Methods for Evaluating Innovative Health Programs (EIHP): A Multi-Country Study

Thomas, Ranjeeta and Jones, Andrew M and Squire, Lyn

University of York, United Kingdom

July 2010

Methods for Evaluating Innovative Health Programs (EIHP): A Multi-Country Study

Ranjeeta Thomas[1], MSc, Andrew M Jones[1], PhD, Lyn Squire[2], PhD,

Affiliations: [1]University of York, UK, [2]Brookings Institution, USA

Corresponding author: Prof Andrew M Jones, Department of Economics and Related Studies, University of York, York, YO10 5DD, United Kingdom. Tel: +44-1904-433766. Fax: +44-1904-433759. Email: amj1@york.ac.uk.

Keywords: Millennium Development Goals; child and maternal health; communicable diseases; impact evaluation; capacity building; Asia; Africa; Latin America

Running title: Methods for Evaluating Innovative Programs

Acknowledgements: The authors would like to thank the Bill & Melinda Gates Foundation for their generous funding of this project.

Authorship: LS and RT were involved in the planning and management of the EIHP project throughout its duration. AJ was involved in the reporting and dissemination phases of the project. All authors have participated in the interpretation of the studies and the reporting stage of this manuscript and seen and approved the final version.

Conflict of interest: We declare that we have no conflict of interest.

Methods for Evaluating Innovative Health Programs (EIHP): A Multi-Country Study

ABSTRACT

Designed as a global research initiative, the EIHP project aims at adding to the evidence base of health interventions that have the potential to improve health outcomes in Africa and Asia. The project focuses on rigorous, quantitative evaluations of innovative local initiatives that address the Millennium Development Goals for health: reductions in child and maternal mortality and communicable diseases. This overview brings together the outcomes and lessons from the project for evaluation methods. It draws together the methodological implications of carrying out impact evaluations under very different settings and emphasizes the need to build in evaluations in project designs.

INTRODUCTION

Over the last few decades billions of dollars of international aid have been channelled to Africa and Asia. However, despite these resources these regions continue to remain the poorest in the world (Chen and Ravallion 2007) with the worst health outcomes. The continuing poverty and poor health reflects the limited impact aid has had in improving the conditions in these regions. Despite a large amount of research on aid effectiveness the causal link between aid and improved outcomes is at best still murky (Bourguignon and Sundberg 2007). Research on the last link in the casual chain, on the evaluation of development programs and policies, is a critical tool in channelling resources to where they are likely to have the greatest impact (White 2006). The combined need of improved health and aid effectiveness point to the urgency of identifying and evaluating the impact of innovative programs and policy measures to address each health goal in ways that are appropriate in the specific context of Africa and Asia.

The Global Development Network's (GDN) project "Evaluating Innovative Health Programs" (EIHP), funded by the Bill & Melinda Gates Foundation, seeks to inform policy on the effectiveness of health solutions that have the potential to improve health outcomes in developing countries. It evaluates the impact of nineteen programs from across developing and transition countries that focus on the health-related Millennium Development Goals (MDGs) of reducing child and maternal mortality, and halting and reversing the trend of communicable diseases such as HIV/AIDS, malaria and other diseases (United Nations 2008).

The EIHP project includes in its portfolio a range of programs including national programs and small scale non-governmental programs and both supply and demand-side interventions. Table 1 summarizes the classification of the programs in terms of their location, the health issue they address and the agency responsible for their implementation.

INSERT TABLE 1 AROUND HERE

METHODOLOGICAL ISSUES

Impact evaluation involves answering policy-relevant questions using counterfactual analysis (see Heckman and Vytlacil 2007, Blundell and Costa-Dias 2008, Imbens and Wooldridge 2008, for overviews of program evaluation methods). Specifically, the evaluations aim to assess whether changes in outcomes for an individual, household or institution are attributable to a specific program or intervention. They are based on the notion of causal inference, which aims to isolate the impact of the program from other confounding factors that influence the outcomes (see for example: Rubin 1974, Holland 1986, Heckman 2008). The methods available to assess impact vary according to the nature of the program under evaluation, for example whether it is new or ongoing. Also, they differ in terms of the policy questions they can answer, their data requirements and the costs involved in carrying out the evaluations.

The programs evaluated in this project differ greatly in size and scope. Most of the national programs have been operational for several years and cover a large proportion of the target population, while others are either new programs implemented as small scale pilots or on-going local initiatives. The identification strategy adopted to evaluate a program depends on the nature and scope of the intervention and the availability of data. For the new programs, where the investigators had some discretion over the assignment of the program, randomization was used to create a source of exogenous variation, with the collection of baseline and follow-up data built into the evaluation process. More often, in the case of national programs which are implemented to achieve a specific policy target or for a specific group, treatment is not randomly assigned but variation in the geographical coverage and intensity of treatment is exploited to isolate program effects. The growth in availability of observational data has resulted in significant developments in techniques for evaluations using non-experimental approaches. Where suitable longitudinal data are available difference-in-differences (DiD) is used, sometimes combined with adjustments for observed confounders by regression or matching methods. In cases where longitudinal data are not available, matching methods or the regression discontinuity approach is applied using cross-sectional data.

The experimental and non-experimental approaches have their benefits and limitations in terms of their applicability to a particular situation and in the costs and resources required to carry them out. Banerjee and Duflo (2008) review the advantages of randomized experiments. Experimental designs allow the assignment of treatment to be isolated and controlled by researchers, improving the scope to provide internally valid estimates of program effectiveness. They also seek to address some of the common criticisms of randomized experiments: that the results of an experiment are dependent on the specific environment where it is carried out, limiting their generalizability to other contexts; that there are often problems with compliance in experimental studies; that randomization itself may affect the outcomes (for example, through Hawthorne effects); that small-scale experiments may only reflect partial equilibrium effects and fail to capture general equilibrium or spillover effects that may occur when policies are implemented on a larger scale; and that experiments may not be able to capture heterogeneity in treatment effects. The drawbacks of a mechanical reliance on the experimental approach, and of 'quasi-experimental' approaches that use instrumental variables, with instruments selected to mimic a randomized experiment, rather than being drawn from a structural economic model, are reviewed by Deaton (2008). He is critical of the use of these methods to evaluate projects *per se* and favours their use as tools to aid our understanding of the underlying theoretical mechanisms that drive behaviour. The EIHP project lies within the tradition of applying impact evaluation methods to specific interventions.

