

Measuring impact in the Millennium Development Goal era and beyond: a new approach to large-scale effectiveness evaluations



Cesar G Victora, Robert E Black, J Ties Boerma, Jennifer Bryce

Evaluation of large-scale programmes and initiatives aimed at improvement of health in countries of low and middle income needs a new approach. Traditional designs, which compare areas with and without a given programme, are no longer relevant at a time when many programmes are being scaled up in virtually every district in the world. We propose an evolution in evaluation design, a national platform approach that: uses the district as the unit of design and analysis; is based on continuous monitoring of different levels of indicators; gathers additional data before, during, and after the period to be assessed by multiple methods; uses several analytical techniques to deal with various data gaps and biases; and includes interim and summative evaluation analyses. This new approach will promote country ownership, transparency, and donor coordination while providing a rigorous comparison of the cost-effectiveness of different scale-up approaches.

Introduction

The Millennium Development Goals (MDGs) have stimulated interest and increased funding for programmes aimed at reduction of maternal and child mortality and the burden of HIV/AIDS, tuberculosis, and malaria. At the same time, the realisation that few programmes and initiatives have been evaluated properly,¹⁻³ and interest in results-based financing approaches,⁴ are increasing pressure on implementers to undertake effectiveness evaluations.

A common evaluation framework is needed to allow future comparison of the performance of different initiatives—measured in terms of increasing coverage and achieving health effects—and their cost.⁵ Such a framework should include: (1) a conceptual model outlining pathways through which the initiative is expected to affect the MDGs; (2) a list of standard indicators of inputs, processes, outputs, outcomes, and impact, with clear measurement plans; and (3) guidelines for design of evaluations in a compatible way. Much progress has already been made on the first two objectives.⁶ This Health Policy article focuses on the third topic, reporting on evaluations of large-scale public-health programmes.

Our objective is not to establish the efficacy of new biological or behavioural interventions, or of packages of such interventions; these aims are best achieved with randomised controlled trials. We are interested in assessment of how well large-scale complex programmes deliver efficacious interventions, using different delivery channels in various health-system contexts.

Programme success is defined as gains in intervention coverage and in health effects under real-world conditions, when implementation tends to be less intense and more variable than in efficacy trials.⁷ Observational designs are needed because evaluators cannot control where, when, and how rapidly programmes will be implemented at scale by governments, international or bilateral agencies, and private voluntary organisations.

Design of health-programme evaluations has been dominated traditionally by experimental approaches used in medicine, in which specific individuals or clusters of people receive an intervention whereas others do not. Studies tend to be undertaken in controlled environments in which the influence of external factors is kept to a minimum or eliminated. In the real world, however, the intervention or programme of interest usually accounts for only a small part of variability in health outcomes. Figure 1 presents a simplified framework showing that maternal and child health outcomes can also be affected by socioeconomic and contextual factors, by changes in existing health services in the public and private sectors that are outside the scope of the programme of interest, and by other initiatives or interventions in health or other sectors present in the same geographic areas. Because changes in all the above factors can be happening concurrently with implementation of the programme under assessment, real-world effectiveness evaluations present challenges that cannot be properly addressed by the traditional approach of intervention versus comparison group. In addition to the reality that programmes are not scaled up in a vacuum, they also rarely start from a blank sheet. Pre-existing baseline levels and, particularly, trends in key indicators need to be taken into account.

We start by describing typical approaches to evaluation of large-scale programmes and move on to propose a new approach that addresses the mosaic of concurrent programmes and initiatives characteristic of most low-income countries with high rates of maternal and child mortality. We draw heavily on an evaluation study published in *The Lancet*,⁸ the Multi-Country Evaluation of Integrated Management of Childhood Illness (MCE-IMCI),^{9,10} and subsequent efforts to develop designs for independent assessments of the Catalytic Initiative to Save a Million Lives,¹¹ including the three-country rapid scale up of maternal, newborn, and child health funded by the Bill & Melinda Gates Foundation via WHO (the Partnership for Maternal, Newborn, and Child Health).

Published Online

July 9, 2010

DOI:10.1016/S0140-

6736(10)60810-0

Post-Graduate Program in Epidemiology, Federal University of Pelotas, Pelotas, Brazil (Prof C G Victora MD); Institute for International Programs, Bloomberg School of Public Health, Johns Hopkins University, Baltimore, MD, USA

(Prof C G Victora,

Prof R E Black MD, J Bryce EdD);

and World Health

Organization, Geneva,

Switzerland (J T Boerma MD)

Correspondence to:

Prof Cesar Victora, Federal

University of Pelotas, Pelotas,

Brazil

cvictora@terra.com.br

For the Partnership for Maternal, Newborn, and Child Health see <http://www.who.int/pmnch/en/>

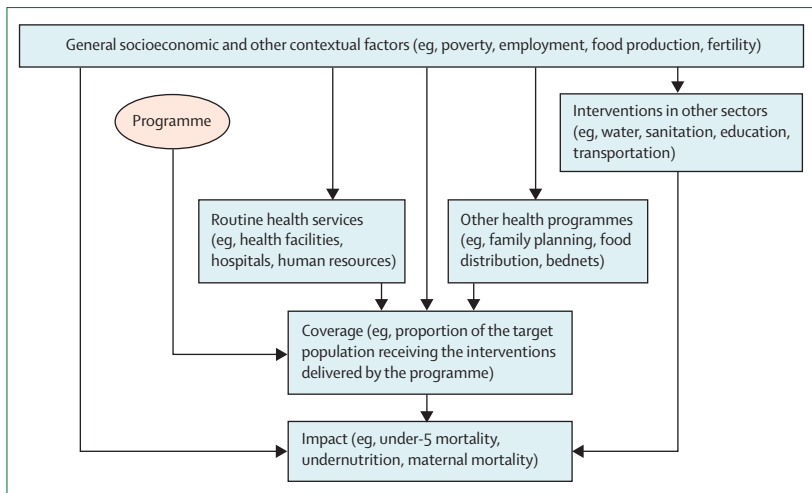


Figure 1: Outline of factors affecting maternal and child health and nutrition

Common approaches to large-scale evaluations Before and after assessment

Programme areas only

This evaluation design is one of the simplest and entails recording changes over time in the intervention area, in terms of indicators of health service provision, utilisation, coverage, and impact. These evaluations usually attempt to note goals in terms of coverage (eg, 80% of children aged 6–59 months receiving two doses of vitamin A in the previous 12 months) or impact (eg, a 25% reduction in mortality over a specific period), and have been described as adequacy designs, because evaluators focus on assessment of how well the stated programme goals are met.⁷

The counterfactual for this type of evaluation is that changes would not have happened without the programme, that is, indicators would not have improved because of pre-existing routine health services, other programmes in health or related sectors, or in broader determinants of health—eg, socioeconomic or environmental factors (figure 1). The major limitation of before and after designs is that, without a comparison group, there is no way to determine whether changes are attributable to the programme or to other factors.

