Health Policy and Planning, 31, 2016, 860–867
doi: 10.1093/heapol/czw001
Advance Access Publication Date: 9 March 2016
Original Article

Measuring implementation strength: lessons from the evaluation of public health strategies in low- and middle-income settings

James R Hargreaves,[1,*] Catherine Goodman,[1] Calum Davey,[1] Barbara A Willey,[2] Bilal Iqbal Avan[3] and Joanna R M Armstrong Schellenberg[3]

[1]Faculty of Public Health and Policy, [2]Faculty of Epidemiology and Population Health and [3]Faculty of Infectious and Tropical Diseases, London School of Hygiene & Tropical Medicine, Keppel Street, London, WC1E 7HT, UK

*Corresponding author. Faculty of Public Health and Policy, London School of Hygiene & Tropical Medicine, Keppel Street, London WC1E 7HT, UK. E-mail: james.hargreaves@lshtm.ac.uk

Accepted on 29 December 2015

Abstract

Evaluation of strategies to ensure evidence-based, low-cost interventions reach those in need is critical. One approach is to measure the strength, or intensity, with which packages of interventions are delivered, in order to explore the association between implementation strength and public health gains. A recent systematic review suggested methodological guidance was needed. We described the approaches used in three examples of measures of implementation strength in evaluation. These addressed important public health topics with a substantial disease burden in low-and middle-income countries; they involved large-scale implementation; and featured evaluation designs without comparison areas. Strengths and weaknesses of the approaches were discussed. In the evaluation of Ethiopia's Health Extension Programme, implementation strength scoring for each kebele (ward) was based on aggregated data from interviews with mothers of children aged 12–23 months, reflecting their reports of contact with four elements of the programme. An evaluation of the Avahan HIV prevention programme in India used the cumulative amount of Avahan funding per HIV-infected person spent each year in each district. In these cases, a single measure was developed and the association with hypothesised programme outcomes presented. In the evaluation of the Affordable Medicines Facility—malaria, several implementation strength measures were developed based on the duration of activity of the programme and the level of implementation of supporting interventions. Measuring the strength of programme implementation and assessing its association with outcomes is a promising approach to strengthen pragmatic impact evaluation. Five key aspects of developing an implementation strength measure are to: (a) develop a logic model; (b) identify aspects of implementation to be assessed; (c) design and implement data collection from a range of data sources; (d) decide whether and how to combine data into a single measure; and, (e) plan whether and how to use the measure(s) in outcome analysis.

Key words: Evaluation, implementation strength, malaria, HIV/AIDS, maternal, newborn and child health, Africa, Asia

Key Messages

- Measuring the strength of programme implementation and assessing its association with outcomes is a promising approach to strengthen pragmatic impact evaluation, both to assess impact and to identify which aspects of a programme need to be strengthened.
- We suggest a five-step approach for developing a measure of implementation strength: (a) develop a logic model; (b) identify the aspects of implementation to be assessed; (c) design and implement data collection from a range of data sources; (d) decide whether and how to use a single measure; and (e) plan whether and how to use the measure in statistical analysis.

Background

In low-and middle-income settings, weak health care systems are common (Mills 2004), and the gap between knowledge of proven interventions and routine practice is wide. For example, in the 75 countries which account for 95% of annual global maternal and child deaths, just 62% of births have a skilled attendant, 26% of babies have postnatal care within 2 days, and only 42% of children with suspected pneumonia receive antibiotic treatment (Countdown to 2015, 2013). Identifying successful implementation approaches is essential to help ensure that evidence-based, low-cost interventions reach those in need.

Evaluations of strategies to improve public health are conducted for multiple reasons including impact evaluation, course correction and accountability to donors (Bryce *et al.* 2005; Horton *et al.* 2008). It may be important to know which interventions can be delivered at scale; how to optimize these in a new setting; and whether scale-up efforts result in public health gains. Randomized trials and plausibility studies (Habicht *et al.* 1999) based on a comparison of changes over time in both intervention and comparison areas may not be feasible. One evaluation approach is to measure the strength or intensity with which packages of interventions are delivered as they are rolled out, with a view to exploring the association between implementation strength and public health gains (Victora *et al.* 2011; Schellenberg *et al.* 2012). For this purpose a measure of implementation strength might ideally range from zero to the level sufficient to achieve a specified change in either intervention coverage or in health outcomes in a particular context (Schellenberg *et al.* 2012). An impact evaluation might assess whether and how changes in health outcomes vary by implementation strength. Evidence of an association between an improvement in health outcomes and implementation strength could provide evidence that the improvement in outcomes is due to the programme. Such an assessment might form either the primary analysis strategy, or an approach complementary to other approaches such as before–after studies, or as a secondary analysis conducted within a trial.