A critical element of these evaluations is the cost involved in implementing them. In resource constrained settings such as those where these programs are operational, costs can be a significant barrier in undertaking an evaluation. Each of the evaluations submitted detailed budgets covering costs for personnel, data collection and administration. Table 2 shows the reported costs of the evaluations by the method applied.

INSERT TABLE 2 AROUND HERE

EVALUATING NEW PROGRAMS

Evaluating new ideas with the potential for improving health outcomes involves implementing pilots of untested ideas and subjecting them to evaluations using the most rigorous techniques available. Randomized experiments have widely been considered the 'gold standard' in evaluating programs and when conducted well offer the best estimates of impact.

Randomized Experiments

Four of the EIHP programs use *randomized experimental designs* to examine a variety of outcomes and the economic implications of the programs. In this approach, a sample of individuals is selected from the population of interest, where the selection may be done on a set of observable variables that makes the sample homogeneous. This sample is then divided randomly into two or more groups. In the simplest case one group receives the treatment and the other serves as the control. The random assignment should ensure balanced control and treatment groups and should remove any selection bias in participation. The key assumption here is that all other variables – either observed or unobserved are independent of treatment allocation (Ravillion 2008). The impact is estimated as the difference in means between the two groups before and after the program is implemented (i.e. the difference between the differences). Primary data collection for both treatment and control groups before and after the program is implemented is a key component of experiments. Randomization's adaptability to different definitions of treatment and multiple treatment arms can be seen in the four studies.

Randomized experiments can overcome potential bias from adverse selection in programs involving economic incentives such as health insurance. In the case of health insurance, individuals with greater risks of illness are more likely to enrol in insurance programs. By randomizing the provision of insurance this unobserved (by the researcher) risk is on average equal in the treatment and control groups. The experimental program in Nicaragua randomizes 'incentives' to obtain health insurance for informal sector workers in the capital city of Managua (Thornton et al. 2010). It uses multiple treatment arms and randomizes information on the insurance, costs of insurance and convenience of signing-up for insurance. The treatment arms include, a brochure detailing the insurance program, a brochure with a 6-month insurance subsidy with instructions to sign up at Nicaraguan Social Security Institute's (INSS) central office, and a 6-month insurance subsidy with instructions to sign up at any one of three participating microfinance institutions (MFI). The impact of the treatment is relatively straightforward to recover by comparing the difference in the means between the treatment and control groups for the outcome of interest (in this case the take-up of insurance). Comparisons between the MFI group with the INSS and the baseline is used to evaluate the effectiveness of MFIs as delivery agents for health insurance. In order to measure the effect of having insurance on utilization and expenditures they use an instrumental variables framework with the treatment status as an instrument for having insurance.

In contrast to economic incentives, the experimental program in Thailand analyzes the take-up of HIV testing when the service is initiated by providers (Teerawattananon et al. 2009). Closer in design to a pure experimental setting, a cluster randomized trial design allocated hospitals with low and high HIV prevalence into treatment and controls. The intervention in this program involves presenting each patient between the ages of 13 and 64 who visited an outpatient department with an invitation card for free counselling and HIV/AIDS test. The validity of the results of experiments depends critically on the success of the randomization in balancing the control and treatment arms across observable (and unobservable) characteristics. Most often in social experiments differences do remain between the groups and regression methods are applied to correct for these differences. In the case of the Thailand experiment, the Generalized Estimating Equation (GEE) approach is used to estimate the impact of the intervention on average acceptance rates, while accounting for variations in cluster variables.

Of the other two experimental evaluations the Malawi conditional cash transfer program focuses on economic incentives (Baird et al. 2010), while the Colombia program randomizes childcare centres to test the impact of frequent hand washing, using hand gels, on preventing diseases in children (Correa et al. 2010).

Successful randomized evaluations provide the best basis for internally valid estimates of treatment effects by simultaneously controlling for differences in observables and unobservables. On this basis the four programs by this classification score the highest in terms of internal validity. The four experimental programs are all pilots, with sample sizes ranging from 3,821 (CCT in Malawi), 4,001 respondents at the baseline and 2608 at the follow-up for the Nicaragua program, 16 hospitals in the Thailand program and 46 childcare centres (1,671 children) in the Colombian program. Each of these programs was implemented in a single district or metropolitan area and collected baseline and follow-up data for the treatment and control groups. One common factor that distinguishes these experimental evaluations from the others in the project is the much larger financial outlay required to estimate the impact after one year of the program. This however is driven largely by the data collection required in these experiments which as shown in Table 3 are a significant component of total costs for these evaluations. This financial outlay must be placed in the light of the benefits from estimating accurate impact of pilot programs through well designed experiments prior to their expansion or implementation on a larger scale.

Randomized experiments however are limited in their time scale as well as in their ability to evaluate on-going programs. For example, successful pilots that are gradually expanded to cover the entire target population allow randomization until the control group is exhausted or until the randomization process is contaminated. Long term, large scale evaluations require randomization to be complemented by the non-experimental methods discussed in the following sections.

EVALUATING ON-GOING PROGRAMS

Fifteen programs in this project fall under the category of on-going interventions. These interventions have been evaluated using a range of non-experimental methods. The methods applied depend on the way that the programs are implemented as well the data available. These programs are classified below based on the data availability.

Pre- and Post-Intervention Data

The basic difference-in-differences approach used to estimate the treatment effect in the randomized experiments is applicable to evaluations based on observational data. This approach is the most widely used in identifying treatment effects in non-experimental settings, particularly for large programs implemented in multiple regions. In this project, six of the programs are evaluated using the method of DiD (Ashenfelter 1978, Ashenfelter and Card 1985, Heckman and Robb 1985). These programs are national interventions that cover large portions of the population and have been active for several years. As in the case of randomized experiments, data are required before and after the intervention for both treatment and control groups. The data may either be longitudinal for a particular set of individuals (as in randomized experiments) or repeated cross-sections from the target population.