This type of evaluation continues to be popular despite its limitations, particularly with implementers undertaking internal assessments of their own programmes. Such studies are better than no assessment, if results are interpreted with the necessary caveats. They are especially useful when no changes are found in coverage or impact indicators, which suggests that the programme needs to be reformulated or (rarely) managed to offset the negative effect of other determinants of health, such as natural or man-made crises.

This design can be extended in several ways. First, if data are available on preintervention trends, the counterfactual is continuation of these trends. Favourable changes in indicators after initiation of interventions are an important

evaluation finding. Second, in case of several units of analysis, such as districts, provinces, or countries, a dose–response analysis that examines (for instance) the association between funding levels and service coverage could further contribute to the evaluation.

Comparison of programme and non-programme areas

This design has been used frequently in rigorous programme evaluations. It allows investigators to increase plausibility of a causal effect by inclusion of an external comparison to assess whether trends in coverage and impact indicators differ in programme and non-programme areas. A variant of this approach is the stepped-wedge design,¹² in which the new programme is implemented gradually, ideally with random allocation, and areas without the programme constitute the comparison group up until the time they start implementation.

The counterfactual is that programme areas, had the programme not been implemented, would show trends similar to those seen in non-programme areas. Returning to figure 1, this design assumes that background conditions (eg, socioeconomic and environmental factors), levels of existing health services (eg, number of health facilities or staffing patterns), other health programmes (eg, mass immunisation campaigns), or non-health-sector initiatives (eg, water and sanitation) would evolve similarly over time in both programme and non-programme areas. Four scenarios highlight the limitations of this design in view of the current public-health landscape in most low-income countries. These scenarios are illustrative and not mutually exclusive.

First, the untouched comparison area scenario assumes that comparison areas do not have the programme under evaluation, and that neither intervention nor comparison areas have other health programmes affecting intervention coverage or mortality. Intervention and comparison areas are assumed to be similar in terms of public and private health services and of non-health-sector interventions, and with respect to the broader determinants of health. Of the four scenarios presented here, this one is the closest to the experimental approach, but it is increasingly uncommon because many programmes are now being implemented in most if not all districts in low-income countries. So-called virgin comparison districts, if they ever existed, are unlikely to exist now. Furthermore, global health partnerships—such as the Global Fund to Fight AIDS, Tuberculosis and Malaria (Global Fund) and the Global Alliance for Vaccines and Immunization (GAVI)—generally support programmes with national coverage.

Second, the parallel health programmes scenario takes into account that health programmes other than the one being assessed—but addressing the same causes of morbidity and mortality—are underway in both intervention and comparison districts. An example is evaluation of IMCI in Tanzania.¹³ The main component of this

intervention was improved case-management of malaria, pneumonia, and diarrhoea in government health facilities. At the same time, other programmes were present in both intervention and comparison areas, including mass distribution of vitamin A and social marketing of insecticide-treated nets. The IMCI evaluation measured coverage of these parallel programmes and took them into account when interpreting impact results, but complete separation of effects was impossible because vitamin A and insecticide-treated nets were also being promoted by IMCI.

Third, the rapid socioeconomic progress scenario builds on the previous setting but also recognises that health outcomes are affected strongly by socioeconomic progress. Before introduction of efficacious biological interventions currently promoted in countries of low and middle income, the striking decline of under-5 mortality in high-income countries in the first half of the 20th century attested to the importance of economic progress, nutrition, sanitation, and improved living conditions.¹⁴ In today's world, socioeconomic progress might boost child survival in countries of low and middle income not only through the pathways depicted in figure 1 but also by increasing access to antibiotics and services provided by the private sector, improved transportation (eg, motorcycles and better roads), and communications (eg, mobile phones), all of which can result in better access to health services. A rigorous before-and-after design with a comparison group will measure socioeconomic and other contextual factors at the beginning and end of the evaluation, identify imbalances between the two groups, and attempt to take these into account when analysing results.

However, socioeconomic progress and mortality declines can be so rapid that an additional effect of the programme under evaluation might be impossible to detect. The Bangladesh IMCI evaluation provides a good example.¹⁰ This study used a cluster-randomised design with intervention and comparison areas to show annual declines in under-5 mortality of 8.6% and 7.8%, respectively. Such massive rates of decrease were not foreseen when the study was designed in 1999, because before that time annual rates of mortality decline were less than 2%. In addition to socioeconomic progress, Bangladesh has wide availability of lifesaving drugs in a thriving private sector, and strong vertical programmes (eg, oral rehydration therapy, vitamin A, vaccination, etc) that were present in both IMCI and comparison areas. This example shows how the effect of a particular intervention can be difficult to detect with rapid socioeconomic and general health service improvement.

Contextual epidemiological changes can also affect evaluation. An example is assessment of a set of preventive HIV interventions in rural Zimbabwe during a period of rapid secular decline in HIV prevalence and incidence.¹⁵ The secular decrease could have been related to general behavioural changes or a phase of

reduced HIV transmission in the epidemic, because populations at highest risk have died. Under these circumstances, a carefully planned local intervention study is much less likely to be able to record a significant effect on HIV transmission.¹⁶ Contextual epidemiological changes can also affect evaluations for other diseases such as malaria, which is affected by rainfall patterns, and measles or meningitis, which have multi-year cyclic patterns of transmission.

