Economics and Epidemiology: Two Sides of the Same Coin or Different Currencies for Evaluating Impact?

Marcella Vigneri[1], Edoardo Masset[1], Mike Clarke[2], Josephine Exley[3], Peter Tugwell[4], Vivian Welch[5], Howard White[1,5]

November 2018

[1] Centre of Excellence for Development Impact and Learning, UK
[2] Queen's University Belfast, UK
[3] London School of Hygiene and Tropical Medicine, UK
[4] University of Ottawa, Canada
[5] Campbell Collaboration, Canada and India

Table of Contents

Abstract

This paper discusses differences in the practice, conduct and reporting of impact evaluation studies by economists and epidemiologists. Recognizing and defining the boundaries in these disciplines' differences, and the missed opportunities for collaborative research, is an important conversation for the Centre of Excellence in Impact Evaluation and Learning (CEDIL) and its goal to promote interdisciplinary research to strengthen evaluation evidence in international development.

This paper develops its key arguments from the proceedings of a workshop organized in November 2017 by CEDIL, which was attended by evaluation specialists from both disciplines. The methodology used in this short study is to compare evaluation studies in similar areas from a selection of epi-econ pairs of peer reviewed papers, discuss evaluation practices reported in systematic reviews in each discipline, and suggest where identified differences could be used as opportunities for cross learning. The paper also suggests the rationale for an econ-epi evaluation glossary to overcome seeming differences in terminology and highlight similar methods that are present in both disciplines but for which the term widely used in one discipline is not known in the other.

Introduction

─────

Economics and epidemiology share a research interest in investigating and measuring the impact of interventions in international development. In a 2016 contribution to the *American Journal of Public Health*, Spiegelman (2016) stated that similarities in the conduct of impact evaluations across disciplines like epidemiology and economics overwhelm the differences, and that the distinctions that have been made confound the underlying common ground. A similar claim can be made in relation to the differences between economics and clinical epidemiology, as defined by Feinstein and Sackett (Feinstein, 1995; Sackett, 2002). Spiegelman defined impact evaluation as a method for *assessing the efficacy and effectiveness of an intervention in terms of intended and unintended health, social, and economic outcomes, and involving the explicit statement of a counterfactual* (Spiegelman, 2016).[1]

While the principles of Spiegelman's definition are clear, it is also important to recognise and understand the differences that exist between how economists and epidemiologists conduct evaluations. Although these disciplines often share an interest in similar research topics and apply statistical tools to measure impact, they often use evaluation tools differently to answer the same questions, because of the research nature of each discipline. This, in turn, is reflected in different practices in the conduct and reporting of evaluations.

This paper is motivated by the desire to identify some of the boundaries between the two disciplines, explain their rationale with some examples drawn from peer reviewed papers, and suggest a roadmap of activities to support cross-discipline learning in the shared practice of evaluation work. Given the high degree of heterogeneity in each discipline, the overarching thematic focus of the paper is to look at interventions conducted and evaluated on global health issues, which have been researched by both economists and epidemiologists.

The paper also suggests the development of an Econ-Epi evaluation glossary, a tool to facilitate a better understanding of the terms used in each discipline and to compare definitions across different evaluation practices. The annex presents a table with suggestive examples of where this glossary could be a valuable addition for evaluation practitioners across disciplines, and how this could become an online tool to help achieve the CEDIL goal to promote cross disciplinary work in impact evaluation.

The rest of the paper is organised as follows. Section 2 presents the methodology used to define boundaries and identify potential for cross learning across disciplines.

─────────────────

[1] Evaluators distinguish between outcomes and impact, with the latter being longer-run welfare measures. Neither economists nor epidemiologists generally make this distinction, so impact evaluation refers to the impact on outcomes of interest which may be either short or long-run (for further discussion see White, 2010).

Findings are presented in Section 3 and 4. Section 3 provides an overview of areas where economists and epidemiologists differ in the practice of impact evaluations, while Section 4 compares differences in the conduct of systematic reviews undertaken by economists and epidemiologist. Section 5 introduces and discusses the importance of a unified Econ-Epi evaluation glossary (included in the annex) to support a common lexicon across the two disciplines. Section 6 concludes the paper, reflecting on learning opportunities across the two approaches, and suggesting a roadmap for future cross-disciplinary work. Finally, Section 7 discusses the paper's implications for CEDIL and the UK's Department for International Development (DFID).