Depending on the data available slightly different sets of assumptions are required for the DiD estimator. Particularly in the case of repeated cross-sections, two critical assumptions are (1) that the composition of the observations in the cross-sections has not changed over time (2) the treatment group can be clearly identified in the first time period. The second assumption would not hold if a program was implemented nationwide in a single time period. However, this assumption usually holds in most development policy scenarios where programs are typically targeted at a particular group of individuals or implemented in specific regions. The DiD approach falls under the category of methods that allows for selection on unobservable confounders, but any unobservable factors that are likely to influence participation are assumed to be time invariant. Without the exogenous variation provided by randomization to treatment and control, this approach relies on variation in implementation and timing to identify treatment effects. However, endogeneity can often be a problem in this approach if program allocation has been purposive based on pre-existing characteristics or other dynamic characteristics, for example if assignment of the program is influenced by a dip in pre-treatment outcomes.

The Family Health Program (PSF) in Brazil illustrates the different dimensions of using a DiD approach (Rocha and Soares 2010). It uses municipality level data from five different sources to capture information on different aspects of mortality (outcomes), timing and implementation of the program (treatment-variation) and municipality level controls (covariates). It uses variation in which municipalities adopted the program and the length of time they have been exposed to it for identification of treatment effects. The endogeneity problem discussed earlier is critical in this approach. In the case of this program, if the program was first implemented in municipalities with high mortality rates then allocation to treatment can no longer be used as an exogenous source of variation. However, since the program in Brazil was eventually expanded to cover all the municipalities it would seem this source of endogeneity is not an issue. Similarly, if the timing of adoption in a municipality followed the occurrence of a negative health shock that resulted in a spike in the mortality rates the treatment effects identified will be biased. To address this problem the evaluation uses municipality level and state level fixed-effects respectively to control for differences leading to adoption and timing of the program. The study also carries out a simple test for the existence of endogeneity to verify whether the above concerns of endogeneity are indeed an issue. The test involves a hazard rate analysis of the determinants of the probability of a municipality joining the program. The probability is estimated as a function of municipality fixed characteristics, changes in health variables, political variables and socioeconomic variables. The findings show a very small correlation between participation and previous health shocks while political considerations were found to have larger impacts indicating that the above concerns of endogeneity may not be too serious an empirical issue.

The DiD estimator can be biased by factors other than endogeneity of the policy variable. Particularly in the case of national programs, where this approach is most often used, identifying causal impact of the program under consideration can prove difficult. Most often governments are implementing multiple programs in the same region addressing different policy issues, but the impact of these programs could have indirect effects on outcomes not directly targeted by it. This potential problem is addressed in the Mother and Infant Health Program (MIHP) (Nizalova and Vyshnya 2010) in Ukraine which evaluates the impact on the program on infant mortality and maternal health related outcomes such as anaemia, blood circulation and late toxicosis. In this case existence of other health programs could result in an upward-bias of the estimated results of the MIHP evaluation. In order to test this contamination of the identified treatment effect, the evaluation of the MIHP program uses a placebo test by estimating the DiD model on pregnancy unrelated outcomes (in this case prevalence of diabetes and hepatitis). The lack of statistically significant impacts on health improvement for these unrelated outcomes is used as a measure of the validity of the DiD estimator on pregnancy related outcomes.

A further source of contamination in non-experimental settings is the contamination of the control group due to spillover effects. In health related development programs improvement of supply-side factors is a key intervention. These could include building health houses or improvements in procedures. In such non-experimental settings there is little to prevent individuals from non-intervened locations from accessing services in a ‘treated’ location. The only barrier in such cases would be the distance to the ‘treated’ facility. In such cases the recovered treatment effect is a biased estimate. One example of this problem is addressed in the evaluation of the Family Planning Program in Iran (Salehi-Isfahani et al. 2010). Implemented at the village level, this program involves the construction of rural health houses to provide family planning services to residents of the village. Contamination of the control group would occur if a treated village is located in close proximity to a non-treated village whose residents could easily access these services. To test for spillover effects the DiD regression includes a control variable for the interaction between coverage of the program at the (higher) district level and the characteristics influencing the likelihood of having a health house. This variable shows a positive impact on fertility (outcome) indicating that spillovers were not a major cause for concern in this case.

Similar to these three programs, the PARSalud in Peru (Díaz and Jaramillo 2009) and the CFW program in Kenya (Oduor et al. 2009) also use DiD to control for observable differences. For large programs it is particularly useful in identifying impact on aggregate statistics such as maternal mortality ratios, fertility rates or death rates from malaria. But it is easily extended to household level analysis as well. In the PARSalud evaluation, both levels of analysis are carried out. The household level analysis requires microdata which are less likely to be readily available and, as in this case, may have to be collected directly.

The accuracy of the DiD approach in generating unbiased measures of impact depends on the exogeneity of the variation in the areas adopting the program or in the timing of adoption and on controlling for observable factors that may bias the outcomes. Based on these criteria, an evaluation such as the Family Health Program in Brazil, which successfully tests and controls for these issues, could be given the highest credibility among the non-experimental evaluations. In contrast the program in Kenya has limited data on confounding factors, making the evaluation more prone to potential biases.

The final program in this section is Performance Based Financing for general health services in Rwanda (Basinga et al. 2009). This program was originally designed as a large scale district level randomized experiment to remunerate providers according to their performance on a given set of quantity and quality indicators. The intervention implemented at the health facility level covered 19 administrative districts across Rwanda. Due to the experimental design data collection at the facility level and exit interviews of service users was done at baseline and follow-up. However, prior to implementing the intervention a change in the demarcation of district boundaries led to some districts originally allocated to treatment having some health facilities from the control group. For equity purposes all facilities in such districts were classified as treatment. This break in the randomization results in a change in the analysis from a pure experiment to a DiD combining data from Rwanda's Health Information Management System with the primary data collected controlling for facility level and time fixed effects.

