

Strengthening Non-Randomized Studies of Health Communication Strategies for HIV Prevention

Calum Davey¹

Marc Boulay²

James R Hargreaves¹

¹ London School of Hygiene and Tropical Medicine

² Health Communication Capacity Collaborative, Johns Hopkins Bloomberg School of Public Health

Contact information for corresponding author

Calum Davey

Calum.davey@lshtm.ac.uk

+44 020 7 927 2292

London School of Hygiene and Tropical Medicine

Keppel Street

London WC1E 7HT

UK

Word count:

Abstract: 244

Text: 3,777 (excluding box and figure)

Box 1: 515

Figures: 1

ABSTRACT

Background

There is growing interest in impact evaluations of health communication (HC) interventions for HIV prevention. While cluster randomized trials may be optimal in terms of internal validity, they are often unfeasible for political, practical or ethical reasons. However, a common alternative, the observational study of individuals who do and do not self-report HC intervention exposure, is prone to bias by confounding. Cluster-level quasi-experimental study designs offer promising alternatives to these extremes.

Methods

We identified common roll-out strategies for HC initiatives. We mapped these scenarios against established quasi-experimental evaluation designs. We identified key issues for implementers and evaluators if these designs are to be more frequently adopted in HC intervention evaluations with high internal validity.

Results

Stronger evaluations will document the planned intervention components in advance of delivery and will implement interventions in clusters according to a pre-defined, systematic allocation plan. We identify four types of allocation plan and their associated designs. Where some places get the HC intervention while others do not, a non-randomised-controlled study may be feasible. Where HC is introduced everywhere at a defined point in time, an interrupted time-series may be appropriate. Where the HC intervention is introduced in phases, a non-randomised phased-implementation or stepped-wedge design may be used. Finally, where there is variation in strength of implementation of HC a non-randomised, dose-response study can be planned.

Discussion

Our framework will assist teams planning such evaluations by identifying critical decisions for the implementers and for the evaluators of HC interventions.

INTRODUCTION

Health communication (HC) is a central component of the HIV prevention agenda. Early in the epidemic, policymakers identified promoting awareness as a priority global response to HIV,¹ and most national programs responded with information campaigns.² HIV-related HC efforts have evolved from straightforward media campaigns to encompass a range of communication activities seeking to influence behaviors associated with disease transmission^{3,4} and characteristics of the broader social environment in which these behaviors are embedded, such as stigma^{5,6} and gender norms.⁷

Several factors complicate impact evaluation of HC interventions. Individuals are often able to self-select for exposure to these interventions, potentially biasing comparisons of outcomes between those exposed and those unexposed.⁸ Also, evidence suggests that these messages may diffuse through informal community networks to influence individuals who do not directly see or hear program materials.^{9,10} Furthermore, many HIV-related communication initiatives have multiple simultaneous elements using overlapping messages and channels. Cluster-randomized trial designs (cRCTs) respond to many of these issues. In such trials, the outcome distribution among those allocated to the control arm can be interpreted as the potential outcome distribution that would have been observed in the intervention arm if the intervention had not been allocated. A comparison of outcome distributions between places with and without the intervention can therefore be interpreted as the average causal effect of allocation to the intervention.¹¹ Integrated process evaluation is an essential part of such studies since HC programs delivered in real-life settings do not guarantee that randomization alone will ensure useful results.¹²⁻¹⁴ While cRCTs offer the least-biased and simplest approach to estimating and understanding the intention-to-treat effect¹⁵, investigator-controlled random allocation of HC interventions is often not an option.