Methodology

To identify differences and similarities between the disciplines we examined recent work undertaken by groups known to the authors, reviewed the literature to identify evaluations and systematic reviews undertaken by economists and epidemiologist on similar topics and finally held a workshop to reflect on the emerging findings.

Recent work
First, we examined reflections from two recent pieces of work on differences in the evaluation practices between the disciplines. One piece looked at the gap between the research methods of Impact Evaluation in Economics and Clinical Epidemiology and concluded that whilst both disciplines are interested in generating evidence on global health and international development and use approaches that are valid and reproducible, their differences in research approaches have led to a lack of acceptance of each other's work and a longer time for the transfer of innovations from one area to the other. This debate proposed the development of a methodology that compares the different approaches, highlights similarities and differences, and suggests where each field can learn from the other. The second piece of work was a recently published paper jointly written by a group of economists and epidemiologists based at the London School of Hygiene and Tropical Medicine (Powell-Jackson *et al.*, 2018), which specifically looked at differences in the practice of randomised trials across the two disciplines. The authors suggest that evidence-based public health can be strengthened by a better understanding of the differences in the use, practice and reporting of results from randomised trials, and in so doing promote cross-disciplinary learning.

Review
In the second stage we reviewed published evaluations on topics studied by both economists and epidemiologist that addressed similar research questions. The four topics identified were: (1) bed nets, (2) WASH (Water, Sanitation and Hygiene), (3) anti-retroviral treatment (ART) for the reduction of HIV incidence, and (4) gender interventions to reduce the risk of Intimate Partner Violence (IPV) .We reviewed pairs of papers (one economic and one epidemiological) that evaluated similar public health interventions in each of the specific topics. To be classified as an economics paper, authors had to be affiliated to an economics department and the paper had to be published in an economics journal, while epidemiological papers had to be published in a medical or epidemiological journal.

In parallel, we identified systematic reviews of interventions that both economists and epidemiologists have an interest in and looked at how these are differently used across disciplines. The topics considered were: (1) conditional cash transfers, (2) school feeding, and (3) health insurance programmes.

It is important to stress that the paired papers and systematic reviews used to illustrate this paper were not selected to produce a set of generalisable conclusions on the differences across the two disciplines reviewed below. The intention was to use them as examples to illustrate the occurrence of different methods and approaches in the evaluation practice, and to suggest how these differences could inspire future opportunities for interdisciplinary work. The studies cited, therefore, must not be taken as representative of the *general* impact evaluation practice adopted by each discipline.

Workshop

The third step was a workshop organised by CEDIL in November 2017. The event gathered economists and epidemiologists (both clinical and those specialising in public health) with extensive experience in undertaking evaluation research[2] to reflect on the evidence from the first two stages to identify a roadmap of possible activities for cross learning.

This paper emerged as a write up of and elaboration from the workshop proceedings and, unless stated otherwise, largely reflects the opinions and suggestions of the workshop attendees. The objective of this paper is therefore to contribute to the thinking around how innovative interdisciplinary impact evaluation methods can be strengthened.

---

[2] Participants in the CEDIL workshop: Marcella Vigneri (CEDIL), and Edoardo Masset (CEDIL), Orazio Attanasio (Institute for Fiscal Studies), Mike Clarke (Queen's University Belfast), Tessa Edejer (World Health Organization), Josephine Exley, Richard Hayes, and Tim Powell Jackson (all at the London School of Hygiene & Tropical Medicine), Peter Tugwell (University of Ottawa), Hugh Waddington (International Initiative for Impact Evaluation), Vivian Welch (Campbell Collaboration and University of Ottawa), and Howard White (Campbell Collaboration).

Comparing Economic and Epidemiological Approaches to Evaluation

A great degree of heterogeneity exits within economics and epidemiology; this paper focuses on a specific category of each discipline. For epidemiology, the focus is on clinical and social epidemiology[3] relevant for public health issues in developing countries. For economics, the focus is on development economics applied to measuring the impact of socio-economic interventions. These categories define the conceptual framework of the discussion in the rest of the paper.