The term 'implementation strength' has been used interchangeably with 'implementation intensity' and is defined here as a quantitative measure of the amount of input to, or activity to support, the implementation of a programme. This is related to, but differs from, the concept of intervention fidelity. Fidelity relates to the degree to which a specified intervention or programme is delivered as planned, e.g. as documented in a research protocol or programme document (Carroll *et al.* 2007). Although there is common ground in the concepts of fidelity and strength, our focus here is less on whether implementation realities match up to pre-defined plans, and more on documenting the strength of implementation put in place in practice. In further contrast to implementation fidelity, in some research designs implementation strength might be deliberately intended to vary from one place to the next.

Implementation strength is an appealing concept, yet it can be unclear how to put it into practice. A systematic review of 25 studies identified 3 common approaches (Schellenberg *et al.* 2012). Some authors developed scales with detailed descriptors of absolute levels of implementation (e.g. '2 h of contact per week'). Others developed relative strength scores often converted to a percentage scale. Finally, some scales identified whether particular aspects of a programme were in place and to what extent. A major conclusion of the review was that more guidance was required in this area. In this article, we reflect on three studies, none of which were included in the review. Two of these studies come from work with which our author group had been directly involved (Tougher *et al.* 2012; Karim *et al.* 2013), while the third came from a study that we had previously appraised (Ng *et al.* 2011). We reflect on how measures of implementation strength had been used in these large-scale programme evaluations. We explore the advantages and disadvantages of the approaches and in discussion offer guidance for how measures of implementation strength could be developed in the future.

Methods

We focused on three case study evaluations that addressed important public health topics that carry a substantial morbidity and mortality burden in low- and middle-income countries. They all involved large-scale implementation and featured evaluation designs without comparison areas. For each we identified the questions addressed by the evaluations and developed our own simple logic model to describe our understanding of the intended pathway between the intervention package and its intended outcomes (WHO 2007; Bryce *et al.* 2011). We distinguished between inputs, processes, outputs, outcomes and impacts. For example, 'inputs' such as money, medicines and health staff together with 'processes' such as training health staff to use the medicines or mass media to promote use of services might have been intended to lead to 'outputs' such as utilization of health services by patients in need and 'outcomes' such as sick children correctly treated and 'impact' such as lives saved. Typically, implementation strength measures captured information on inputs and/or processes while programme effect measures related to outputs, outcomes or impacts.

We identified the main measure(s) of implementation strength used by the authors, and describe the data collection approach deployed and how the implementation strength measures were developed. Finally, we identified how the relationship between implementation strength and public health outcomes was investigated, or (in one case) why this was judged not to be possible. Since outcomes and impacts are usually affected by issues other than the programme under study, impact analysis will also require

Table 1. Five steps in developing and using an implementation strength measure in impact evaluation studies, illustrated with three case study examples

	HEW Programme, Ethiopia	Avahan, India	Affordable Medicines Facility—malaria, multi-country
1. Develop a Logic Model	See Figure 1	See Figure 2	See Figure 3
2. Identify aspects of implementation strength to be measured	Strength of health extension programme's outreach activities	Amount of spending on Avahan HIV prevention approach	Duration of implementation, supporting intervention disbursements and coverage
3. Data collection	Surveys with mothers (independent of outcome assessment)	Accounting data	Document review, key informant interviews, accounting information and provider survey
4. Develop measures	Kebele-level score developed from four indicators	District-level metric of cumulative amount of Avahan funding per HIV-infected person spent each year	Country-level assessments of each aspect of implementation; no attempt to combine in a single score
5. Use in outcome analysis	Secondary analysis correlated with outcomes, adjusting for measured potential confounders	Primary analysis correlated with outcomes, adjusting for measured potential confounders	No formal association with outcomes attempted, but overall patterns of implementation and outcomes were compared across countries

measurement and adjustment for contextual factors that change over time and that might act as potential confounders or modifiers of the relationship between implementation strength and programme effects. Although we describe and reflect on the approaches used in the case studies for this purpose, it was not our aim in this article to critically appraise the overall study designs, assess the potential for bias through uncontrolled confounding in each specific example, or appraise the approaches used to control confounding. Rather our aim was to focus on the development and use of the measures of implementation strength.

Results

Table 1 provides a summary of the three cases.

