. Mann V, Fazzio I, King R, et al. The EPICS Trial: Enabling Parents to Increase Child Survival through the introduction of community-based health interventions in rural Guinea Bissau. *BMC Public Health* 2009;9:279.

45. Lewycka S, Mwansambo C, Kazembe P, et al. A cluster randomised controlled trial of the community effectiveness of two interventions in rural Malawi to improve health care and to reduce maternal, newborn and infant mortality. *Trials* 2010;11:88.