How do economists and epidemiologists conduct evaluation studies? When the practice of impact evaluations was new to economists, the interest was in evaluating interventions to measure the success of a program, and this was reflected in the content of evaluations of the specific intervention (Miguel and Kremer, 2004). More recently, the focus on evaluations in economics has shifted to testing economic theory using empirical work, and to differentiate less effective interventions from successful ones to help improve existing programs. This has become a key requirement to publish economic evaluations in a top economic journal, alongside the interest to determine whether a programme has an impact (on a set of key health and social outcomes), and to quantify *how large* that impact is. Epidemiologists operating in public health conduct evaluations as a process for determining, as systematically and objectively as possible, the likely future relevance, effectiveness, efficiency, and impact of activities with respect to established health outcomes (Centres for Disease Control and Prevention).

Although impact evaluations conducted in both disciplines share the common goal of identifying the presence or absence of programme success, their journey to the estimate of impact differs in many ways. The CEDIL workshop identified eight key areas where differences in evaluation practices were most obvious: (1) the use of theory to underpin empirical models used in the evaluation; (2) the generalisability and transferability of findings from the specifics of an individual case study; (3) the definition of research questions; (4) study design, and endpoint/outcome measure of impact; (5) power calculations to ensure that studies have adequate power to detect statistical significance; (6) publication of protocols and pre-specified outcomes; (7) reporting findings, and timing of publications; and (8) replicability. We discuss each of these below

---

[3] Clinical Epidemiology is a medical science that studies the frequency and determinants of disease development, as well as the diagnostic and therapeutic approaches to disease management in clinical practice (Sackett, 2002).
Social epidemiology focuses on the effects of social-structural factors on states of health. The central question to be answered is what effect do social factors have on individual and population health (Cassidy *et al.*, 2011)

to highlight differences, with a reminder that any comparisons suggested are not intended to claim universal truths or to generalise across the spectrum. Counter examples no doubt exist but we believe the issues we highlight may resonate with readers and support constructive thinking in building interdisciplinary methods to generate stronger evidence from impact evaluation research.

## 2.1 Using Theory to Underpin the Evaluation Problem

In best practice economics, theory is used to provide a rationale for the analysis, interpret the experimental results of an evaluation, and expand the usefulness of the experiment. Economists also use theory to generalise beyond the experiment, for example by making predictions about the effect of a future change in the design of the policy.

Although the practice of grounding research in theory is not universal among economists, it is generally considered the gold standard to inform the design of a study, guide the empirical testing and strengthen the analysis underlying impact evaluations. Papers in economics journals will usually have a section titled 'The Model' which outlines the theory being tested. In the jargon of economics, the model provides the assumptions underlying the data analysis. One example of this is work by Attanasio, Meghir and Santiago (2012): an evaluation of the PROGRESA Conditional Cash Transfer (CCT) programme. The intent of their paper is to estimate a structural model of education choices using data from the PROGRESA randomised experiment, and to then use the model to simulate the effect of changes to some of the parameters of the programme (the use of monetary incentives) to incentivise school enrolment. In this case, the use of experimental variation in the data allows estimation of structural models that offer a richer policy analysis, and a better understanding of the mechanisms driving the effects. For example, another study by Attanasio *et al.*, (2015) estimates a structural model of skill production that finds a positive impact of an early childhood stimulation programme (which was trialled using a randomised design), and explained by increases in the level of parental investment, rather than the way in which skills are produced.

In epidemiology there are also examples where theory is developed and used to test the assumptions behind behavioural interventions in a similar way to economists (Eccles *et al.*, 2005). However, social and clinical epidemiologists also use theory to explain the health status of populations in societal and ecological context, rather than explaining why specific individuals become ill or stay healthy. In clinical epidemiology, for example, this translates to looking at transmission mechanisms capturing the dynamic nature and spread of diseases and to incorporate positive and negative feedback characteristics of infectious processes (Dawood *et al.*, 2012). Epidemiology also uses theory to inform clinical decision analysis (e.g. by using decision trees, Markov models to assess the probability of transitions from one state to another, or Bayes modelling in network meta-analyses) or for modelling from surrogate outcomes (e.g. blood tests such as viral load in HIV to humanely-meaningful outcomes). There are also instances where epidemiology uses theory to forecast outcomes that occur under different scenarios without making any assumption, as seen in economics, about

individual behaviour and social interactions. Figure 1, for example, illustrates a typical section of a theory of change; carrying out the intervention brings about behaviour change, which leads to better health.