The DiD approach is only applicable in situations where data are available before and after the program for both groups. Such data however are not available in all developing countries. The growth in the use of the DiD approach has been driven by the growth of national data collection efforts such as the Living Standards Measurement Survey (LSMS) and the Demographic and Health Survey (DHS). It's applicability across all settings will remain limited, until data collection efforts are improved and standardized in all developing countries.

On the positive side, in the absence of randomization, the DiD approach will often give reliable estimates of program impact for on-going programs. It is serves as an excellent alternative for new programs where randomization is not possible or as a complement to randomization when programs run for several years and randomization is no longer possible. For on-going programs it provides the quickest turn-around time for policy messages. In this project the DiD evaluations reached completion in just over one year with preliminary results for some being available as quickly as eight months. The major time consuming activity is the merging of large data sets from multiple sources. All of the evaluations used multiple sources of data that needed to be matched and merged. Particularly in the case of the Ukraine and Iran studies, data had to be manually transcribed to electronic format from physical records held by government departments.

The biggest benefit of the DiD approach is the high return to investment in evaluating large scale, long-term programs. Table 3 shows the evaluation cost of the six programs using DiD. Except for the Rwanda evaluation that was designed as a randomized experiment which includes two rounds of data collection, five of the evaluations have costs ranging from $93,740 for the evaluation of Brazil's large scale national program to $205,747 for Peru's national program. A point to note about the Peru program is that household data was collected for the individual level analysis which accounted for $72,360 of the costs. These evaluation costs must be viewed in light of the size of the programs being evaluated and the years of operation of the program; covering almost 85-90% of the target locations in the case of Brazil and Iran and multiple regions in Ukraine, Kenya and Peru. The Ukraine program, though not collecting primary data, has high costs compared to the other DiD evaluations because of expenses for transcribing data to electronic form. A note of caution here is that these should not be interpreted as full economic costs as much of the staff costs are not accounted and it is not possible to compare costs of evaluation and program costs. These programs have also been operational for several years allowing for long-term impacts to be estimated and heterogeneity in treatment effects. The identified impacts also pertain to a much larger population or in some cases the entire target population.

Post-intervention Data

As mentioned earlier, the longitudinal data required by DiD estimators is still not widely available in all regions, limiting the applicability of this approach for many programs. A range of cross-sectional estimators are available to overcome data limitations. Broadly they can be classified into those rely on 'selection on observables' (matching, regression analysis) and those that also account for 'selection on unobservables' (instrumental variables, regression discontinuity, control function approach). Applications of two of these approaches (matching and regression discontinuity) are discussed below.

The first alternative to DiD used in this project is that of matching. This approach is based on the assumption that selection into treatment is fully reflected by observable variables. This cross-sectional approach is applicable when experimental control groups are not available for direct comparison of outcomes. In this case a suitable comparison group is constructed from non-participant individuals based on the similarity of observables characteristics with the treatment group under evaluation. The key assumption underlying this approach is that conditional on a set of covariates, selection into treatment is independent of the outcomes being evaluated. This conditional independence, or ignorability, assumption when combined with an assumption of overlap or common support, that is, the availability of comparison individuals with similar covariate values to the treated group, permits non-parametric identification of the treatment effect (see for example: Cochran and Rubin 1973, Rubin 1973a, 1973b, 2006, LaLonde 1986, Heckman, Ichimura and Todd 1998, Deheija and Wahba 1999, 2002). A range of methods are available for matching. In cases where the number of variables affecting treatment assignment and outcomes is small the two groups can be matched directly on these variables. Most often this is not the case and a large number of variables impact treatment and outcomes. In this case propensity scores are used to reduce the dimensionality of the matching (Rosenbaum and Rubin 1983, 1984). The propensity score is a balancing score representing the conditional probability of allocation to treatment given the observed confounders.

In this project all the programs evaluated using matching rely on propensity scores. The wide range of programs that can be evaluated using this approach is evident in those selected for this project. The Ghana National Health Insurance is similar to those in the DiD category and is a large scale, nationwide program covering 55% of the national population (Mensah et al. 2010). The Ghana program is similar to the Safe Motherhood program in Thailand, implemented in central, general and community hospitals and the Health Services Extension program in Ethiopia (Chandoevwit and Vacharanukulkieti 2009, Admassie et al. 2009). The other program in this category - 'Yeshasvini' insurance program in India (Aggarwal 2010) is smaller in scale and implemented in a specific region as a joint enterprise between the government and other organizations.

The accuracy of matching estimators relies heavily on the similarity between the treatment groups and the selected comparison group. In order to enhance the quality of the analysis the Young Medial Volunteers Program evaluation applies two rounds of matching (Ngoc and Quoc 2010), the first to prune the sample of available treatment and control sites prior to data collection, and the second at the analysis stage. This evaluation uses a combination of primary and secondary data. In the first stage, secondary data from a rural census is used to match intervention and non-intervention communes (villages). At this stage matching is carried out using nearest-neighbours without replacement resulting in 213 pairs of treated and comparison communes. 180 treatment communes with the highest propensity scores are then selected into a further round of matching including all comparison communes (213). The resulting 180 pairs were selected for the survey. The household survey consisted of 13,365 respondents. The second round of matching was used to identify the program impacts by applying three different matching methods – kernel matching, single nearest-neighbour and 3 nearest-neighbours.

In contrast to the YMV program, none of the other matching evaluations use secondary data, primary data on both treatment and controls were collected for the purpose of the evaluations. The matching estimator identifies the average treatment effect over a region of common support. In the case of propensity scores, common support requires that for levels of the probability of participation (propensity score), the probability of observing a non-participant is positive. One way of checking whether the common support requirement is met i.e. the extent of the similarity between the treatment and comparison groups is by comparing the estimated propensity scores across blocks or intervals of the common support region. For example, the Ghana NHIS evaluation was carried out in two administrative districts each (one urban and one rural) from two regions of the country. A random sample of 400 participants and 1600 non-participants from these districts were surveyed and data was collected for eight outcome and ten control variables that impact both participation and treatment. Once propensity scores were estimated, the common support region was divided into four blocks with each block containing both treated and untreated individuals. The mean of the propensity scores between the two groups within each block is then compared. The study finds no significant difference in the averages in any of the groups indicating that the covariates are likely to be balanced between the two groups for the different blocks. The study then uses nearest-neighbour matching and kernel matching to estimate the average treatment effect on the treated.