Finally, the Balkanisation scenario results from the fact that various development partners and international organisations agree—usually with government consent—to support programmes in specific provinces or districts within a country. This process of Balkanisation means that different partners promote and support similar interventions in different geographic areas, sometimes with variable approaches to intervention delivery. Comparison of districts receiving a given programme with either the rest of the country or with another set of neighbouring districts could end up comparing very similar sets of interventions. This situation arose with evaluation of the UNICEF Accelerated Child Survival and Development (ACSD) programme in Mali, which showed a 24% fall in under-5 mortality during the intervention period in the ACSD focus districts and a 31% decline during the same period in the national comparison area.⁸ Documentation of programme implementation and contextual factors showed that other funding agencies (mainly the US Agency for International Development [USAID]) were supporting scale-up of some of the same maternal and child survival interventions as ACSD in many comparison area districts, suggesting that the presence of ACSD had displaced another programme that was at least as effective, or possibly more so. Because an evaluation focused on one programme almost never gathers information on another, an intervention-comparison design might fail to detect that comparison areas were, in fact, profiting from a more effective programme.

A way forward: the evaluation platform design

In the preceding section, we addressed limitations of intervention-comparison designs in contexts for which many programmes and interventions are being scaled up simultaneously. We argued that a reductionist approach to evaluation based on isolation of programme effects is no longer appropriate for scaling up of initiatives to reach the MDGs in most low-income countries.

What is the alternative? How can evaluations of large-scale programmes be designed in ways that are scientifically rigorous and yet reflect the real-world context? First, one must stop thinking about comparison of a few programme districts (or populations) with a handful of others that do not have that particular programme—because the comparators are likely to have, to some degree, initiatives similar to those under scrutiny. Second, one must try to understand why certain



Figure 2: Focus districts for selected development partners in maternal and child health, Mozambique, 2008

programmes are implemented in some areas rather than others, because this reason could affect how likely they are to succeed. Finally, one must be able to answer one of the most important questions of all: which of the various programmes or delivery approaches implemented by different partners works best in a given country? An improved approach to evaluation will build heavily on existing monitoring data and complement them as needed, with thorough quality checks, additional data collection, and enhanced data analyses.

Figure 2 shows the geographic distribution by province of health programmes supported by partners in Mozambique in 2008. Most of these programmes cover a few districts within every province. Moreover, several initiatives—such as the Global Fund, GAVI,

Stop TB, the US President’s Emergency Plan for AIDS Relief (PEPFAR), etc—attempt to cover the whole country but are implemented with variable intensity in each district or province. All these programmes merit independent evaluation to inform the Mozambican people and programme funders about their effectiveness and, where possible, value for money. Yet, in a review of several evaluations, the effectiveness of most health programmes is not assessed at all, and even when evaluations are undertaken they generally lack scientific rigor.² In the rare event that more than one programme in the same country is assessed rigorously, results of the different evaluations can be difficult to compare because the same indicators were not used or methods were inconsistent.

The new evaluation platform proposed here has several features. First, it uses the district as the unit of design and analysis. Second, it is based on continuous monitoring of the different levels of indicators. Third, additional data are gathered before, during, and after the period to be evaluated by various methods. Fourth, a range of analytical techniques are used to deal with data gaps and biases. Finally, interim and summative evaluation analyses are undertaken. Panel 1 provides details of an evaluation platform approach that was developed for Mozambique.

Discussion on the design and applications of the evaluation platform is focused on countries of low and middle income, in contexts for which several agencies and partners are likely to be active in different regions of the country. The main concepts of the platform design are also relevant to developed countries, and indeed have already been incorporated in ecological studies of programme impact on the basis of existing data (eg, vital statistics).

Districts as units

The preferred unit of study design and analysis for evaluations is usually the district, because this is the core administrative unit for government health and other programmes in many countries. Districts are easily identifiable geographically and typically have some level of sociocultural or economic homogeneity. Population sizes vary from 100 000 to 500 000 in most nations. Some countries have subdistricts as the smallest administrative units and these might serve as units for the evaluation. Data for a range of indicators from programme inputs to service delivery and coverage can be gathered at the district level.

Several limitations to this idea exist. Use of services does not follow district boundaries strictly, because people can cross boundaries to obtain the nearest or better services. Some countries redefine district boundaries periodically. District sizes vary widely by country and, in most cases, aggregation of districts is necessary to obtain adequate numbers. Finally, health status indicators, and some coverage indicators, might

Panel 1: Planning the evaluation platform design in Mozambique

Mozambique provides us with examples of practical opportunities and challenges of the platform design. Many agencies are present in different parts of the country (figure 2), and several existing databases can provide data at district and provincial level that are useful for evaluation.

The platform design can be used to evaluate different initiatives. The reach every district (RED) approach, coordinated by the Ministry of Health, started implementation of a selected set of maternal and child interventions in 33 districts in 2008, with support from different partners. These interventions include longlasting insecticide-treated mosquito nets, immunisations, breastfeeding promotion, vitamin A supplementation, and integrated management of childhood illnesses in health facilities. New cohorts of 33 districts will start implementation every year until all 148 districts are covered, allowing a non-randomised stepped-wedge analysis to be done, building on the evaluation platform, which will include baseline data for contextual factors, coverage, and mortality levels and will allow analyses of coverage and health impact outcomes relative to programme implementation strength.

A separate initiative, the rapid scale-up of the partnership for maternal, newborn, and child health, is concentrating its efforts in 12 districts. Since information on all districts in the country will be gathered, the platform approach could be used to evaluate the two programmes described above and other existing or future initiatives.

To complement official data available at national level, focal points will be hired in each of the 11 provinces (each with about 13 districts) to report monthly on changes in health facilities, training of facility and community health workers, supervision, and staffing patterns. They will also report on distribution of commodities (eg, insecticide-treated nets, drugs, vaccines, nutrition supplements, etc), provided by governmental and non-governmental sources. For quality control, district-level information obtained by the provincial focal points will be cross-checked with official data provided by sources at national level and with denominator data (eg, population size); discrepancies will be identified and discussed with health managers to contribute to data quality improvement over time.

Panel 2 shows an initial list of available variables from different data sources in Mozambique. The proposed network of informants at provincial and district level will help update information on contextual factors in the database and report unpredicted events, such as food shortages or natural or man-made emergencies that could affect maternal and child health.

(Continues in next column)

(Continued from previous column)

Coverage data will be obtained from existing and future surveys. The 2003 demographic and health survey (DHS) and 2008 multiple indicator cluster survey included more than 12 000 and 14 000 households, respectively, in their samples. In 2008, over 3000 households belonged to the 33 districts included in the first wave of the RED initiative. This number will provide a sufficient sample for precise measurement of intervention coverage for all key maternal and child health interventions. A new DHS is planned for 2011, which can provide additional coverage information and allow measurement of changes.