Figure 1. A Causal Chain representation of atypical logic model in Epidemiology

| Intervention X | — | Behaviour change | — | Better health |
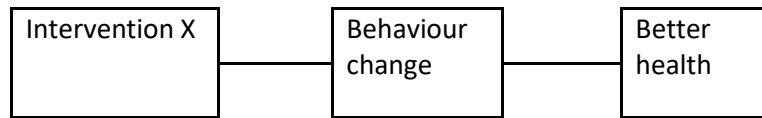
A statement of causal effects depicted as above would not be seen in how economists use theory. Economists most usually present their models as sets of equations, often introducing and explaining the functional form of each equation. The causal pathways - which may include non-linear relationships, lags and feedback loops (known as simultaneity) – are embedded in the equations.

Davies (2018) provides a critique of conventional representations of theory of change, such as in Figure 1, because of the inadequate attention paid to the causal mechanisms implied by the arrows joining the boxes. A theoretical approach would situate the analysis of how the intervention might be expected to affect behaviour by using a theory of behaviour change. There may indeed be competing theories. For example, information campaigns might provide new knowledge, but may also motivate people to act on knowledge they already have or encourage peer effects through discussion of the information. Knowing which of these theories is correct can help shape the messaging to reinforce the active causal mechanism.

One example of how epidemiologists use theory is in the paper by Morris *et al.* (2004), which looks at the impact of *Bolsa Alimentação,* (a national health-related CCT programme in Brazil since 2003) on growth among children of beneficiary households. The paper opens by outlining how poverty-related factors such as lack of access to nutritionally rich diets, inadequate infant feeding practices, and repeated illness, are all contributing factors to stunting. Therefore, based on the assumption that children from poorer families would benefit from significant improvements in living conditions, such as food supplements or their adoption, the authors test the hypothesis that the direct transfer of money to very poor families through the programme would lead to an improvement in the growth of their children. This resonates the causal chain in Figure 1: cash transfers (the intervention) improve living standards (by incentivising behaviour change) and so improve child nutrition (better health). Theory here is not used to *model* the impact of the change in incentives on household *decision making* in the way Attanasio et al. do (as described above), but rather points to the order of causation that follows from introducing a specific incentive that is expected to alter behaviour.

The different use of theory across the two disciplines is important as it reveals opportunities for cross learning: both disciplines recognise that identification of causal links is crucial for evaluation, however this is achieved through a different set of tools. In economics, theory plays an important role in achieving identification by looking at the

most plausible model and set of assumptions that will ensure identification and establish causality. In epidemiology, theory emphasizes the causality between exposure to an event, change in subjects' behaviour, whilst the health outcome is the variable of interest, against which to measure impact of an intervention.

A recent scoping review found 82 different theories of behaviour change (Davis *et al.*, 2015). The adoption of a mid-level theory approach being developed as part of CEDIL's work (Davey *et al.*, 2018) encourages researchers to engage more seriously with theory. As argued below, referring to a general theory to frame the evaluation of a specific intervention helps with the transferability of findings.

The proposed use of mid-level theory resonates with Deaton's critique of impact evaluations that we should be interested in mechanisms, not whether a specific project in a specific time and place works or not. For example, he wrote "the analysis of projects needs to be refocused towards the investigation of potentially generalizable mechanisms that explain why and in what contexts projects can be expected to work" (Piketty and Deaton, 2014). And, indeed, some authors have explicitly stated that their focus is not the intervention per se (which may not be replicable) but the insight which comes from the study of its impact (Card, DellaVigna and Malmendier, 2011).

There are several examples of an evaluation conducted in economics where the authors extrapolate from the specifics of their findings to test theory and provide recommendations on how to improve the design of similar interventions in the future. One well-known example is the study of teacher absenteeism in India by Duflo, Hanna and Ryan (2012). In this piece, teachers' attendance was monitored by requiring them to take a date and time stamped photo each day with their class, and pay was conditional upon attendance. This scheme was more effective in reducing absenteeism than an award scheme administered by head teachers, since head teachers gave everyone the award regardless of performance. In their conclusion, the authors explicitly recognise that using cameras more widely may prove difficult.