Propensity score matching is also applicable to multiple treatment arms and comparison groups as used in the Yeshasvini health insurance program. In this program health insurance is provided to members of a co-operative society who can sign-up on a voluntary basis. The evaluation focuses on a range of outcomes covering healthcare utilization, economic well-being, financial protection and surgery/treatment outcomes. The coverage of the program is also broad and includes out-patient care, surgery and maternal health. Thus, the definitions of treatment vary by duration of membership and on being a beneficiary or not – households with membership at the time of the survey, households who had not renewed their membership in the last 3 or more years, and households who had been beneficiaries in the last 4 years. In this program selection of an appropriate comparison group was also not straight forward. Theoretically, co-operative society members eligible for the insurance but who do not participate may be the closest in terms of observables to the treatment group. However, due to the voluntary nature of the program there is the danger of selection bias from time varying unobservable characteristics. For instance, non-participant households eligible to sign-up (members of a co-operative society) may not participate if they have recently undergone surgery and do not expect to face catastrophic health costs in the near future. This group would then be less suitable as a comparison group as opposed to non-eligible households. To examine the sensitivity of the results to different specifications of comparison groups, two different groups were used – co-operative society members who did not take up insurance and non-cooperative society members (who are not eligible for the insurance). By comparing the distribution of the propensity scores from the different treatment arms with both the comparison groups, the study finds that non-participant –eligible households are more similar in distribution to the treatment arms than the non-eligible households. The evaluation collected data on 4,109 households for 400 control variables and the outcome variables.

The choice of observed characteristics is important and should capture all factors that affect both treatment assignment and outcomes of interest but are not affected by the treatment itself (to avoid post-treatment bias). Where this is not the case, matching fails to control for treatment selection. To ensure the conditional independence assumption holds, matching should be carried out on baseline characteristics of both control and treatment groups. Often, pre-treatment (lagged) outcome variables are included in estimating the propensity scores. Such information was however not available for all of the programs in this category. To circumvent this problem the Health Extension Workers Program (HSEP) in Ethiopia included a village level survey of leaders and senior residents to elicit information on pre-intervention village level characteristics that could account for differences in the treatment and control groups. This survey was also used to obtain information on access to markets and social infrastructure. The HSEP is a national program covering about 50% of the rural villages. The selection of the regions was purposive and based on the availability of both treated and non-treated districts in the region. The total sample size includes 3,396 children between the ages of 0-5 and 3,540 women between 15-49 years.

In the absence of random assignment and longitudinal data, the cross-sectional estimator of matching relies on post-intervention level characteristics to select comparable comparison groups. The validity of the conditional independence assumption in such applications must be justified by including all relevant covariates that are likely to influence treatment selection and outcomes and justifying that treatment allocation to different regions or districts was carried out randomly.

Program allocation based on certain criteria such as cut-off points can also be exploited to recover impact estimates. This cross-sectional approach is regression discontinuity design (RDD) where controlling for an observable variable occurs in circumstances where the probability of assignment to the treatment group is a discontinuous function of one or more observable variables (Duflo et al. 2008). The treatment effect is a comparison of mean outcomes of individuals just below with those just above the cut-off points. This approach is applied in the HIV/AIDS teacher training program in Cameroon (Arcand and Wouabe 2010). It exploits the natural experiment generated by the program implementation where teachers in towns having between 1 and 4 secondary schools were trained in communicating HIV/AIDS related information to students while those towns with more than 4 schools received no training. Since the number of schools in a town was determined several years earlier by independent factors, the target population had no control over this factor, the threshold then generates a sort of 'local randomized experiment'. The sample surveyed consists of 2,279 15-17 year old and 2,267 12-13 year olds between grades 1 to 6. 108 schools were surveyed, 56 schools received the teacher training component and 52 served as the control.

As mentioned earlier all these evaluations involved primary data collection. Costs of these evaluations are provided in Table 2. The range of costs for this approach lies between the DiD evaluations at the lower end and the randomized experiments at the upper end. However, in comparison to the DiD approach, the return to investment in matching is compromised if the validity of the approach is questionable due to the lack of pre-program information to control for likely differences. This approach would then score relatively low amongst the viable options and should be implemented only if none of the others are possible. Alternatively, considerable time and resources must be spent to elicit such pre-program characteristics through surveys, as was done in the Ethiopia evaluation.

Mixed Methods

Matching methods and DiD both seek to ensure comparability between the treatment and control groups. Matching aims to improve the balancing of observed characteristics, using methods that are more semiparametric than standard linear regression. DiD allows for imbalance in unobservable characteristics, so long as they remain constant over time, but imbalance in observed covariates is typically handled through linear regression of the DiD specification. The strengths of the two methods can be combined in the evaluation process using a mixed approach: matching can be used to improve the balance in observed covariates and then the DiD regression can be applied to the matched sample to allow for time invariant unobservables that are not captured by the covariates. Two studies in this project – Evaluation of the Safe Motherhood Program in China (Feng et al. 2010) and School Based Malaria Treatment in Mangochi district of Malawi (Simwaka et al. 2009) apply matching methods and DiD, while the Government Social Franchise (GSF) evaluation at commune health stations in Vietnam (Ngo et al. 2009) combines one-to-one matching with multivariate regression and factor analysis. In these evaluations matching is used to select appropriate comparison groups, following which DiD or multivariate regression is applied to estimate program effects. The Safe Motherhood Program in

China did not select counties for treatment based on random assignment so, in order to correct for the possible endogeneity of program placement, treatment and non-treatment counties were matched on baseline characteristics using radius matching on propensity scores. 283 treatment counties were matched with a sample of 1,583 comparison counties. The pruned sample was then used in the DiD estimation applying two estimation strategies, the first, to establish a dose-response relationship between years of treatment and outcomes by including a set of dummy variables representing years, and the second, to explore the channels through which the outcome (maternal mortality ratio) is affected by the treatment (MCH, mother and child health services) the dummy variables are interacted with the MCH variables.