A major challenge for the evaluation platform in Mozambique is to strengthen local capacity for data collection, management, and analyses. Another, perhaps even greater, challenge is to ensure cooperation and support of the many local and international partners with a role in implementation of health programmes.

only be available or sufficiently precise at higher levels of aggregation, such as provinces or subnational regions.

The manner in which districts are selected to receive the programme under evaluation needs to be taken into account. Documentation of criteria that guided this selection is essential.

Monitoring and databases

Evaluation needs a continuous, strong monitoring effort that entails careful and systematic documentation of contextual variables, health system inputs and service delivery, intervention coverage, risk factors, and health status. This procedure requires strengthening of country efforts to build district health information systems.

Most low-income and middle-income countries already have several databases maintained by governmental, international, or partner institutions (panel 2). These might include data for health inputs (financing, human resources), health facility reports (services provided), facility assessments (geocodes, drug availability, etc), the socioeconomic and demographic situation (including poverty maps), and partner presence. These databases contain information disaggregated at provincial or state level, and sometimes at district level.

We propose that relevant information from different existing databases should be integrated in a continuous manner to lay the foundation for evaluations. The main data repository for the evaluation platform is a database in which geographic units constitute the rows, and indicators that are relevant to the evaluation make up the columns—one for every year for which data on that indicator are available. New information about programme implementation by different agencies (government, bilaterals, multilaterals, non-governmental organisations [NGOs]) would also be added. In most countries, the evaluation platform would include

Panel 2: Examples of data available from existing databases at district and provincial level in Mozambique, and sources of information

Socioeconomic factors

- Household assets
- Family income and poverty
- Parental education and occupation
- Unemployment
- Land tenure

Sources: 2007 census; economic censuses and surveys

- Economic crises (inflation rates, crop failures, floods)

Source: National Institute of Statistics

Demographic factors

- Population density
- Fertility patterns
- Family size
- Ethnic groups

Source: 2007 census

Environmental characteristics

- Water supply
- Sanitation
- Urbanisation
- Housing

Source: 2007 census

- Rainfall
- Altitude

Source: National Meteorological Institute

Baseline health characteristics

- Under-5 mortality
- Prevalence of malnutrition
- HIV prevalence
- Malaria transmission patterns

Sources: 2007 census; 2008 MICS; malaria and HIV surveys

Health services characteristics

- Availability of health services (hospitals, clinics, etc) in governmental and private sectors
- Population/facility ratio
- Health worker staffing patterns
- Health worker pay
- Drug supply
- Baseline utilisation rates
- Availability of referral services
- Strength of district health management team
- District health budget (overall and for child health)

Sources: Health Metrics Network; Ministry of Health Information Systems; UNFPA Needs Assessment Survey; WHO Service Availability Mapping

(Continues in next column)

(Continued from previous column)

Presence of other projects and programmes that could affect health status

- Micronutrients
- Indoor residual spraying
- Immunisations
- HIV programmes
- Others

Sources: UNICEF; WHO; NAIMA; ODAMOZ

MICS= multiple indicator cluster survey. UNFPA= UN Population Fund. NAIMA= Network of Organizations working in Health and HIV/AIDS. ODAMOZ= Official Development Assistance to Mozambique Database.

programmes are concentrated, particularly in the case of large countries with striking regional disparities. For example, a platform-like approach was used in northeast Brazil to evaluate the effectiveness of IMCI.^{17,18}

The evaluation platform will not replace existing databases, but instead will build on and contribute to their improvement. It could be described as a comprehensive database that covers information on all relevant monitoring indicators, pays systematic attention to data quality, and includes contextual and qualitative information to answer evaluation questions.

Additional data collection

Although solid monitoring systems provide essential information on trends and indicators, the platform design will need additional data collection to record programme implementation, assess data quality and make statistical adjustments, fill data gaps, and address specific evaluation questions. Various methods might be required, such as health-facility assessments, household surveys and oversampling of districts, longitudinal designs, and qualitative research. The prospective design of the platform provides greater opportunities to gather essential additional data than retrospective studies, when only post scale-up data can be obtained.⁸

Detailed documentation of inputs, training, supervision, quality of care, and delivery channels is essential for understanding why programmes succeed or fail. This component of health evaluations is frequently neglected because efforts to measure coverage or mortality consume most of the assessors' time and resources. Data for implementation can be obtained at the national level (eg, Ministry of Health, international agencies, or NGOs supporting the programme), but records of what is actually reaching the population at district level are also important. For example, the IMCI evaluation showed that even simple data for training coverage are sometimes hard to obtain. The numerator—how many health workers were trained in every district—is usually known, but to find out whether workers remained in their posts after training is difficult, because turnover tends to be high in many countries. Even the denominator for training coverage—the total number of health workers in

information about all districts, thus allowing assessors to understand how programmes are being deployed. However, the platform can also be restricted to a subnational region in which health and development

the cadres eligible for training—can be difficult to ascertain at national level.¹⁹ In the planned platform design in Mozambique (panel 1), key informants at provincial level will record programme implementation and check the quality of information available at national level. Documentation of all programmes being implemented in every district permits researchers (with this platform design) to test the possibility that a given initiative could have displaced a more effective one, as discussed in the Mali ACSD example.⁸

Programme performance scores can provide a quantitative measure of implementation strength on the basis of information gathered through the documentation process. For example, the score for a programme promoting community case-management of pneumonia could be derived from the number of community health workers trained per population, availability of drugs, and the intensity of supervision. These scores can be used to measure the so-called dose of a programme in dose–response analyses.

Documentation of costs associated with programmes is also important and can be done by obtaining accurate information on government investments and those by international and bilateral agencies at country level. Donor atlases are available in some nations and can contribute useful information on aid flow by province.²⁰ Disaggregation at district level is also possible.