A second example is the Ashraf, Bandiera and Jack (2014) study of social marketing of female condoms in Lusaka. The study compared financial incentives to hairdressers versus social recognition through public display of stars and awards. Social recognition had a larger impact on sales than financial incentives. In their conclusions, the authors do not discuss the specific intervention at all, instead, they discuss aspects of context which they believe will affect the effectiveness of non-financial incentives, recommending they be adopted by others: "While we implemented a specific type of non-financial rewards, the general design principles are easily replicable and adaptable to other settings" (Ashraf, Bandiera and Jack, 2014).

But there are obvious limitations to this approach. Finding a case which may be interpreted as consistent with a specific theory may not rule out competing theories which have not been tested. Moreover, we do not know the extent to which the findings can be generalized to other settings. The generalisability of impact evaluation results is another issue of central importance that we now turn to.

2.2 Generalisability and Transferability

Impact evaluations should be internally valid (i.e., design and conduct should attempt to minimize the possibility of bias) but to be useful, whether assessing clinical, global health, or social policy interventions, the results must also be relevant to a definable group of units in similar settings. The second requirement is generally referred to as external validity, applicability, or generalisability.

In clinical epidemiology, the careful review of trials for internal validity is regarded as essential for assessing the quality of the research. However, the question of whether the findings of a study can be applied to different populations or under different conditions is of equal importance for establishing whether comparisons of treatment are different between patients groups or clinical situations. Because empirical data to test effect modifiers are rare, external validity often remains a case of clinical judgement. According to a recent contribution in the economics literature (Banerjee, Chassang and Snowberg, 2016), external policy advice is equally subjective but can and should be laid out as structured speculation, but the best that can be done is for economic experiments to lay out beliefs of what might work and where, to avoid idle speculation for example by explicitly stating four dimensions of external validity: (1) is the experiment scalable, (2) is what are the treatment effects on a different population, (3) what are the treatment effects on the same population under different circumstances, and (4) what is the effect of a different by-related treatment? This approach would go a long way to avoiding generalizable claims of interventions that may work universally despite plenty of evidence from systematic reviews of their limited impact elsewhere (e.g. Welch *et al.*, 2013) .

We now illustrate another aspect of the different approaches to generalisability and transferability in the two disciplines with examples from two studies on bed net interventions. The epidemiology paper, which looks at bed net use (Krezanoski, Comfort and Hamer, 2010), finds that the provision of incentives for the use of ITNs (insecticide treated nets) significantly increased the probability of ITN use. The discussion in the paper does not comment on how generalizable or under what conditions this outcome might be transferable to predict the change in the proportion of households using ITNs in other contexts. The epidemiologists attending the CEDIL workshop suggested this to be consistent with the practice in epidemiology to avoid translating the results of one study to other settings, especially when it is not possible to discount how complex generalisation might be. The economic paper selected (Dupas, 2009) looks at take-up rates of ITNs following their introduction as subsidised malaria-control devices among rural households in Kenya. The papers objective was to test whether the demand for the devices was price elastic (that is, to see if it varied depending on changes in the cost of bednets), and if the variation in demand was correlated with how the marketing messages were framed. In her concluding remarks, Dupas commented on the presence of significant price elasticity to the cost of ITN and on the absence of variation in take-up rates due to the marketing messages used. The findings of the impact evaluation are discussed in relation to those reported in other economic case studies and are used to explain more widely the reason why poor households, who have limited savings, systematically underinvest in health services.

The argument above about the generalizability of evaluation findings in economics and epidemiology, however, does not intend to draw a universal judgment on the issue from the few examples in the literature that we have drawn on. Rothwell (2005), for example, rightly suggests that generalizability (i.e. external validity) is a slippery concept in epidemiology and medical science; the understanding of the determinants of the external validity of a randomised trial require clinical expertise and usually depends on a detailed understanding of the specific clinical condition under study and its management in routine clinical practice. For example, to be clinically useful, the result of a trial must also be relevant to clinical practice, i.e. be reasonably likely to be replicated when applied to a definable group of patients in a particular clinical setting. For some interventions (such as lowering blood pressure in chronic uncontrolled hypertension), the benefits have been shown to be generalizable to most patients and settings, but the effects of other interventions will often depend on factors such as the characteristics of the patient, or the setting of treatment. Rothwell concludes that the greatest current problem is, perhaps, that clinical trials are often not reported in sufficient detail to allow clinicians to judge to whom the results can reasonably be applied (see also (Hoffmann *et al.*, 2014).