In contrast to the other large programs applying DiD, the School Based Malaria Treatment Program is implemented by the Save the Children Fund in Mangochi district of Malawi. The program covers half the schools in the district. The evaluation uses administrative records at school and student level from 2001/02 to 2005/06. In this application, which is quite similar to the Safe Motherhood Program in China, treated schools were first matched with other non-treated schools. From the pruned sample, 10 student records from each school were randomly selected. Regression adjusted DiD was then applied at the student level to estimate impact of the program on sickness and absenteeism.

The evaluation of the GSF model at health commune stations in Vietnam uses household surveys of potential users and client surveys of actual users to assess improvements in reproductive health and family planning service quality. Unlike the other two evaluations in this category, this evaluation uses primary data collected over three rounds – baseline, six months and one year after the franchise network was established. The sample was selected from two treatment and two control provinces using 1:1 matching of 38 treatment and control commune health stations. Multivariate regression was fitted for the community level indicators – perception of service quality and staff expertise. At the client level, factor analysis was applied to various outcomes – client satisfaction, likeliness to return to the franchise and other quality related factors.

CONCLUSIONS

Choosing an evaluation strategy

Well-designed and conducted randomized experiments recover the average treatment effect on the treated for a specifically chosen set of subjects. An extension of these results is not necessarily guaranteed when such programs are implemented on a much larger scale in the same location and even less when transported for replication in another region or country. The changes in effects could be induced in many ways – scaling up in the same location could lead to general equilibrium effects; different factors could influence outcomes in different locations, especially in developing countries. Identifying universally consistent impacts of specific factors requires replication of these programs in different settings. Further financial investments still need to be made in re-evaluating these programs in local settings either as randomized experiments wherever possible or using non-experimental methods where experiments are not feasible.

Programs that seek to have behavioural impacts on the target population have long-turnaround times; several years of operation may be required before impact can be measured. The Malawi conditional cash transfer program is such an example (Baird et al. 2010). The costs provided in Table 2 represent one year of evaluation and do not include personnel cost of the researchers (these were provided free). The program has been extended through other sources of funding for a further two years of implementation and data collection. Despite the costs and the administrative requirements of implementing experiments, randomized experiments are crucial to evaluating ideas that have not been implemented before or exploring the impact from replicating successful ideas. However, their contribution in evaluating programs over a long horizon or on-going programs is limited and this task typically requires the use of observational data.

Answers to the problems facing health policy in developing countries do not necessarily lie in new programs. Often existing programs can provide insights into success. Within particular countries the results from DiD evaluations of large programs are likely to be more generalisable as compared to pilot randomized experimental evaluations. However, caution must be used in extending lessons from one country to another, often in vastly differing settings. In such cases the external validity of the DiD approach is no more than that of a randomized evaluation. The programs in this project show that if sufficient data are available to control for observable differences and program placement is not endogenous, or factors affecting endogenous program placements can be controlled for, then the DiD approach is most favourable in non-experimental settings. It is economical and, if implemented well, can provide accurate impact estimates with quick turn-around times. They are the least cost way of monitoring impacts over several years and serve as an excellent substitute when randomization is no longer feasible.

The EIHP programs that use pre- and post-intervention data show that, when program allocation is not randomized but longitudinal data are available, selection issues and the endogeneity problem can be addressed by combining methods that capture selection on observables and unobservables. Particularly when treatment assignment is not at the individual level, it is usually based by program operators on observable characteristics – such as pre-program outcome levels, poverty or infrastructure. Matching methods can then be applied on baseline characteristics to control for differences in selection. The DiD approach then controls for time-invariant unobservables that are likely to be confounding factors. In cases where it is possible to exploit a natural experiment, RDD is an option, limited however in identifying the effects for the population immediately around the threshold.

The post-intervention methods such as matching used in the EIHP programs make an important contribution to identifying solutions from existing programs. It encourages evaluations in the many cases where longitudinal data may not be available. However, they depend on finding a suitable comparison group and typically require significant efforts in data collection for both treatment and comparison groups. These methods also require a clear justification of the selection on observables. Until it becomes standard to collect data pre and post intervention for programs implemented in the future, this method is an efficient solution to the evaluation of program impact.

The EIHP project provides policy-makers looking to identify successful innovative programs with evidence on the effectiveness of 19 health programs. These evaluations provide a starting point for evidence based replication of successful programs. However, there are several methodological messages that emerge. Caution must be exercised in replicating programs, as success in one country does not guarantee success elsewhere. Results from evaluations are relevant to the population they are evaluating. Replications must be built around pilot evaluations. Irrespective of the type of program and its scale, evaluations are critical to identifying good ideas. Data collection for the purpose of evaluations must be built into program designs. New programs must look to randomize where possible and if planned for the long term must build in non-experimental methods to continue the evaluations. In the case of existing programs this project shows that a range of methods are available that can be adapted to evaluating a wide array of programs.

Notes:

1. The EIHP project's contribution was 4% of the total cost of the Rwanda study. The remaining was contributed by several other donors.

REFERENCES

Admassie, A., Abebaw, D. and Woldemichael, A.D., 2009. Impact evaluation of the Ethiopian Health Services Extension Program. *Journal of Development Effectiveness*, 1(4), 430-449.

Aggarwal, A., 2010. Impact evaluation of India's 'Yeshasvini' community based health insurance program. *Health Economics*, in press, 10.1002/hec.1605.

Arcand, J-L. and Wouabe, E.D., 2010. Teacher Training and HIV/AIDS prevention in West Africa: Regression discontinuity design evidence from Cameroon. *Health Economics*, in press.

Ashenfelter, O., 1978. Estimating the effect of training programs on earnings. *Review of Economics and Statistics,* 60, 47-57.