Analyses of existing survey data

Measurement of intervention coverage in low-income countries requires population-based surveys. Demographic and health surveys (DHS)²¹ or UNICEF's multiple indicator cluster surveys (MICS) are undertaken every 3–5 years in most low-income countries, and an increasing number of additional surveys are being implemented to assess broader development indicators (eg, living standards measurement studies surveys)²² or specific vertical disease programmes (eg, malaria indicator surveys,²³ HIV/AIDS indicators surveys).²⁴ The frequency of such surveys is likely to increase as the MDG 2015 deadline approaches. In most countries, survey samples are insufficient to provide precise estimates of coverage at district level. Rare exceptions include Malawi's 2010 DHS with 1000 households per district,²⁵ and India's reproductive and child health surveys with a total sample of more than 600 000 households.²⁶

However, even if the number of sampled households per district is small, few programmes are implemented in just one district, and pooling across several districts could result in sufficient numbers of individuals to assess coverage. Generally, researchers believe that survey results should not be pooled for groups of districts, unless these districts were defined a priori, because few national surveys are designed to provide probability samples at district level. We question this logic, especially since survey reports systematically break down national

results according to age, socioeconomic, and ethnic categories, even though the sample was not designed to be strictly representative of such subgroups. In practice, most nationally representative surveys introduce implicit stratification within every district by listing enumeration areas in a geographic sequence and systematically sampling these areas; as a result, households included in the sample tend to be spread throughout the districts. With due attention to sampling weights, groups of districts in which a programme was implemented can be separated from a national survey for the assessment of coverage.^{27,28} Dose–response analyses of programme implementation strength and coverage are also possible; these are discussed below (see Analyses for programme enhancement and evaluation). For programmes covering a few districts for which even the pooled sample size is insufficient, coverage estimates will need national DHS or MICS with oversampling, as was done in the ACSD evaluation, or separate surveys will have to be done, using comparable methodologies in programme districts.

The primary indicator for MDG 4 (reduction of under-5 mortality by two-thirds) is the mortality rate in children younger than 5 years, and for MDG 5 (reduction of maternal mortality by three-quarters and universal access to reproductive health) it is the maternal mortality ratio. Most DHS and some MICS include full birth or pregnancy histories, in which women of reproductive age report on how many children they ever had, their dates of birth, whether they have died, and if so the dates of death. This information allows retrospective construction of annual mortality rates for young children for a period of up to 10 calendar years before the survey, calculated similarly to the 5-year rates published by DHS.²⁹

The sample size limitations discussed in the context of coverage estimation are even more vital for estimates of under-5 mortality; in this case, oversampling of selected districts with a programme when a national survey is undertaken, pooling results of such surveys across several districts, or doing stand-alone surveys might be needed. In the ACSD evaluation,⁸ national DHS were oversampled in the programme districts in Benin and Mali, and this action allowed comparison of trends in mortality in these areas to those in the rest of the country.

A commonly used impact indicator is prevalence of undernutrition (underweight, stunting, and wasting) in children younger than 5 years, which is related to MDG 1 (eradication of poverty and hunger). Sample size issues are not as important for these indicators because their frequency tends to be high and because their denominator includes all children younger than 5 years in the sample. Use of mean values of weight-for-age, height-for-age, or weight-for-height Z scores further reduces need for large samples, compared with estimation of the prevalence below a given cutoff.³⁰

Maternal mortality ratios can also be estimated on the basis of surveys. However, the required sample sizes are

Panel 3: Examples of questions answerable by the platform approach**Interim evaluation questions (early in the implementation cycle)**

Are programmes being deployed where need is greatest?

Implementers usually assert that their programmes are being deployed in areas of the country where mortality is highest or poverty most frequent, but this statement is not always true. Indicators of baseline mortality and poverty levels and of the strength of health systems available in the platform database will be linked through the platform to data for implementation, thus supporting assessments of whether programmes are indeed deployed where they are most needed. An example of how useful such simple analyses can be comes from the Integrated Management of Childhood Illness (IMCI) evaluation in Brazil, where implementation was stronger in municipalities that were close to their state capitals than in those that were poorest and had high mortality—an example of placement bias of health programmes.¹⁸ In the west African Accelerated Child Survival and Development evaluation, some countries implemented the programme in the neediest areas, but others did not.⁸

Is implementation strong enough to have an impact?

Insufficient implementation is a common reason for absence of impact. Analyses of implementation data gathered at both national and local level will allow linking of inputs and outputs to target populations in the evaluation platform, thus obtaining an estimate of the strength of implementation. Although a high ratio of outputs to population is not necessarily indicative of high coverage, because there could be leakages or other inefficiencies, low ratios surely suggest that implementation is unlikely to lead to a measureable impact. Simulation exercises can contribute to answering this question. The lives saved tool (LiST), which allows users to estimate the impact of changes in coverage for proven interventions on maternal, neonatal, and under-5 mortality,³¹ is a useful adjunct to these analyses.

What approaches lead to rapid coverage increases in the short term?

Different implementing agencies and partners generally rely on diverse delivery channels for increasing coverage—eg, facility-based approaches, outreach sessions, community health workers, involving the private sector, community groups, etc.³² Interim analyses of trends in coverage based on mid-term surveys, linked through the platform to implementation data, could help assess which of these approaches are most effective in the short term. Data gathered through the platform might also identify districts that are doing well and those that are lagging behind, and motivate further in-depth analyses with a so-called mixed-methods approach to improve implementation.³³

Summative evaluation questions (at the end of a programme cycle)

Are coverage rises sustained?

Analyses of short-term coverage increases, associated with specific delivery channels (as discussed above), must be complemented by longer term assessments of whether these rises are sustained. This research is especially important in view of findings on the damaging effects of stock-outs of essential commodities (such as drugs or insecticide-treated nets), shifting priorities and funding patterns, and other barriers to continuity of a programme.^{8,34,35} The evaluation platform design will support analyses of coverage trends in areas with every type of programme, recording associations between the intensity of implementation activities and coverage levels. Indices of programme effort can be used to summarise implementation strength.⁶

(Continues on next page)

very large and estimates refer to a period that is too far backdated to be useful for evaluation. As an alternative, coverage of maternal interventions can be measured and used as a proxy for mortality.

Household surveys are an important source of data relating to progress towards the goals of MDG 6 (to halt

and reduce spread of HIV/AIDS, malaria, and other diseases and provide universal access to treatment for HIV/AIDS). They might include data for coverage of malaria interventions (such as use of insecticide-treated bednets, indoor residual spraying, intermittent preventive therapy during pregnancy, and treatment of children with fever) and selected HIV interventions (such as HIV testing coverage in adults and pregnant women attending antenatal clinics). Surveys with biological data collection can provide information about prevalence rates of HIV infection and parasitaemia. Other indicators, such as tuberculosis treatment success rates and use of antiretroviral therapy, are based on reports from health facilities.