Broadly speaking, randomized trials control for confounding by design, while observational studies do so by analysis. Observational studies of HC interventions can have great value, and in some cases they are the only option available. Analytic approaches are described in the literature to adjust for confounding in such studies.¹⁶ However, evaluators also worry about confounding by unmeasured and poorly-understood factors. Instrumental variable approaches offer the chance to control confounding without measurement of all confounding variables, but, in practice, naturally-occurring valid instruments are rarely identifiable.¹⁷

If evaluators do not control allocation of the intervention, then evaluation “design” principally refers to making decisions about what, where, when, and from whom data are acquired. Collecting data is a necessary but not sufficient component of impact evaluation design, which also includes plans for the analysis of such data. The

most significant problem for an evaluation seeking to estimate intention-to-treat effects in a manner analogous to a cRCT is residual confounding by unknown, unmeasured, and/or imprecisely measured factors.¹⁸⁻²¹ This paper aims to improve and encourage the use of quasi-experimental designs in evaluations of HC strategies.

METHODS

HC interventions are naturally delivered in clusters—groups or areas—rather than to individuals. We refer to these units as “clusters”, which may be districts, towns, schools or any other politically or physically-determined unit. We emphasize situations where the primary aim of an evaluation is to estimate the causal effect of a defined program on HIV-related endpoints using the intention-to-treat principle (ITT).²³ Often, many people within intervention clusters will not be exposed, but the primary concern of an ITT-analysis is to estimate the overall effect in the target population. The scope of this paper does not permit discussion of several other critical design elements, e.g., population sampling, data validity or sample size estimation. Rather, we hope to help teams navigate design options and recognize critical decision points where impact evaluation may be strengthened.

We start by outlining two extreme but recognizable scenarios and associated evaluation designs, anticipating that readers will recognize both and agree that often neither design will match their needs or the real-life conditions in which evaluations are planned (Box 1). At one extreme is the cRCT which can produce internally-valid intention-to-treat estimates of intervention effects. At the other extreme is the observational study, susceptible to bias, that must rely on associations between self-reported exposure and end-points to estimate effects.

We argue that there is a “middle ground” of cluster-level quasi-experimental designs. These designs can be adapted for use in HC intervention roll-out scenarios commonly encountered by implementers and can give rise to valid effect estimates. The designs will require evaluators and implementers to work together and make informed compromises. Implementers should consider evaluation as part of intervention planning. Evaluations will be better able to produce valid estimates of impact, without randomization, where, in advance of deploying the intervention, the following are clearly documented:

1. The intervention components
2. Criteria that determine which clusters are eligible to receive the intervention
3. Criteria that determine which eligible clusters will actually receive the intervention; we refer to this as the presence of an allocation scheme.

Our call mirrors concerns in the causal inference literature, where it is argued that a counterfactual approach in public health requires that causal effects are defined in terms of contrasts between health outcomes

corresponding to different “well-defined” intervention conditions and, where analysis and design strategies allow, appropriate control of confounding.²²

Not all situations are amenable to evaluation but when evaluation is a major concern then satisfying these conditions should be feasible. Where these conditions are met, we discuss four quasi-experimental research designs that are based on implementation scenarios which are commonly encountered in real-life. These are shown in Figure 1 and described in more detail below.

Design 1: Non-randomized, controlled, comparison

The implementation plan may allocate some eligible clusters to receive the HC intervention but not others. The evaluation design may exploit this variation between clusters. For example, community-based HC programs, such as those that involve community drama; peer educators, or other change agents recruited in the community; or other interpersonal channels of communication, are typically implemented in a subset of communities within an overall project area. While mass media programs that use national broadcast channels would not fit this scenario, programs relying on community radio stations, with circumscribed broadcast areas, may reach just a subset of clusters. For example, a current trial in Burkina Faso is testing the effectiveness of a community radio-based intervention by defining the non-overlapping geographic catchment areas for 14 community radio stations and randomly allocating seven areas to receive messages on key health issues.

Random allocation is not always feasible and other considerations, such as a desire to target areas with less favorable health or economic indicators, may influence the selection of areas for implementation. An evaluation will be strengthened when the factors determining whether or not eligible clusters are allocated to receive the interventions are determined in advance and/or easily measured. Causal attribution is much harder in situations where allocation is driven by unknown factors or is chaotic and unplanned. The challenge facing evaluation teams is to measure outcomes in places that are and are not allocated to receive the interventions and, crucially, to be able to convincingly argue that differences in the outcome distributions between these places, after adjusted analysis, arise because of the intervention allocation in question, i.e., that the difference is not confounded by other factors.