Ashenfelter, O. and Card, D., 1985. Using the longitudinal structure of earnings to estimate the effect of training programs. *Review of Economics and Statistics*, 67, 648-60.

Baird, S., Chirwa, E., McIntosh, C. and Ozler, B., 2010. The short-term impacts of a schooling conditional cash transfer program on the sexual behaviour of young women. *Health Economics*, in press, DOI: 10.1002/hec.1569.

Banerjee, A. V. and Duflo E., 2008 The experimental approach to development economics. NBER working paper w14467. National Bureau of Economic Research, Inc.

Basinga, P., Gertler, P.J., Binagwaho, A., Soucat, A.L.B., Sturdy J.R. and Vermeersch, C.M.J., 2009. Impact of performance based financing in Rwanda: health facility level analysis. GDN working paper 32. Global Development Network.

Blundell, R. and Costa-Dias, M., 2008. Alternative approaches to evaluation in empirical microeconomics. CEMMAP working paper CWP26/08. Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Bourguignon, F. and Sundberg, M., 2007. Aid effectiveness - opening the black box. *American Economic Review*, 97, 316–321.

Chandoevwit, W. and Vacharanukulkieti, K., 2009. An evaluation of a safe motherhood hospital program. GDN working paper 23. Global Development Network.

Chen, S. and Ravallion, M., 2007. Absolute poverty measures for the developing world, 1981-2004. Policy Research Working Paper WPS4211. The World Bank.

Cochran, W. and Rubin, D., 1973. Controlling bias in observational studies: a review. *Sankhya: The Indian Journal of Statistics*, Series A, 35(4), 417-46.

Correa C.J., Pinto, D., Camacho, J., Quintero, J., Rondón, M. and Salas, L., 2010. A couple of squirts a day keep the doctor away: a cluster randomized controlled trial of alcohol based hand sanitizer gel for prevention of infectious diseases in children. GDN working paper 39. Global Development Network.

Deaton, A.S., 2008. Instruments of development: randomization in the tropics, and the search for the elusive keys to economic development. NBER working paper w14690. National Bureau of Economic Research, Inc.

Deheija, R. and Wahba, S., 1999. Causal effects in nonexperimental studies: re-evaluating the evaluation of training programs. *Journal of the American Statistical Association,* 94, 1053-62.

Deheija, R. and Wahba, S., 2002. Propensity score matching methods for nonexperimental causal studies. *Review of Economic Studies,* 84, 151-61.

Díaz, J.J. and Jaramillo, M., 2009 Evaluating interventions to reduce maternal mortality: evidence from Peru's PARSalud programme. *Journal of Development Effectiveness*, 1(4), 387-412.

Duflo, E., Glennerster, R., and Kremer, M., 2007.Using Randomization in Development Economics Research: A Toolkit. *In*: T.P. Schultz and J.A. Strauss, ed. *Handbook of Development Economics*. Amsterdam: Elsevier, vol. 4, chap. 61, 3895–3962.

Feng, X.L., Shi, G., Wang, Y., Xu, L., Luo, H., Shen, J., Yin, H. and Guo, Y., 2010. An impact evaluation of the Safe Motherhood Program in China. *Health Economics*, in press, DOI: 10.1002/hec.1593.

Heckman, J. J., 2008. Econometric causality. CEMMAP working paper CWP1/08. Centre for Microdata Methods and Practice, Institute for Fiscal Studies.

Heckman, J.J., Ichimura, H. and Todd, P. E., 1998. Matching as an econometric evaluation estimator. *Review of Economic Studies, 65, 261-94.*

Heckman, J.J. and Robb, R., 1985. Alternative models for evaluating the impact of interventions, *In*: J. J. Heckman and B. Singer (eds.), *Longitudinal Analysis of Labour Market Data.* Cambridge: Cambridge University Press, chap. 4, 156

Heckman, J.J., and Vytlacil, E., 2007. Econometric evaluation of social programs. *In*: J,J Heckman and E. Leamer, ed. *Handbook of Econometrics*, Amsterdam: Elsevier, vol 6B.

Imbens, G.W. and Wooldridge, J.M., 2008. Recent developments in the econometrics of program evaluation. NBER Working Paper 14251. National Bureau of Economic Research, Inc.

Holland, P.W., 1986. Statistics and causal inference. *Journal of the American Statistical Association,* 81, 945-60.

Lalonde, R.J., 1986. Evaluating the econometric evaluations of training programs with experimental data. *American Economic Review,* 76, 604-20.

Mensah, J., Oppong, J.R. and Schmidt, C.M., 2010 Ghana's National Health Insurance Scheme in the Context of the Health MDGS – An Empirical Evaluation Using Propensity Score Matching. *Health Economics*, in press.

Ngo, A., Phan, H., Pham, V., Trinh, T. and Truong, K., 2009. Impacts of a government social franchise model on perceptions of service quality and client satisfaction at commune health stations in Vietnam. *Journal of Development Effectiveness*, 1(4),413-429.

Ngoc, U.V., and Huy, V.Q., 2010. Health impact evaluation: Project Young Medical Volunteers for Vietnam Rural Mountain. GDN working paper 38. Global Development Network.

Nizalova, O. and Vyshnya, M., 2010. Evaluation of the impact of the Mother and Infant Health Project in Ukraine. *Health Economics*, in press, DOI: 10.1002/hec.1609.

Oduor, J., Kamau, A. and Mathenge, E., 2009. Evaluating the impact of micro-franchising the distribution of anti-malarial drugs in Kenya on malaria mortality and morbidity. *Journal of Development Effectiveness*, 1(3), 353-377.

Ravillion, M., 2007. Evaluating Anti-Poverty Programs. *In*: T. P. Schultz and J. A. Strauss, ed. *Handbook of Development Economics*, Amsterdam: Elsevier, vol. 4, chap. 59, 3787-3846.

Rocha, R. and Soares, R., 2010. Evaluating the impact of community based health interventions: Evidence from Brazil's Family Health Program. *Health Economics*, in press, DOI: 10.1002/hec.1607.