Disaggregation of survey data at district level might not be possible if information about geographic location of every cluster is deleted from the database—a process known as scrambling. This procedure is common when sensitive data are obtained (such as HIV serology). Deletion of information will severely limit the potential usefulness of surveys for evaluation of all outcomes, including those unrelated to AIDS, and in our view this limitation is strong justification for not including serological status in multipurpose surveys.

Analyses for programme enhancement and evaluation

The evaluation platform will ideally summarise all types of available data for factors affecting outcomes of interest (figure 1), including baseline and regular updates on information related to geographic, epidemiological, socioeconomic, demographic, and other relevant characteristics of every district. It will also include variables related to programme intensity—in particular, implementation scores—and coverage and impact data.

The new design allows for various analytical techniques. First, data can be pooled from several districts to obtain robust coverage and health outcome estimates. Second, dose–response analyses can be undertaken, in which every district is a datapoint. Third, poor baseline comparability attributable to so-called placement bias (or the fact that districts receiving the programme might differ from other districts in the country) can be dealt with. Fourth, consistency of indicators can be assessed across different levels of the framework (figure 1)—eg, relation of programme implementation scores, coverage, and impact indicators to each other. Fifth, modelling exercises can be implemented, such as those supported by the lives saved tool (LiST).³¹ Finally, changes in contextual factors can be incorporated into analyses.

Districts are the primary units of statistical analysis for the national evaluation platform approach. We envisage two main types of data analysis: (1) interim (or formative) analyses could be undertaken as soon as the database is set up with process indicators and results fed back to implementers to allow mid-course corrections; and (2) endline (or summative) analyses could be done at the

end of a programme cycle to assess how it affected coverage and impact indicators.

Interim (or formative) analyses based on the evaluation platform approach can produce results that not only are useful to governmental health managers (and all agencies that have a role in implementation) for making mid-course corrections but also increase the probability that health system barriers to full implementation are identified and addressed. Feedback based on interim analyses is an important departure from traditional evaluation approaches, in which external evaluators are advised to refrain from interfering with programmes under assessment. Double-blinding, which is strongly recommended for testing new interventions, is not feasible in the context of evaluation of scale-up of proven interventions.

At the end of a programme cycle, the evaluation platform can enable summative questions to be answered. The timeline for assessment of a programme usually takes 5–7 years to complete, including time needed for full implementation of the programme, for the biological effect of the intervention to take place, and for impact to be measured. Panel 3 shows illustrative questions and approaches to interim (or formative) and summative analyses.

The national evaluation platform design will also help answer broader policy-relevant questions that are sometimes difficult to address with traditional designs. Examples include assessment of the equity effects of a programme (which can be undertaken by analysis of coverage and impact, broken down by sex, socioeconomic status, urban or rural residence, and ethnic group) and investigation of unintended effects of programmes (by looking at other health outcomes in the programme area). In view of its multi-programme, multiple-outcome nature, the platform is ideally suited to answering such questions or those related to programme sustainability. If the number of districts permits, researchers could also assess whether the effect of a programme differs according to district characteristics, such as baseline levels of mortality or socioeconomic development. A national evaluation platform, maintained over the long term, can provide answers to questions that cannot be obtained from short-term, single-programme evaluations.

Sample size

Calculations of sample size and statistical power for the evaluation platform will depend on the types of comparisons planned. We begin by considering use of the platform to undertake a traditional comparison of districts with and without a given programme. Because the number of districts is fixed and analyses will rely on existing surveys (in which the number of households per district is also fixed), sample size calculations will allow estimation of study power, or the likelihood that a true effect is picked up in analyses. Evidence of low power will

(Continued from previous page)

Did programmes have an impact?

By linking implementation data to information on mortality and nutritional status, the platform design will allow assessment of a programme's impact. For example, in IMCI evaluations in Peru and Brazil, training coverage of health professionals was not associated with child mortality indicators in dose–response analyses.^{17,36} Variable strengths of programme implementation in different districts will favour ecological, dose–response, time-trend analyses in the platform design. However, the more traditional approach of comparing areas with and without each programme is also possible.

Was coverage associated with impact?

Although the answer to this question might seem obvious, reasons why increased coverage might not have an effect do exist, such as wrong choice of interventions in view of the epidemiological profile of the population, lack of essential cofactors, or low quality of intervention delivery. Dose–response time-series analyses correlating coverage levels with measures of impact across all districts in the country will help answer this question. Simulations with LiST³¹ will be especially useful, by allowing comparison of estimated impact on the basis of coverage changes with actual reductions in mortality or undernutrition.

Do alternative explanations exist for the findings?

The wealth of data for contextual factors and their change over time available in the platform can help rule out alternative explanations for impact findings, enhancing the plausibility that the noted effect is attributable to a given programme.⁷ This process can be done by incorporation of relevant contextual factors as confounding variables in regression models, as was done in the Brazil and Peru IMCI evaluations.^{17,36}

Which programmes were most cost effective?

This question will need information on costs gathered as part of programme documentation related to health impact results.

require enrolment of increased numbers of households within each district, or oversampling. Even if oversampling is not feasible and study power is less than what would be desirable, the platform approach can still contribute to interpretation of results by providing additional information on placement bias, confounding variables, and presence of similar programmes in the comparison areas.

If dose–response analyses are used, which will usually be the case, then sample sizes will be affected by how much the implementation score varies across the different districts, how much variability there is in baseline coverage, and how strongly implementation affects coverage. Calculations for an evaluation platform design in Malawi suggested that there would be 80% power for detection of a change in coverage of antibiotic treatment for pneumonia in the presence of a community case-management programme, in view of the following assumptions: 28 districts in the country, samples of 1000 households per district, and an increase in coverage of seven percentage points for every ten percentage point increase in the implementation score. Malawi is a special case, with relatively few districts compared with other countries and large survey sample sizes within each district; calculations must be done on a country-by-country basis.