In some cases the rules determining allocation may be complex. When the clusters allocated to receive the intervention are defined in advance by such rules, the design challenge is to identify other eligible clusters that will not receive the intervention that can act as controls. The intention is to select clusters that are alike, prior to intervention, in respects relevant to the outcome distribution. Matching may be used, incorporating geopolitical

factors such as cultural, health system, and political contexts. For example, the Young People’s Development Programme in the United Kingdom was an intervention delivered to schools that were selected through a competitive tendering process.²⁴ After sites were allocated to receive the intervention, comparison sites were drawn from among unsuccessful applications, matched to the intervention sites by region, local deprivation, teenage pregnancy rates, urban/rural/seaside residence, and sector (voluntary or statutory). The evaluators compared outcomes in intervention and matched schools.

It is often convenient to match on a small number of strong predictors of the outcome; however, where no single factor is strongly predictive of the end point, a “propensity score” approach may be used. A propensity score is calculated for each eligible cluster, usually using a logistic regression of potential confounding factors that “predict” whether or not a cluster will actually be allocated to receive the intervention. Clusters with similar scores to those of intervention clusters—i.e., judged to have a similar propensity to be allocated to the intervention—are considered, along with the intervention clusters, to have been effectively randomly allocated to receive the intervention or not. The propensity score can then be used in both design (e.g., for matching or defining eligibility) or analysis (e.g. as an independent variable). These approaches are described in more detail in a wide literature on the subject.^{25,26} Matched studies are more complex to analyze than unmatched designs and may have less statistical power.²⁷

In other situations, deterministic rules may be applied to define whether clusters actually receive the intervention. For example, Arcand et al. evaluated an HIV-education training module for school teachers in Cameroon.²⁸ For pragmatic reasons, although there were villages with between 1 and 8 schools each, only villages with 4 or fewer schools received the intervention. Thus, allocation was determined by a simple parameter. This had advantages; it was (a) specific, (b) on a continuous ordered scale, and (c) not closely related to the end point of interest. Villages with 3 or 4 schools were compared with villages with 5 or 6 schools, i.e., those either side of the arbitrary cut-off. The evaluators sought to show that clusters either side of the allocation criterion cut-off were similar and chose narrow inclusion bounds to reduce confounding. Arbitrary cut-offs may be utilized in a real-life evaluation; when evaluators and implementers are discussing the allocation schema, it may suit both parties to consider including such a cut-off.

In non-randomised-controlled studies, baseline data are especially useful to verify the comparability of the places with and without the intervention, to make adjustments in the analysis, and to identify particular places with important baseline differences that may need to be excluded.

Design 2: Interrupted Time Series

HC intervention implementation plans may result in all eligible clusters being allocated to receive the intervention at a given time, i.e. that there is variation in allocation status in time, but not between clusters. This scenario may occur for programs relying primarily on national mass media channels, which typically have well-defined phases separated by periods of time with little or no activity, but little or no planned variation geographically. For example, in Brazil PRO-PATER implemented three separate mass media campaigns promoting vasectomy in 1983, 1985, and 1989, and the number of vasectomies increased markedly during each campaign period.²⁹ More recently, a time-series analysis of condom sales in Ghana demonstrated an abrupt upward shift corresponding with the start of the Stop AIDS Love Life communication program.³⁰

An evaluation design for this scenario cannot make comparisons between clusters to estimate impact. Instead, the outcome time-trend before intervention is used to estimate the outcome trend if the intervention had not been implemented. This is distinct from a simple before-after comparison which does not account for temporal changes in the outcome distribution. Evidence of the effect of the intervention comes from an “interruption” in the prevailing outcome trend coinciding with the intervention.³¹ The interruption may be a break in the trend line or a change in the gradient of the trend. For example, in Ghana a time-series design was used to investigate the effect of two policy decisions on the proportion of pregnant women having deliveries that were assisted by a skilled attendant.³² In 2005 a delivery-fee exemption was rolled-out, then in 2008 the government exempted pregnant women from national insurance fees so that they were entitled to antenatal, childbirth, and postnatal care without charge. Data on time trends in the proportion of women giving birth in a facility were plotted over time. Even though there was an upward secular trend in the outcome, it was possible to convincingly isolate the impact of the policy changes on the outcome of interest.