Rosenbaum, P.R. and Rubin, D.B., 1983. The central role of the propensity score in observational studies for causal effects. *Biometrika,* 70, 41-55.

Rosenbaum, P.R. and Rubin, D.B., 1984. Reducing the bias in observational studies using subclassification on the propensity score. *Journal of the American Statistical Association,* 79, 516-24.

Rubin, D. B., 1973a. Matching to remove bias in observational studies. *Biometrics,* 29, 159-83.

Rubin, D.B., 1973b. The use of matched sampling and regression adjustments to remove bias in observational studies. *Biometrics,* 29, 185-203.

Rubin, D.B., 1974. Estimating causal effects of treatments in randomized and non-randomized studies. *Journal of Educational Psychology,* 66, 688-701.

Rubin, D.B., 2006. Matched sampling for causal effects. Cambridge: Cambridge University Press.

Salehi-Isfahani, D., Abbasi-Shavazi, M.J. and Hosseini-Chavoshi, M., 2010. Family planning and rural fertility decline in Iran: A study in program evaluation. *Health Economics*, in press, DOI: 10.1002/hec.1613.

Simwaka, B. N., Simwaka, K. and Bello, G., 2009. Retrospective analysis of a school-based malaria treatment programme demonstrates a positive impact on health and education outcomes in Mangochi district, Malawi. *Journal of Development Effectiveness*, 1(4), 492–506.

Teerawattananon, Y., Leelukhanaveera, Y., Thavorncharoensap, M., Hanvoravovongchai, P., Ingsrisawang, L., Tantivess, S., Chaikledkaew, U., Mohara, A., Lertpiriyasuwat, C. and Pimsawan, N., 2009. Provider-initiated HIV/AIDS counselling and testing at healthcare facilities in Thailand: a cluster-randomization trial. *Journal of Development Effectiveness*, 1(4), 450-469.

Thornton, R., Hatt, L., Islam, M., Field, E., Solís, F. and González Moncada, M.A., 2010. Social security health insurance for the informal sector in Nicaragua: A randomized evaluation. *Health Economics*, in press, DOI: 10.1002/hec.1635.

United Nations Department of Economic and Social Affairs, 2008. Millennium Development Goals Report.

White, H., 2006. Impact evaluation: The experience of the Independent Evaluation Group of the World Bank. Technical Report 1111. The World Bank.

Table 1: The EIHP Programs

Program	**Country**	**Health Issue**	**Implementing Agency**
Safe Motherhood Program	China	Maternal Care	Government
Reproductive Health Capacity	Vietnam	Maternal Care	Government
The PARSalud Program	Peru	Maternal Health	Government
Mother and Infant Health Project	Ukraine	Maternal Health	Government
Safe Motherhood Hospital Program	Thailand	Maternal Health	Government
Alcohol Based Hand Sanitizers	Colombia	Child health	Pilot randomized evaluation
PSF Family Health Program	Brazil	Healthcare Delivery	Government
Health Services Extension Program	Ethiopia	Healthcare Delivery	Government
Family Planning Program	Iran	Healthcare Delivery	Government
Young Medical Volunteers	Vietnam	Healthcare Delivery	Government
Performance Based Financing	Rwanda	Healthcare Delivery, HIV/AIDS	Government
'Yeshasvini' Community Based Health Insurance	India	Health Insurance	Local initiative
National Health Insurance Scheme	Ghana	Health Insurance	Government
Social Security Health Insurance	Nicaragua	Health Insurance	Pilot randomized evaluation
Micro-franchising the distribution of anti-malarial drugs	Kenya	Malaria	Government
School-based Malaria Program	Malawi	Malaria	International NGO
Conditional Cash Transfers	Malawi	Schooling, Sexual Behaviour, HIV/AIDS	Pilot randomized evaluation
HIV/AIDS Education Program	Cameroon	HIV/AIDS	International Organizations/Government
Provider-initiated Voluntary HIV Counselling and Testing	Thailand	HIV/AIDS	Pilot randomized evaluation

Table 2: Evaluation Costs

Primary Data:			
Randomized Experimental Evaluations			
	Total Cost (US$)	**Primary Data Collection (US$)**	**Primary Data Collection as % of budget**
Social Security Health Insurance (Nicaragua)	405,392	171,370	42%
Conditional Cash Transfers (Malawi)	353,560	341,658	97%
Provider-initiated Voluntary HIV Counselling and Testing (Thailand)	318,680	119,800	37 %
Alcohol Based Hand Sanitizers (Colombia)	261,010	81,300	31%
Non-experimental studies: Post- intervention Data (Matching)			
	Total Cost	**Primary Data Collection**	
Safe Motherhood Hospital Program (Thailand)	188,730	49,300	26%
Young Medical Volunteers (Vietnam)	249,860	195,140	78%
'Yeshasvini' Community Based Health Insurance (India)	153,542	76,186	50%
National Health Insurance Scheme (Ghana)	148,525	79,180	53%
Health Services Extension Program (Ethiopia)	132,596	53,505	40%
Regression Discontinuity Design			
	Total Cost	**Primary Data Collection**	
HIV/AIDS Education Program (Cameroon)	65,530	33,030	50%
Secondary Data:			
Non-experimental studies: Pre-and Post- intervention Data (Difference-in-Differences)			
	Total Cost	**Data Collection**	
Performance Based Contracting (Rwanda)	2,348,981[1]	1,534,443	65%
The PARSALUD Program (Peru)	205,747	72,360 (data collected for individual level analysis)	
PSF Family Health Program (Brazil)	93,740	----	

Mother and Infant Health Project (Ukraine)	187,088	64,860 (transcribing data and collection from regions)	
Family Planning Program (Iran)	120,000	----	
Micro-franchising for the Distribution of Anti-malaria Drugs (Kenya)	136,601	----	
Mixed Methods			
	Total Cost	**Data Collection**	
School Based Malaria Program (Malawi)	97,507	49,800 (cost benefit analysis)	
Safe Motherhood Program (China)	135,000	----	
Reproductive Health Capacity (Vietnam)	216,373	111,673	