Conclusions

We propose a systematic approach for evaluation of scale-up of national programmes for maternal and child survival and potentially other public-health programmes that address specific diseases (such as HIV/AIDS, tuberculosis, or malaria). A national evaluation platform shows how ideas have evolved in response to changes in the development landscape in most countries of low and middle income, where governments are working with several partners to implement many simultaneous public-health programmes with overlapping aims and geographic boundaries.

Why does this approach make sense? First, setting up a broad platform design that includes documentation of contextual factors and implementation of many programmes—and indicators of coverage, impact, and cost—puts governments in the driving seat and provides them with information needed to make wise decisions about how they should use scarce resources. The platform approach supports country ownership of national programmes and their evaluation, helps build local capacity, and promotes donor coordination in the spirit of the Paris Declaration.³⁷

Second, the approach is likely to be cost effective. Rather than replacing existing databases and data repositories, the platform approach will allow them to communicate with one another, with the specific objective of answering evaluation questions. This strategy can serve as an organising framework for large-scale survey datasets supported by various partners and with several aims, permitting full use of every survey to address broad public-health questions rather than focus on one disease, programme, or population, and complementing other initiatives and proposals for strengthening of district-level data use in support of programmes.³⁸ Although the initial investment in setting up the platform could be fairly small, large surveys will be necessary in some countries to measure coverage and impact indicators with adequate precision for groups of districts implementing a given programme. On the other hand, large surveys that allow assessment of many programmes are likely to be cheaper than stand-alone traditional evaluations of these programmes, each with an intervention and comparison group of districts, which is the *modus operandi* at present. Even if costs of large surveys needed for the platform approach exceed current evaluation investments, substantial gains will arise in terms of geographic coverage and the scope of findings.

Third, the platform design promotes evaluation as a continuous process aimed at improvement of implementation, adapted to the current realities of simultaneous undertaking of several programmes and the resulting absence of untouched comparison areas. The flexible and comprehensive design of the platform allows evaluation to respond to changes in implementation, avoiding potential blows to rigid designs, such as strong implementation of a new

programme in a previously defined comparison area. Feedback to all levels and linkage to programming cycles are essential features of our approach.

Finally, the platform enhances the independence of evaluators, who are not being commissioned and financed by the one agency whose programme is under scrutiny. It will help governments use the many surveys done in their countries as a source of broader learning about the relative effectiveness and costs of alternative approaches implemented with support from various partners. To ensure independence, the platform should be hosted by a national academic or research institution that is not directly involved in the programmes being evaluated, with a steering committee of national (Ministry of Health, national statistical office, etc) and international (UN agencies, bilateral agencies, NGOs, etc) partners. The research institution would employ and supervise the network of key informants at district level.

One limitation of the national evaluation platform approach is its observational design. No other alternative is feasible, however, in view of the focus on evaluation of real-world programmes as they go to scale and the resulting inability of assessors to control the pace and location of implementation.

A second limitation is the platform's reliance on transparency and collaboration between government and development partners, and within many agencies. Funding of an evaluation platform has risks for donors, because the results will reflect the contributions of many and will be in the public domain.

Lastly, development of a national information system for programme evaluation that integrates survey results with facility-based and administrative statistics to provide usable district-level information on an ongoing basis is not a trivial matter. It will need an enormous change within countries concerning the way facility and administrative statistics are used. Development of the platform will also require major investments in local capacity building, leading to a shift from stand-alone evaluations commissioned by external parties to country-led assessments that will allow each country to know more about its own programmes. Guidance on how platforms should be made operational is specific to the country and context, and provision of detailed recommendations is beyond the scope of this article. Indeed, implementation of evaluation platforms is an intervention that needs sufficient funding and deserves its own external evaluation.

Balancing these limitations is the realism and rigor of the platform design. The current reductionist approach of single-programme, intervention-versus-control-area evaluations has a limited role in the programmatic and development contexts of countries where most avoidable deaths happen.

Contributors

CGV prepared the first draft of the article, based on discussions with JB and REB arising from their joint work on evaluation of maternal and child health programmes over several years. JTB expanded the article to

cover other health outcomes, and to place it in the context of national information systems. All authors provided comments on subsequent drafts and approved the final version of the manuscript.

Conflicts of interest

We declare that we have no conflicts of interest.

Acknowledgments

We thank Hilde De Graeve and Bert Schreuder for allowing reproduction of figure 2; Baltazar Chilundo, Fatima Abacassamo, and Iná Santos for insights into the adaptation of the method to the Mozambique situation; Jennifer Requejo, Kate Gilroy, and David Peters for comments on the text; Agbessi Amouzou and Larry Moulton for inputs on sample size calculations; and Gareth Jones for comments on the plan of statistical analyses. Members of the Catalytic Initiative Evaluation project group at the Institute for International Programs of Johns Hopkins Bloomberg School of Public Health have contributed in important ways to the experience and thinking described in this report. This work was supported by the Canadian International Development Agency and the Bill & Melinda Gates Foundation through subcontracts to the Institute for International Programs at the Johns Hopkins Bloomberg School of Public Health for independent evaluations of the Catalytic Initiative to Save a Million Lives and of the Rapid Scale Up of the Partnership for Maternal, Newborn and Child Health.