This approach requires multiple data points before and after the introduction of the intervention; the number needed depends on a number of factors. Since the design relies on a good characterization of the prevailing trend in the outcome, the evaluators may need to draw on routine or surveillance data, e.g., antenatal clinic data on HIV infections, clinic registers, or data gathered for another study. Evaluators may look for specific places where sufficient data have been collected so that this design can be used and then work with implementers to balance the requirements of the design against their priorities for roll-out.

Since the dynamics of infectious diseases rarely conform to simple linear trends, mathematical models can draw on other data to help predict trends.³³ A lag between the interruption and a change in outcomes can make analysis more complicated and also widens the period when other events that potentially explain changes in the outcomes could have taken place. Optimal interventions for a time-series approach will be implemented at a

defined time-point, rapidly taken up by the target populations, and could feasibly cause changes in outcomes quickly.

Design 3: Phased-Implementation

The implementation plan may allocate clusters to initiate the intervention at different times, with eventual initiation in all clusters. A typical HC case is one where a mixture of community-level and mass media programs are initiated at different times in different places, possibly because of limitations in an NGO's capacity to train community leaders and produce locally relevant health messages. As an example, the Bridge Project in Malawi, between 2001 and 2008, used mass media and community-level interventions to communicate HIV prevention messages. Initially implemented in 8 of Malawi's 28 districts, it has since expanded to 11 more districts. When randomization determines when places are allocated to receive interventions, the evaluation design is known as a "stepped-wedge" or "phased-implementation" cluster-randomized trial design.

As in non-randomised-controlled studies (design 1), a challenge for evaluations of this scenario is ensuring that during the time periods where clusters do not change allocation status (phases) there is "balance" between arms on important characteristics. As in the time-series design (design 2), phased-implementation studies often include multiple measurements over time. Analysis of the data from such studies can thus be thought of in two ways. A "horizontal" approach estimates the secular trend in the outcome in the clusters that are not changing intervention condition, and accounts for this trend in the before and after comparison of outcome data in clusters changing intervention allocation status. The challenge is ensuring that the measured trend is a valid estimate of the expected trend in the clusters that change intervention status. The second approach compares clusters with and without the intervention within phases and combines the within-phase estimates, making no assumptions about the nature of the secular trend—a "vertical approach." The analyses of stepped-wedge trials with randomized start times will often combine horizontal and vertical approaches.³⁴

The probability of detecting an effect may be reduced if there is a lag between the introduction of the intervention and a change in the outcomes, as with HIV prevention, or if the full intervention is not realized during the time between steps.³⁵

Design 4: Implementation Strength

The implementation plan may entail variation in the strength of the intervention allocated to clusters, with some clusters allocated a greater "dose" of activities than others. An evaluation design that relies on this variation

could be appropriate in situations where a program uses a single channel, for instance, small-group activities at the community level, to communicate messages, with activities occurring more often in some communities than in others. Alternatively, since most large-scale programs now use multiple channels to communicate health messages, this design may also apply when the number of program channels differs across clusters. For instance, the COMMIT project in Tanzania used both mass media and community-based activities to communicate messages promoting behaviors to reduce the transmission of malaria. The program's mass media messages reached all communities, but only some clusters had community-based group activities and, in even fewer communities, the project recruited community members to serve as local change agents promoting malaria prevention. The allocation of these channels across communities would allow program evaluators to measure the dose of the intervention for each cluster.