Case Study 1: Evaluation of the effect of outreach activities through Ethiopia's health extension programme on maternal and newborn care practices, Karim *et al.* (2013).

Overview of intervention

Ethiopia's health extension programme was launched in 2003 with the aim of providing universal access to primary health care services, particularly preventive care, through over 34 000 female salaried health extension workers (HEWs). Two HEWs in each 'kebele' (ward) served around 5000 people, and spent 75% of their time on outreach activities: household visits, educating families to adopt healthy behaviours and serve as 'model families' in their area, and organizing others to participate in health extension programme services. Volunteers drawn from 'model family' households supported the HEWs by promoting health messages in the community. Promoted child survival strategies included immunization, vitamin A, oral rehydration therapy, malaria prevention through treated mosquito nets and nutrition education. Although promotion of essential newborn care including clean childbirth, cord care, thermal care, immediate and exclusive breastfeeding and extra care for low-birthweight babies was part of the health extension programme, prior to 2009 the HEWs had no relevant skills in this area. In 2009 the 'Last 10 km' project (The Last Ten Kilometers Project. http://l10k.jsi.com/, accessed 11 March 2015) introduced a programme to give HEWs skills and tools to promote essential newborn care practices in 101 'woredas' (districts), with a population of around 11.6 million people, ~16% of Ethiopia's population. Figure 1 shows a simplified logic model for one of the maternal and newborn care practices, initiating breastfeeding immediately after birth, a practice which has been shown to save lives (The Partnership for Maternal NCH 2011).

Evaluation design

This study evaluated the effect of the health extension programme's outreach activities on maternal and newborn care practices at household level, through an assessment of whether changes over time in care practices varied by the intensity of the health extension programme's outreach activities. The evaluation used a before–after comparison in 101 woredas, comparing household maternal and newborn care knowledge and practices at baseline (2008) with those reported 2 years later (2010). In further analysis, changes from baseline in care practices in kebeles with relatively low programme intensity were compared with changes from baseline in kebeles with relatively high programme intensity.

Definition and measurement of implementation strength

Implementation strength was expressed as 'programme intensity' and gave a measure of exposure to the health extension programme services. The implementation strength score was developed for each 'kebele' from interviews with mothers of children aged 12–23 months, whereas household, maternal and newborn care knowledge and practices were assessed in women with a child aged 0–11 months. Implementation strength scoring was based on four outreach activities: (a) %age of women reporting household visits by HEWs in the 6 months prior to the survey; (b) %age of women reporting household visits by community volunteers in the 6 months prior to the survey; (c) the proportion of families with a family health card; and (d) the proportion of households that were either 'model families', or were working towards 'model family' status. The data were combined and a score was calibrated to range between 0 and 10, with a higher score indicating greater strength of implementation.

Association of implementation strength with outcomes

Between baseline and endline surveys, the percentage of mothers who had initiated breastfeeding immediately after birth increased by

Inputs	Processes	Outputs	Outcomes	Impact
Health extension programme training, tools, and services	Home visits Health card ownership Model families	Pregnant women counselled on immediate breastfeeding	Initiating breastfeeding immediately after birth	Newborn survival

Figure 1. Simplified logic model for the effect of Ethiopia's health extension programme services on early initiation of breastfeeding.

Inputs	Processes	Outputs	Outcomes	Impact
Money and 'Avahan' design and methods	Contract and monitor organisations to deliver a package of HIV prevention interventions according to common minimum program	Number and % of target population met monthly, Number of condoms distributed against estimated gap, number and percent of target population visiting the clinic every 3 months	High quality prevention interventions used by high-risk groups Safer behaviours	HIV infections averted among high-risk populations HIV infections averted among the general population

Figure 2. Simplified logic model for the effect of the Avahan initiative on HIV prevention in India at the district level.

an average of 8% points [95% confidence interval (CI): 5–12, from 46 to 54%]. This change over time could have been due to a secular trend or to other health programmes. Analysis of implementation strength showed that the increase was greater in areas with higher implementation strength score, with a 10% point increase in implementation strength being associated with a 10% point increase in the odds of immediate breastfeeding (95% CI: 1–20%, $P = 0.017$). Other outcomes showed similar results.

Case Study 2: Evaluation of an HIV prevention initiative in India: Ng *et al.* (2011).