The general conclusion on generalizability which emerged from the CEDIL workshop was that, while it is true that epidemiologists tend to be cautious in making statements about generalizability and transferability of findings because of the complexity of over concluding from the results of one study to other settings (e.g. Awasthi *et al.*, 2013), they often situate the findings of randomised trials in the context of updated reviews of similar research (e.g. by using systematic reviews). This practice, on the other hand, is less common in economics, although the complexity of the interventions they assess is certainly no less than that for epidemiology. The economists attending the CEDIL workshop also noted that economists remain reluctant to use and produce systematic reviews because of their scepticism on the robustness of the methods employed in this type of research, and because they recognise the challenges involved in comparing highly heterogeneous studies.[4] The more general observation which emerged from the workshop on the issue of transferability was for both disciplines to adopt the good practice of establishing under what conditions the findings of one study may be generalizable to other contexts.

2.3 Research Questions

The third difference identified between evaluations in economics and epidemiology is in the type of research questions that each discipline seeks to answer. On this matter, participants in the CEDIL workshop noted that impact evaluations in economics will test models of human behaviour using real life data to explain behaviour, and phrase their research questions accordingly. Epidemiologists will focus directly on how a given intervention has impacted the outcome of interest, and less frequently elaborate the

---

[4] Having said that, some economists are increasingly endorsing the review methodology, especially where there are sufficient studies to analyse heterogeneity. A good example is the Campbell review by Baird *et al.* (2014) of conditional cash transfers where the authors code interventions by the degree to which conditionality is monitored and enforced, showing the greater the enforcement the greater the impact.

discussion on the behavioural change mechanism that led to that impact. This approach is reflected not just in how the research questions are framed for each discipline's evaluation studies (and in the type of outcome variables used to measure impact), but also in the context information and supporting information that is provided to justify the need to answer these questions.

To illustrate this point, we looked at two studies examining similar intervention which looked at similar research questions: the impact of WASH interventions on diarrhoea-related outcomes among younger siblings of school-going children (epidemiology paper), and the impact of providing universal access within a village to hygienic latrines and in-home piped water on the incidence of treated diarrhoea episodes (in the economics paper). The debate emerging in the CEDIL workshop from discussing these two examples was that evaluations in epidemiology remain very focused on the specifics of the intervention that is being assessed (e.g Hoffmann *et al.*, 2014), limiting the details around the wider context of a study, possibly assuming that the specialised readers of these papers will have sufficient existing knowledge of the topic or intervention being presented.

The epidemiology paper by Dreibelbis *et al.* (2014) assessed the health and educational impacts of two WASH improvement interventions carried out in schools in Kenya on the prevalence of diarrhoea. The indicator selected to measure impact was a 1-week period prevalence rate of diarrhoea episodes among children of school age, measured from a sample of individuals who were interviewed twice in 26 months. The body of the paper focuses mainly on explaining the study design and the methods adopted for the trial, with the section on results explaining in detail different statistics for the odds of diarrhoea associated with the survey population in different arms of the trial.

In the economic paper sampled in the same thematic area, Duflo *et al.* (2015) evaluated a village-level intervention that promotes adoption of household latrine and bathing facilities, a community water tank, and a distribution system that supplies piped water to household taps. The authors detail the typical sequence of events for implementing the intervention, explaining in detail the procedures adopted by extension workers, the expected responses of village leaders, with an elaboration of the process by which villagers begin construction of household latrines and bathing facilities under the intervention, and a costing exercise to show the financial requirements for households constructing latrines and bathing facilities. The paper offers a great deal of detailed explanation of the economic study design, of the sampling and evaluative method procedures, which takes the reader through the rationale of why and how the intervention is expected to change the behaviour of the targeted households.

This difference between the two papers may reflect differences in both the experience of their authors and the expected readership. Dreibelbis has published over twenty papers on WASH, and so is very familiar with the nature of the intervention and likely assumes the same of readers. Duflo has published randomised trials across a wide range of sectors, but the specifics of the WASH intervention may be new to her, and she would be right to assume the same for readers of the NBER working paper, where she published her findings.