As for a non-randomised-controlled study (design 1), the evaluation design and analysis will need to account for potential confounding arising from differences between clusters that receive different strengths of intervention. Ideally, the variation in implementation strength will be planned so that the results are in keeping with the ITT principle. However, where this variation is not planned, the next best option will be to estimate variations in implementation strength as it happens. Developing an index of implementation strength involves a numerator, e.g., money spent on interventions, and a denominator, e.g., the size of the target population. Few research studies using this design with an ITT approach are found in the literature. An example with a measured index of intensity comes from an evaluation of the impact of Avahan, a large, targeted HIV prevention intervention in India.³⁶ Ng et al. estimated the intensity of the intervention using the money spent in each district per year on targeted interventions (\$) divided by the estimated median number of people living with HIV (PLHIV) in each district.³⁷ The cumulative HIV allocation intensity (\$/PLHIV) was summed from the start of the program until year t , and regressed against HIV prevalence among individuals attending antenatal care clinics in year t . Using a multilevel regression analysis approach, they estimated the association between cumulative resource allocation for interventions (\$/PLHIV) in a district and the odds of a particular woman at an antenatal clinic being HIV-positive. In this design detailed plans and budgets will be useful, and the evaluation may benefit from following how allocation intensity changes with time.

An additional challenge with this design is interpreting the dose effects. We have suggested that intensity can be indexed using a continuous variable such as frequency of radio transmissions or in terms of overlapping components. Although we are more concerned here with identifying simple dose effects of increasing intensity on outcomes, the interpretation of a dose effect may include combination effects from different components

acting together. Process evaluation, as well as a comprehensive theory of change, may help evaluators interpret their results.

Combining Designs

How well the assumptions of each design are met may inform the choice of design for a particular situation. It is unlikely that all of the assumptions of any one of the designs will be completely satisfied, and practical factors such as cost and the availability of data may make one option stand out over another. Combining methods can balance the limitations of each design. However, if different methods find different results, interpretation can be difficult. A combination of methods should not be viewed as mutually exclusive routes but rather as mutually supporting options for evaluating an intervention.

DISCUSSION

HC evaluation teams should more commonly deploy quasi-experimental study designs, as these studies can yield greater validity than purely observational studies. Designing such studies can be organized around common HC roll-out scenarios. Maximizing the utility of these designs will require collaboration from the outset between those primarily concerned with implementation and those primarily concerned with evaluation. Such HC evaluations will be strengthened if, in advance of implementation: (1) the planned intervention components are described, (2) cluster eligibility criteria are defined, and (3) intervention allocation criteria are defined and is driven by predictable and measurable factors.

We advocate better and closer communication between evaluators and implementers, up to and including having evaluators influence roll-out of the intervention. We recognize that this may be difficult when the evaluation is strictly “external” to the implementation, e.g., when evaluators and implementers are based at separate institutions and when there is a mind-set that the “independence” of the evaluation is based largely on the separateness of these two groups. We argue that the face-validity of the evaluation is increased with good design and that procedures such as protocol registration and pre-analysis plans can increase the transparency of the method. We do not wish to argue against the merits of external evaluations, but rather that this should not be pursued at the expense of the simple ways that collaboration can improve the evaluation design.

For each intervention allocation scenario there will be many possible evaluation designs. We have focused on the problem of identifying the intention-to-treat effect. The proposed approaches emphasize evaluation questions seeking to identify *whether* the program had an effect and, in themselves, may not necessarily inform questions seeking to identify *how* the programs may have influenced changes in behavior. Comprehensive evaluations of HC programs ideally include assessments of the applicability of the theoretical hypotheses informing the messages used by a program.³⁸⁻⁴⁰ Assessing the theory of change associated with an HC program provides insight into the relative effectiveness of the specific messages and informs program refinements. The validity of an evaluation is further determined by such factors as monitoring of the intervention as it is delivered, data on intervention availability in comparison places, and data on intermediate factors in the theory of change.