References

- The Lancet. Evaluation: the top priority for global health. *Lancet* 2010; **375**: 526.
- Evaluation Gap Working Group. When will we ever learn? Improving lives through impact evaluation. Washington, DC: Center for Global Development, 2006.
- Oxman AD, Bjørndal A, Becerra-Posada F, et al. A framework for mandatory impact evaluation to ensure well informed public policy decisions. *Lancet* 2010; **375**: 427–31.
- World Bank. Results-based financing (RBF). Nov 11, 2008. <http://go.worldbank.org/UDQRQYSTF0> (accessed Feb 8, 2010).
- Victora CG, Black RE, Bryce J. Evaluating child survival programmes. *Bull World Health Organ* 2009; **87**: 83.
- Catalytic Initiative to Save One Million Lives. Evaluating the scale-up for maternal and child survival: putting science to work for mothers and children. June 16, 2008. http://www.jhsph.edu/dept/ih/IIP/projects/catalytic_initiative/Common_evaluation_framework.pdf (accessed June 29, 2010).
- Habicht JP, Victora CG, Vaughan JP. Evaluation designs for adequacy, plausibility and probability of public health programme performance and impact. *Int J Epidemiol* 1999; **28**: 10–18.
- Bryce J, Gilroy K, Jones G, Hazel E, Black RE, Victora CG. The Accelerated Child Survival and Development programme in west Africa: a retrospective evaluation. *Lancet* 2010; **375**: 572–82.
- Bryce J, Victora CG, and the MCE-IMCI Technical Advisors. Ten methodological lessons from the Multi-Country Evaluation of Integrated Management of Childhood Illness. *Health Policy Plan* 2005; **20** (suppl 1): 194–105.
- Arifeen SE, Hoque DME, Akter T, et al. Effect of the Integrated Management of Childhood Illness strategy on childhood mortality and nutrition in a rural area in Bangladesh: a cluster randomised trial. *Lancet* 2009; **374**: 393–403.
- Canadian International Development Agency. The catalytic initiative to save a million lives. Nov 26, 2007. <http://www.acdi-cida.gc.ca/acdi-cida/acdi-cida.nsf/eng/NAD-1249841-JLG> (accessed Feb 8, 2010).
- Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006; **6**: 54.
- Schellenberg JRMA, Adam T, Mshinda H, et al. Effectiveness and cost of facility-based Integrated Management of Childhood Illness (IMCI) in Tanzania. *Lancet* 2004; **364**: 1583–94.
- McKeown T. The modern rise of population. New York: Academic Press, 1976.
- Gregson S, Adamson S, Papaya S, et al. Impact and process evaluation of integrated community and clinic-based HIV-1 control: a cluster-randomised trial in eastern Zimbabwe. *PLoS Med* 2007; **4**: e102.
- Grassly NC, Garnett GP, Schwartländer B, Gregson S, Anderson RM. The effectiveness of HIV prevention and the epidemiological context. *Bull World Health Organ* 2001; **79**: 1121–32.
- Amaral J, Leite AJ, Cunha AJ, Victora CG. Impact of IMCI health worker training on routinely collected child health indicators in northeast Brazil. *Health Policy Plan* 2005; **20** (suppl 1): i42–48.
- Victora CG, Huicho L, Amaral JJ, et al. Are health interventions implemented where they are most needed? District uptake of the integrated management of childhood illness strategy in Brazil, Peru and the United Republic of Tanzania. *Bull World Health Organ* 2006; **84**: 792–801.
- Huicho L, Dávila M, Campos M, Drasbek C, Bryce J, Victora CG. Scaling up integrated management of childhood illness to the national level: achievements and challenges in Peru. *Health Policy Plan* 2005; **20**: 14–24.
- ODAmoz. Mozambique donor atlas 2008. December, 2008. <http://mozambique.odadata.ampdev.net/> (accessed June 8, 2010).
- Measure DHS. Demographic and health surveys: DHS overview. <http://www.measuredhs.com/aboutsurveys/dhs/start.cfm> (accessed Feb 8, 2010).
- World Bank. Living Standards Measurement Study. <http://go.worldbank.org/IPLXWMCNJ0> (accessed June 8, 2010).
- Measure DHS. Malaria Indicator Survey. <http://www.measuredhs.com/aboutsurveys/mis/start.cfm> (accessed Feb 8, 2010).
- Measure DHS. AIS surveys: AIS overview. <http://www.measuredhs.com/aboutsurveys/ais/start.cfm> (accessed Feb 8, 2010).
- Measure DHS. Malawi: standard DHS, 2010. http://www.measuredhs.com/aboutsurveys/search/metadata.cfm?surv_id=333&ctry_id=24&SrvTp (accessed Feb 8, 2010).
- International Institute for Population Sciences (IIPS) and Macro International. National Family Health Survey (NFHS-3), 2005–06: India—volume I. September, 2007. http://www.nfhsindia.org/NFHS-3%20Data/VOL-1/India_volume_I_corrected_17oct08.pdf (accessed June 28, 2010).
- West BT, Berglund P, Heeringa SG. A closer examination of subpopulation analysis of complex-sample survey data. *Stata J* 2008; **8**: 520–31.
- Rao JKN. Small area estimation. Hoboken: Wiley Interscience, 2003.
- Rutstein SO, Rojas G. Online guide to DHS statistics. October, 2006. <http://www.measuredhs.com/help/Datasets/index.htm> (accessed June 8, 2010).
- WHO. Physical status: the use and interpretation of anthropometry: report of a WHO Expert Committee (technical report series no 854). 1995. http://www.who.int/childgrowth/publications/physical_status/en/index.html (accessed June 8, 2010).
- Department of International Health, Johns Hopkins Bloomberg School of Public Health. LiST: the Lives Saved Tool—an evidence-based tool for estimating intervention impact. <http://www.jhsph.edu/dept/ih/IIP/list/index.html> (accessed Feb 8, 2010).
- Bryce J, el Arifeen S, Pariyo G, Lanata CF, Gwatkin D, Habicht J-P. Reducing child mortality: can public health deliver? *Lancet* 2003; **362**: 159–64.
- Peters DH, El-Saharty S, Sladat B, Janovsky K, Vujicic M. Improving health service delivery in developing countries: from evidence to action. Washington, DC: World Bank, 2009.
- Bryce J, Daelmans B, Dwivedi A, et al, on behalf of the Countdown to 2015 Core Group. Countdown to 2015 for maternal, newborn, and child survival: the 2008 report on tracking coverage of interventions. *Lancet* 2008; **371**: 1247–58.
- Biesma RG, Brugha R, Harmer A, Walsh A, Spicer N, Walt G. The effects of global health initiatives on country health systems: a review of the evidence from HIV/AIDS control. *Health Policy Plan* 2009; **24**: 239–52.
- Huicho L, Dávila M, Gonzales F, Drasbek C, Bryce J, Victora CG. Implementation of the Integrated Management of Childhood Illness strategy in Peru and its association with health indicators: an ecological analysis. *Health Policy Plan* 2005; **20** (suppl 1): i32–41.
- OECD. Paris declaration on aid effectiveness: ownership, harmonisation, alignment, results and mutual accountability. Paris: Organisation for Economic Co-Operation and Development, 2005.
- Rowe AK. Potential of integrated continuous surveys and quality management to support monitoring, evaluation, and the scale-up of health interventions in developing countries. *Am J Trop Med Hyg* 2009; **80**: 971–79.