Overview of intervention

Between 2003 and 2008, $258 million was invested in the Avahan initiative in India, with the aim of reducing transmission of HIV through increased coverage of preventive interventions in high-risk groups. These high-risk populations included female sex workers and their clients and partners, injecting drug users and truck drivers. Components of the intervention included: safe-sex counselling through peer outreach; treatment of sexually transmitted infections; distribution of free condoms; needle and syringe exchange; and advocacy and community mobilization. The interventions were delivered by sub-contracted non-governmental and community-based organizations, co-ordinated by state-level implementing partners and a central team. Our simplified logic model for the initiative is shown in Figure 2. We focus below on Ng *et al.*'s study which used one measure of the strength of Avahan implementation to evaluate the impact of the programme. Several other evaluations of Avahan have been undertaken (Ramakrishnan *et al.* 2010; Thilakavathi *et al.* 2011; Pickles *et al.* 2013) with wide variety in methods, geographic focus and outcomes, but we focus on the Ng *et al.* article to illustrate the use of implementation strength for public health evaluation.

Evaluation design

The Avahan programme was not delivered with the same strength across all of the six target Indian States, including after accounting for the number of people living with HIV in an area. Ng *et al.* take advantage of this variation to estimate the impact of the programme. They used regression methods to assess how variation in implementation strength was associated with trends in HIV prevalence in the general population measured through antenatal-clinic based surveillance, adjusting for a range of potential confounding factors at the individual level. The approach did not explicitly account for whether or not the Avahan programme was the sole provider of services in the districts in which it was implemented.

Definition and measurement of implementation strength

Using programme and demographic data the measure of implementation strength was calculated as the cumulative amount of Avahan funding per HIV-infected person spent each year in each district over the period 2003–2008. This figure varied from $24 to $433 per HIV-infected person.

Association of implementation strength with outcomes

Ng *et al.* (2011) concluded that in the Southern states, where the HIV epidemic was largely concentrated in high-risk sexual networks, every $100 increase in Avahan investment was associated with an 18% (95% CI: 4–32%) reduction in the odds of HIV infection. The effect of Avahan seemed to be less pronounced in the Northeastern states, where the HIV epidemic was concentrated in networks of people who inject drugs, and where a $100 increase in Avahan investment was associated with a 4% (95% CI: −6 to 14%) reduction in the odds of HIV infection.

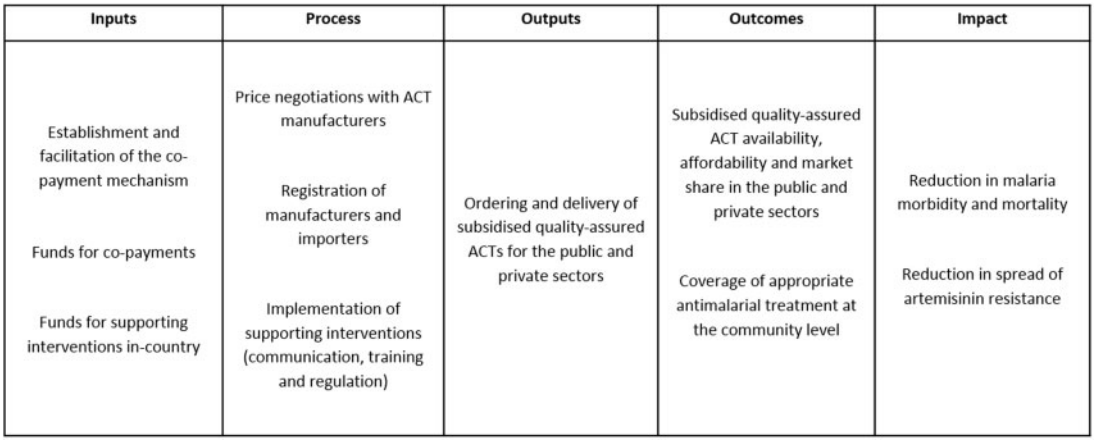

Figure 3. Simplified logic model for the effect of the AMFm on improved malaria treatment.

Case Study 3: Evaluation of the Affordable Medicines Facility—malaria (AMFm), Tougher *et al.* (2012).