One potential limitation of these approaches arises from the difference between intervention allocation and intervention exposure. While HC programs allocate intervention messages at the cluster level, exposure to these messages occurs at the individual level. In situations with high levels of exposure to intervention messages, a high level of correspondence will exist between membership in an intervention cluster and exposure to a program’s messages. When exposure to a program’s messages is relatively low, it may be harder to detect ITT

effects simply because of the low exposure levels. However, in communities with a cohesive social structure and a high level of interpersonal communication about health topics, the diffusion of program messages through peer networks may mitigate the problems associated with low levels of direct message exposure.

Evaluations in real-life contexts may struggle to achieve the internal validity of a cRCT, but quasi-experiments have advantages in terms of their external validity.^{41,42} Evaluations delivered at scale and with the budget and oversight of real-life implementation may have greater external validity than a cRCT performed in limited conditions with an unrealistic implementation budget.

Overcoming the barriers to timely communication between implementing and evaluating partners will go a long way in strengthening evaluation results. Moreover, since donors, civil society, governments, and are increasingly interested in knowing “what works,” we hope that the vision and funding will be available to ensure that implementers and evaluators work as partners.

Box 1. Two Extreme Scenarios and Evaluation Designs

The Cluster RCT

A nongovernmental organization (NGO) has been given funds to design, implement, and evaluate a new multi-component HIV prevention communication strategy. The planned components and their delivery are carefully described. The evaluators help design, and thus understand in detail, the implementation plan and advise use of a cluster-randomized trial to evaluate impact. Communities are randomized to receive or not receive the full implementation package. Cluster size and spacing are carefully considered to allow the full potential impact to be evaluated, including impacts from secondary communications, while ensuring that contamination between clusters is minimized. The length of the trial and the number of clusters and of people sampled allow detection of theoretically plausible differences between trial arms, using cohorts to measure HIV incidence.

The evaluators carefully monitor implementation as the intervention is rolled out, assessing how delivery in practice compares with the initial plan. High-quality data are collected on the level of accessibility and acceptability of the new strategies. Data are collected on levels of exposure to components and on intermediary outcomes such as recognition of intervention materials, knowledge of the content of the intervention, and behaviors that the intervention planned to change. These data are used to estimate the effect of the interventions among those who received it, to allow pathways of change to be investigated, to provide data-informed hypotheses why the intervention had impact or not, and to inform any future scale-up or generalization.

An Observational Study

An NGO has been given funds to design, implement and evaluate the impact of a new multi-component HIV prevention communication strategy. The implementing agency contacts the evaluators, stressing the importance of the evaluation but also the urgency and complexity of the intervention. As the evaluators begin thinking about a design, the implementers identify places where they are most likely to deliver some aspects of the intervention rapidly, and they start channeling funds and human resources to these places. Some aspects of the intervention begin in some places and other aspects in others, to differing degrees depending on local providers, infrastructure, and political and other factors. The implementers put in place a combination of strategies to give extra support to places where things are going poorly and to extend the project in places where things are working well.

The evaluators realize that things have moved on and jettison the idea of a prospective design. They are faced with choices about where to collect data, on whom, and on what. They can see that there is significant variation in what has been implemented, although they are uncertain exactly what has driven the variation. They commission a large survey collecting data on HIV infections. They design survey tools to capture whether study participants have engaged with or seen components of the intervention, but they are concerned that data from individuals on exposure to the interventions is probably highly flawed. They plan an analysis in which associations between exposure and recent HIV infection are assessed, using a variety of complex statistical approaches to try and adjust for measured potential confounding factors.

Table 1: Health communication intervention implementation scenarios and quasi-experimental design options

Implementation scenario	Evaluation Design Option	Design Schematic
1. Some clusters are non-randomly allocated to the program, but others are not	Non-randomized, controlled, comparison	A 1 1 1 1 B 1 1 1 1 C 0 0 0 0 D 0 0 0 0
2. All clusters are allocated to start the program at the same time	Interrupted time-series	A 0 0 1 1 B 0 0 1 1 C 0 0 1 1 D 0 0 1 1
3. Clusters are allocated to start the program at different times	Non-randomized stepped-wedge or phased implementation	A 1 1 1 1 B 0 1 1 1 C 0 0 1 1 D 0 0 0 1
4. All clusters allocated some of the program but at different intensities	Dose-response or implementation-strength	A .2 .2 .2 .2 B .8 .8 .8 .8 C .3 .3 .3 .3 D .5 .5 .5 .5

Key: Letters A-D represent clusters or groups of clusters. Numbers indicate intervention allocation status, 0 = no intervention, 1= intervention, numbers between 0 and 1 represent variation in intensity of intervention. Time is shown along the horizontal axis.

REFERENCES

1. Piot P, Kreiss JK, Ndinya-Achola JO, et al. Heterosexual transmission of HIV. Editorial review. *AIDS*. 1987;1:199–206.
2. Oakley A, Fullerton D, Holland J. Behavioural interventions for HIV/AIDS prevention. *AIDS*. 1995;9:479.
3. Vaughan W, Rogers EM, Singhal A, et al. Entertainment-education and HIV/AIDS prevention: a field experiment in Tanzania. *J Health Commun*. 2000;5:81–100.
4. Rimal RN, Creel AH. Applying social marketing principles to understand the effects of the radio diaries program in reducing HIV/AIDS stigma in Malawi. *Health Mark Q*. 2008;25:119–146.
5. Boulay M, Tweedie I, Fiagbey E. The effectiveness of a national communication campaign using religious leaders to reduce HIV-related stigma in Ghana. *Afr J AIDS Res*. 2008;7:133–141.
6. Hutchinson P, Mahlalela X, Yukich J. Mass media, stigma, and disclosure of HIV test results: multilevel analysis in the Eastern Cape, South Africa. *AIDS Educ Prev*. 2007;19:489–510.
7. Underwood C, Skinner J, Osman N, et al. Structural determinants of adolescent girls' vulnerability to HIV: views from community members in Botswana, Malawi, and Mozambique. *Soc Sci Med*. 2011;73:343–350.
8. Hutchinson P, Wheeler J. Advanced methods for evaluating the impact of family planning communication programs: evidence from Tanzania and Nepal. *Stud Fam Plan*. 2006;37:169–186.
9. Boulay M, Storey JD, Sood S. Indirect exposure to a family planning mass media campaign in Nepal. *J Health Commun*. 2002;7:379–399.
10. Hwang Y. Social diffusion of campaign effects campaign-generated interpersonal communication as a mediator of antitobacco campaign effects. *Commun Res*. 2012;39:120–141.
11. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. *J Educ Psych* 1974;66:688–701.
12. Craig P, Dieppe P, Macintyre S, et al. Developing and evaluating complex interventions: the new Medical Research Council guidance. *BMJ*. 2008;337.
13. Oakley A, Strange V, Bonell C, et al. Process evaluation in randomised controlled trials of complex interventions. *BMJ*. 2006;332:413–416.

14. Bonell C, Fletcher A, Morton M, et al. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. *Soc Sci Med.* 2012;75:2299–2306.
15. Bonell CP, Hargreaves J, Cousens S, et al. Alternatives to randomisation in the evaluation of public health interventions: design challenges and solutions. *J Epidemiol Community Health.* 2011;65:582–587.
16. Little RJ, Rubin DB. Causal Effects in Clinical and Epidemiological Studies Via Potential Outcomes: Concepts and Analytical Approaches. *Annual Review of Public Health* 2000,21:121–145.
17. Greenland S. An introduction to instrumental variables for epidemiologists. *Int J Epidemiol.* 2000;29:722–729.
18. Hernan MA, Hernandez-Diaz S, Werler MM, et al. Causal knowledge as a prerequisite for confounding evaluation: an application to birth defects epidemiology. *Am J Epidemiol.* 2002;155:176–184.
19. Smith GD, Phillips AN. Confounding in epidemiological studies: why "independent" effects may not be all they seem. *BMJ.* 1992;305:757–759.
20. Greenland S, Robins JM. Identifiability, exchangeability, and epidemiological confounding. *Int J Epidemiol.* 1986;15:413–419.
21. Cousens S, Hargreaves J, Bonell C, et al. Alternatives to randomisation in the evaluation of public-health interventions: statistical analysis and causal inference. *J Epidemiol Community Health.* 2011;65:576–581.
22. Glass TA, Goodman SN, Hernan MA, et al. Causal inference in public health. *Annu Rev Public Health.* 2013;34:61–75.
23. Hollis S, Campbell F. What is meant by intention to treat analysis? Survey of published randomised controlled trials. *BMJ.* 1999;319:670–674.
24. Wiggins M, Bonell C, Sawtell M, et al. Health outcomes of youth development programme in England: prospective matched comparison study. *BMJ.* 2009;339:b2534.
25. Do MP, Kincaid DL. Impact of an entertainment-education television drama on health knowledge and behavior in Bangladesh: an application of propensity score matching. *J Health Commun.* 2006;11:301–325.
26. Dehejia RH, Wahba S. Propensity score-matching methods for nonexperimental causal studies. *Rev Econ Stat.* 2002;84:151–161.