Overview of intervention

The AMFm aimed to improve treatment for uncomplicated malaria. AMFm was hosted by the Global Fund to Fight AIDS, Tuberculosis and Malaria (the Global Fund) and designed to improve access to and use of quality-assured artemisinin combination therapies (ACTs). The programme was launched in 2010 through eight national-scale pilots—Ghana, Kenya, Madagascar, Niger, Nigeria, Tanzania mainland, Uganda and Zanzibar—and included providers of antimalarials in all sectors (public, not-for-profit and for-profit). There were three main components: negotiations with manufacturers to reduce factory prices; subsidy of quality-assured ACTs through a co-payment mechanism; and supporting interventions to encourage appropriate use. The co-payment mechanism meant that approved importers could purchase ACTs from approved manufacturers, at between 1 and 20% of the re-negotiated factory price, with the balance paid to the manufacturer by the AMFm co-payment fund. Once medicines arrived in-country they were distributed to end users through standard public and private distribution channels. Supporting interventions in each country included communication campaigns, recommended retail prices, provider training, and strengthening of regulation and pharmacovigilance. All pilots had equal access to the benefits of price negotiations with manufacturers and the establishment of the co-payment mechanism at the Global Fund. Our simplified logic model for AMFm is presented in Figure 3.

Evaluation design

Tougher *et al.*'s (2012) independent evaluation of AMFm reported the effect of the strategy on prices, availability and market share of quality-assured ACTs, and coverage of appropriate antimalarial treatment at the community level. They used a non-experimental design, based on pre–post comparisons, with baseline and endline 1 year apart. Control areas were not possible because implementation was on a national scale. Outcome results were assessed against a set of pre-defined success metrics.

Definition and measurement of implementation strength

The evaluation included a detailed documentation of implementation process and context through document review and key informant interviews, to facilitate interpretation and attribution of study outcomes. At the outset, this component was intended to be primarily qualitative, but during the analysis phase the utility of having quantitative measures of implementation intensity was recognized, both to summarize implementation experience and provide comparable estimates across countries.

The evaluators derived quantitative measures of implementation at the country level based on duration and on implementation of supporting interventions at the time of endline data collection (Willey *et al.* 2014). Duration was felt to be particularly important as the endline evaluation was conducted only 6–15 months after implementation began. The implementation strength measures comprised number of months for which subsidized ACT was available in-country (which ranged from 7 to 15.5 months), months for which communication campaigns were implemented (0–9 months), disbursements for supporting interventions (0.03–0.42 USD per capita) and the proportion of private for-profit providers surveyed at endline who reported being trained on antimalarials with the AMFm symbol (2–50%) (Willey *et al.* 2014). Other potential facets of implementation strength were considered, including the role of stakeholders in facilitating the operation of the AMFm order system and the quality of communications and training, but were less amenable to measurement.

Association of implementation strength with outcomes

There was clear variation in the success of countries in achieving pre-specified benchmarks (Tougher *et al.* 2012). In summary, the largest changes were seen in Ghana and Kenya, followed by Zanzibar and Tanzania mainland. Performance in Nigeria and Uganda was not consistent across outcomes, and limited impact was observed in Niger and Madagascar.

Unlike in the other examples, Tougher *et al.* (2012) made no attempt to do formal statistical analysis of the association between the measure of implementation strength and outcomes. However, the distribution of the implementation strength broadly mirrored the relative performance of the pilots in terms of AMFm outcomes. Ghana and Kenya were considered to have had the strongest implementation, followed by Tanzania mainland, Zanzibar and Nigeria. Niger, Madagascar and Uganda were considered to have the lowest intensity of implementation. This also fed into an assessment of whether changes in outcomes could be attributed to AMFm, as the implementation strength analysis increased plausibility that the large changes seen in some countries were attributable to AMFm.

Discussion

Measuring the strength of implementation of a programme and correlating this with health or other outcomes is a promising approach to pragmatic impact evaluation and is increasingly being used, but methodological guidance is needed. We have described three examples in which measures of implementation strength have been used. In this discussion, we reflect on five aspects of the development and use of these measures and identify some of the potential strengths and weaknesses of the approaches taken.

We suggest that the first step in developing a measure of implementation strength should be to develop an appropriate logic model for the intervention. To illustrate our examples we have described our own simplified logic models for the interventions presented, developed from information provided in the articles. In practice these models would be much more detailed, and the authors of each of our cited examples would have planned their work on the basis of such models. Approaches such as theory of change (De Silva *et al.* 2014) may also be helpful and more amenable to reflecting complex feedback loops, as can consideration of health system building blocks and how they inter-relate in considering inputs and processes (WHO 2007; Adam and de Savigny 2012). We have found the approximate separation of 'implementer-controlled' inputs and processes from the effects of these on services and target populations (outputs, outcomes, impacts) to be especially helpful in this work (Institute for International Programmes 2011).

The second step is to identify those aspects of implementation to be assessed, guided by the logic model. For example, in Avahan a simple, distal measure (dollars spent per HIV-positive person) was used as the implementation strength variable. Money spent has both advantages and disadvantages as a measure of implementation strength. Advantages include that it is relatively simple to assess and should be calculable from programme records. Investigators do not need to make decisions about how to combine information from diverse programme components. Money spent is well aligned to questions of cost-effectiveness and value for money that is primary concerns for policy-makers.

One disadvantage is that money alone does not include information on what the money was spent on, and whether it leads to improvements in service quality, accessibility or acceptability. The Avahan model was based on innovation, efficiency and context-specific tailoring of the delivery of a core package of services. In this context, expenditure may not characterize the intended intensity of appropriate services. Without other data, one cannot answer questions of scalability of different components of the intervention package, nor assess which components are likely to be the most important for effectiveness. Evaluations using money spent as a measure of implementation strength will benefit from additional study of processes or outputs, otherwise the evaluation runs the risk of having a 'black box' approach. It should be noted that other evaluations of the Avahan programme have developed measures of processes and outputs (Verma *et al.* 2010; Thilakavathi *et al.* 2011).

Another disadvantage is that low-intensity spending might partly reflect the availability of services from other providers. Avahan was the not sole provider in many of the districts where it was delivered, Government services were also available. The estimated 'effect' of increasing the intensity of spending might reflect the effect of increased Avahan spending where Avahan is already delivered, or might reflect the replacement of Government services with Avahan. The former (a dose-response relationship) and the latter (the effect relative to Government services) could both support an effect of Avahan, but the evaluation included the two effects in the overall effect estimate making the results hard to interpret.

In contrast to Avahan, the measures for AMFm focused on a mix of 'inputs' and 'processes', and included the duration interventions were in place (time since first arrival of subsidized drugs in-country and duration of communications campaign), the money spent on supporting interventions and a self-reported, provider-level measure of exposure to AMFm-related training. The AMFm evaluation did not include expenditure as a measure of implementation strength, although expenditure on ACT subsidies can be considered a key input (Figure 3). This reflected the demand driven nature of the intervention—money in the co-payment fund was only spent when importers from participating countries made orders, with such orders depending on their perceptions of in-country demand. The amount of subsidized ACT ordered by each country varied considerably, with the number of doses delivered per capita during the study period ranging from 0.08 in Madagascar to 1.01 in Ghana. Order quantities were not under the control of the implementing agencies and could thus be seen as a measure of performance of the intervention in stimulating demand from importers and in the smooth running of the ordering system. Similar challenges with using expenditure as a measure of implementation strength would arise in other interventions that have a demand-creation component. In the case of Avahan it is possible that weaknesses in the services delivered in a certain district would encourage increased funding, especially since a large part of the innovation of the Avahan programme was to have efficient real-time monitoring. Alternatively, efficient contractors may have been awarded additional funds as the quality of their service delivery was evident in the monitoring.

The third step is to design and implement an appropriate data-collection methodology to capture the dimension(s) of implementation identified for measurement. It is important to consider the appropriate level at which to measure (e.g. household, national), as well as the most appropriate point in time for measurement (e.g. multiple time points, baseline and endline only). Data-collection approaches will inevitably draw on a range of data sources. These might include accounting systems, quality assurance or monitoring and evaluation (M&E) data, programme records, or surveys of providers. Data from intended beneficiaries may also be included, but we caution that such measures may reflect not just implementation strength but also the successful coverage on the ground of such interventions, which may in fact be a step further down the logic model. The method used to avoid such bias in the Ethiopia Health Extension Worker study was to measure implementation strength at household level excluding the households of intended beneficiaries. For example, large resources and effort might be spent on establishing a complex *in situ* training regimen for health providers. In some places this may be implemented such that the number of providers trained is high, while in other areas inefficiencies may mean the number trained is low. In spirit a measure of implementation strength may be more concerned with capturing the effort expended than its success in achieving intended outputs (trained providers) since this is itself a question of interest. In some situations, qualitative or participatory data collection approaches might also be useful for assessing implementation strength. Lessons learnt in the AMFm evaluation included the importance of understanding context which had a major role in determining relative performance. This was underlined by the value of qualitative methods in understanding both process and context alongside quantitative process measures to enable ranking by implementation strength.