27. Klar N, Donner A. The merits of matching in community intervention trials: a cautionary tale. *Statistics in Medicine* 1997;16:1753–1764.
28. Arcand JL, Wouabe ED. Teacher training and HIV/AIDS prevention in West Africa: regression discontinuity design evidence from the Cameroon. *Health Econ* 2010;19 Suppl:36–54.
29. Kincaid DL, Merritt AP, Nickerson L, et al. Impact of a mass media vasectomy promotion campaign in Brazil. *Int Fam Plann Persp.* 1996;22:169–175.
30. Bertrand JT, Holtgrave DR, Gregowski A. HIV/AIDS programs in the US and developing countries. In: Mayer KH, Pizer HF, eds. *HIV Prevention: A Comprehensive Approach*. New York: Elsevier; 2009;pp. 571–590.
31. Grijalva CG, Nuorti JP, Arbogast PG, et al. Decline in pneumonia admissions after routine childhood immunisation with pneumococcal conjugate vaccine in the USA: a time-series analysis. *Lancet.* 2007;369:1179–1186.
32. Dzakupasu S, Soremekun S, Manu A, et al. Impact of free delivery care on health facility delivery and insurance coverage in Ghana’s Brong Ahafo region. *PLoS ONE.* 2012;7:e49430.
33. Garnett GP, Cousens S, Hallett TB, et al. Mathematical models in the evaluation of health programmes. *Lancet.* 2011;378:515–525.
34. Hussey MA, Hughes JP. Design and analysis of stepped wedge cluster randomized trials. *Contemp Clin Trials.* 2007;28:182–191.
35. Moulton LH, Golub JE, Durovni B, et al. Statistical design of THRio: a phased implementation clinic-randomized study of a tuberculosis preventive therapy intervention. *Clin Trials* 2007;4:190–199.
36. Verma R, Shekhar A, Khobragade S, et al. Scale-up and coverage of Avahan: a large-scale HIV-prevention programme among female sex workers and men who have sex with men in four Indian states. *Sex Transm Inf.* 2010;86:i76–i82.
37. Ng M, Gakidou E, Levin-Rector A, et al. Assessment of population-level effect of Avahan, an HIV-prevention initiative in India. *Lancet.* 2011;378:1643–1652.
38. Chen H-T, Rossi PH. The multi-goal, theory-driven approach to evaluation: a model linking basic and applied social science. *Soc Forces.* 1980;59:106–122.
39. Weiss CH. Theory- based evaluation: past, present, and future. *New Directions Eval.* 1997;1997:41–55.

40. Kincaid DL, Do MP. Multivariate causal attribution and cost-effectiveness of a national mass media campaign in the Philippines. *Journal of Health Communication* 2006,11:69–90.
41. Cartwright N. Are RCTs the Gold Standard? *BioSocieties* 2007,2:11–20.
42. Deaton AS. Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development. *National Bureau of Economic Research Working Paper Series* 2009, No. 14690.