In the fourth step, decisions must be made about whether and how to use data on different aspects of implementation within a single measure of implementation strength. In the Avahan example a single variable, money per HIV-positive person, was assessed. However, developing this into a district-level measure of

implementation strength implicitly took account both of the length of time the interventions were in place (since the total amount of money spent in district was presumably influenced by the duration of intervention) and the size of the target population. In the Ethiopian example the implementation strength measure included four different aspects of outreach work: two different types of home visits, family health card ownership and the proportion of households with 'model family' status. These reflect a variety of aspects of the programme, and a single score was derived to reflect overall strength. In AMFm the approach was also multi-dimensional, including data on money, time and proportion of the target population trained. However, the AMFm study did not combine these measures into a single index. We advise caution when combining components from different areas of the logic model. Where indicators are combined, choices will have to be made about the method used to combine data, and implicitly or explicitly this will require consideration of weighting. We suggest that these decisions should again ideally be driven by theoretical concerns, e.g. the components hypothesized to be most important, or most expensive might have greater weights. Data-reduction approaches such as principal components analysis may therefore be appropriate but should be approached with care, to ensure the measure is readily understood. One limitation in the measures developed for the Ethiopian programme was that the composite index used to estimate implementation strength was complex and unlikely to be have been readily understood by front-line implementers. A different approach might be needed to reach decisions on which aspects of the programme required strengthening. Other options include review and consensus, which will likely result in a measure with poorer mathematical and statistical properties but better ownership and understanding from programme staff. Once a measure is developed, some basic analysis of its properties, calibration and cross-validation will be necessary. Common-sense checks that the developed measures correspond to realities on the ground in a small number of locations are essential, and may want to be conducted during the course of the study as well as at the end.

The fifth and final step is that decisions must be made about how measures are used in analysis, ideally specified a priori in a statistical analysis plan for the evaluation. This is beyond the remit of this article, but it is worth noting that, using our examples, in Avahan implementation strength was used as a primary exposure measure, while in AMFm and the Ethiopian HEW case studies the measure was used as supporting information to supplement the primary analysis. All three examples suggest that while development of a robust measure of implementation strength had great value, it does not address the potential for confounding when such measures are correlated with health outcomes. The risk of bias from confounding in impact evaluations will need to be judged on a case-by-case basis. In relation to the Ethiopian impact evaluation, the study found strong evidence of a dose-response relationship between implementation strength and better newborn care. However, other factors or programmes affecting the outcomes may have varied over time as well as between kebeles, and without adjustment for these factors the results could be biased in either direction. Contextual factors also influenced AMFm outcomes, and also explained implementation strength itself. For example, the relatively poor implementation and performance in Niger and Madagascar reflected an unfavourable political and economic context, and the nature of the retail antimalarial market which was heavily dominated by unregistered vendors which the authorities did not want to encourage to stock ACT. Given these challenging contexts it is unclear whether AMFm would have been effective in these pilots even if supporting interventions had been fully implemented. Low implementation strength and poor outcomes may be correlated because they both reflect limited local capacity, rather than because of a causal link between implementation strength and outcome.

Conclusion

Evaluation of public health strategies seeking to ensure that evidence-based, low-cost interventions reach those in need is critical. Robust measures of implementation strength for such strategies can be a very useful component of an evaluation strategy, both to assess impact and to identify which aspects of a programme need to be strengthened. By reflecting on three approaches we have identified five critical issues to be considered in guiding the development of these measures in future evaluations.

Acknowledgements

This article forms part of the work of the LSHTM Centre for Evaluation, which aims to improve the design and conduct of public health evaluations through the development, application and dissemination of rigorous methods, and to facilitate the use of robust evidence to inform policy and practice decisions [http://evaluation.lshtm.ac.uk/]. Additionally, JS and BA contributed to this through IDEAS (http://ideas.lshtm.ac.uk/). The authors are grateful to Gina Dallabetta who provided useful comments on the manuscript.

Funding

This work was funded by Wellcome Trust Award 105609/Z/14/Z to the LSHTM Centre for Evaluation; and by the IDEAS grant to LSHTM, which is funded by the Bill & Melinda Gates Foundation.

Conflict of interest statement. None declared.

References

Evaluation: the top priority for global health. 2010. *Lancet* **375**: 526 doi:10.1016/s0140-6736(10)60056-6.

Adam T, de Savigny D. 2012. Systems thinking for strengthening health systems in LMICs: need for a paradigm shift. *Health Policy and Planning* **27**: iv1–3.

Bryce J, Requejo JH, Moulton LH, Ram M, Black RE, Population Health I *et al*. 2013. A common evaluation framework for the African Health Initiative. *BMC Health Services Research* **13**: S10 doi:10.1186/1472-6963-13-s2-s10.

Bryce J, Victora CG, Conference Organizing G. 2005. Child survival: countdown to 2015. *Lancet* **365**: 2153–4. doi:10.1016/s0140-6736(05)66752-9.

Carroll C, Patterson M, Wood S *et al*. 2007. A conceptual framework for implementation fidelity. *Implementation Science* **2**: 40 doi:10.1186/1748-5908-2-40.

Countdown to 2015. *Accountability for Maternal, Newborn and Child Survival: The 2013 Update*. Geneva: WHO, 2013.

De Silva MJ, Breuer E, Lee L *et al*. 2014. Theory of Change: a theory-driven approach to enhance the Medical Research Council's framework for complex interventions. *Trials* **15**: 267 doi:10.1186/1745-6215-15-267.

Habicht J-P, Victora C, Vaughan JP. 1999. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *International Journal of Epidemiology* **28**: 10–8.

Horton R, Murray C, Frenk J. 2008. A new initiative and invitation for health monitoring, tracking, and evaluation. *The Lancet* **371**: 1139–40.

Institute for International Programmes. Measuring the strength of implementation of community case management of childhood illness within the Catalytic Initiative to Save a Million Lives. Working Paper. John Hopkins School of Public Health, Baltimore, 2011. http://www.jhsph.edu/research/centers-and-institutes/institute-for-international-programs/_documents/rapid_scaleup/wp-implementation-strength.pdf

Karim AM, Admassu K, Schellenberg J *et al*. 2013. Effect of ethiopia's health extension program on maternal and newborn health care practices in 101 rural districts: a dose-response study. *PloS One* **8**: e65160 doi:10.1371/journal.pone.0065160.

Mills A. 2014. Health care systems in low-and middle-income countries. *New England Journal of Medicine* **370**: 552–7.

Ng M, Gakidou E, Levin-Rector A *et al*. 2011. Assessment of population-level effect of Avahan, an HIV-prevention initiative in India. *Lancet* **378**: 1643–52. doi:10.1016/s0140-6736(11)61390-1.

Pickles M, Boily MC, Vickerman P *et al*. 2013. Assessment of the population-level effectiveness of the Avahan HIV-prevention programme in South India: a preplanned, causal-pathway-based modelling analysis. *The Lancet Global Health* **1**: e289–99. doi:10.1016/s2214-109x(13)70083-4.

Ramakrishnan L, Gautam A, Goswami P *et al*. 2010. Programme coverage, condom use and STI treatment among FSWs in a large-scale HIV prevention programme: results from cross-sectional surveys in 22 districts in southern India. *Sexually Transmitted Infections* **86**: i62–8. doi:10.1136/sti.2009.038760.

Schellenberg JA, Bobrova N, Avan BI. 2012. Measuring implementation strength: literature review draft report. Working paper, London School of Hygiene and Tropical Medicine, 2012. http://ideas.lshtm.ac.uk/sites/ideas.lshtm.ac.uk/files/Report_implementation_strength_Final_0.pdf.

The Partnership for Maternal NCH. *A Global Review of the Key Interventions Related to Reproductive, Maternal, Newborn and Child Health (RMNCH)*. Geneva, Switzerland: PMNCH, 2011.

Thilakavathi S, Boopathi K, Girish Kumar CP *et al*. 2011. Assessment of the scale, coverage and outcomes of the Avahan HIV prevention program for female sex workers in Tamil Nadu, India: is there evidence of an effect? *BMC Public Health* **11**: S3 doi:10.1186/1471-2458-11-s6-s3.

Tougher S, Group AC, Ye Y *et al*. 2012. Effect of the Affordable Medicines Facility–malaria (AMFm) on the availability, price, and market share of quality-assured artemisinin-based combination therapies in seven countries: a before-and-after analysis of outlet survey data. *Lancet* **380**: 1916–26. doi:10.1016/s0140-6736(12)61732-2.

Verma R, Shekhar A, Khobragade S *et al*. 2010. 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. *Sexually Transmitted Infections* **86**: i76–82.

Victora CG, Black RE, Boerma JT, Bryce J. 2011. Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations. *The Lancet* **377**: 85–95.

WHO. *Everybody's business: Strengthening health systems to improve health outcomes*. Geneva: WHO's framework for action 2007.

Willey BA, Tougher S, Ye Y, ActwatchGroup *et al*. 2014. Communicating the AMFm message: exploring the effect of communication and training interventions on private for-profit provider awareness and knowledge related to a multi-country anti-malarial subsidy intervention. *Malaria Journal* **13**: 46. doi:10.1186/1475-2875-13-46.