The different framing of the research questions across disciplines is also mirrored in the type and number of outcome variables measured. For example, CEDIL workshop participants noted that economic studies often tend to analyse several measures of impact that are correlated with potential externalities of the project, or unintended (positive and negative) effects that result from the intervention. In epidemiology, papers typically report a single primary outcome which is investigated and gets most emphasis, and some secondary measures of impact which may also be discussed and published in separate papers.

Moreover, economists often look at a variety of outcome measures, analysed with both relative and absolute statistics, to capture the primary expected impact and any additional effects that might result from the intervention. For example, Duflo *et al.* (2015) used recorded doctor or nurse visits for episodes of diarrhoea as the primary outcome variable and three additional health outcomes measures (number of monthly cases of diarrhoea, malaria and fever). On the other hand, epidemiologists, given the nature of the problems they research, place more emphasis on relative, rather than absolute statistics because these relative statistics, such as hazard ratios, odds or risk ratios, and survival rates (Barratt *et al.*, 2004), are more stable measures for generalising the results across different prevalence rates. Bor *et al.* (2014), for example, investigated the much-debated epidemiologic question of when to start HIV patients on antiretroviral therapy by using the survival ratio to estimate the causal effect on mortality of early versus deferred treatment in a large South African cohort.

2.4 Study Designs
Differences in study designs are also evident across disciplines. Evaluation studies conducted by economists frequently use quasi-experimental methods, although randomised trials have now become the golden standard. Methods used in economic studies such as instrumental variable models (IV), local average treatment effect (LATE) models, and regression discontinuity design (RDD) are now starting to be considered by epidemiologists (see Bor, 2014 for a review of RDD modelling in epidemiology trials, and also a 2017 special issue of the *Journal of Clinical Epidemiology*). Encouragement designs are less common in epidemiology evaluations, with IV models scoring low on the scale of reliable methods (Martens *et al.*, 2006). Other quasi-experimental methods such as propensity score matching (PSM) are rarely used outside economics, whereas interrupted-time series designs and cohort studies, which are used by epidemiologists, are rarely used in evaluations conducted by economists.

For example, our reading of a systematic review of subsidised health insurance schemes in low and middle-income countries by Acharya *et al.* (2013) provides an illustration of how economists and epidemiologists employ different methods to address selection bias. Health insurance interventions are an interesting case because randomised trials in this area are very challenging to design. In some cases, programmes are implemented at the national level so that randomisation is impossible. In other cases, enrolment occurs on a voluntary basis so that the well-known problems of *moral hazard* and *adverse selection* emerge when estimating impact for the treated population. The review selected 19 papers for inclusion (mostly authored by economists and health economists and published in journals such as *Health Economics* and *Social Science and Medicine*), while it excluded 15 studies on methodological grounds (mostly

published by non-economists). The studies accepted for inclusion by the review used the following methods: matching methods (11 studies, two of which in combination with RDD and one in combination with IV), instrumental variables (4 studies), regression discontinuity design (2 studies), local average treatment effect (LATE) with encouragement design (1 study), triple difference (1 study).

As noted above, most of these methods for observational studies that are in common use in economics (possibly with the exception of the triple difference estimator, which is also rarely used in economics) are not widely used in epidemiology. Studies not included in the review on methodological grounds used the following methods to address self-selection of participants in the health insurance programmes: simulations, interviews to assess the scale of bias, and difference-in-difference regressions of enrolled and not enrolled populations with fixed effects.

When issues of selection bias arise within randomised trials, economists and epidemiologists tend to adopt different approaches. The workshop participants observed that in the context of these trials, as well as using an intention-to-treat analysis, epidemiologists may use either a per-protocol analysis (where those who adhere to the allocated intervention are compared to those in the control group who do not receive it), or an "as treated" analysis (where those who adhere to the intervention are compared to those who do not adhere it), in order to identify the impact of an intervention on the treated. Economists instead tend to address the problem of selection bias, even within randomised trials, with instrumental variables. Another practice adopted by economists is to create an instrument by design through an 'encouragement design' randomised trial, whereby participants are randomised to the offering of the intervention and then proceed to estimate a local-average treatment effect, which does not assess the impact of the intervention on the population, but only the impact on those that were encouraged to take-up the treatment. The encouragement design is quite common among economists; a classic example of encouragement design is the randomised offering of school vouchers in Colombia (Angrist, 2002), and more examples and a didactical non-technical introduction to encouragement design can be found in Gertler *et al.*, 2017).

2.5 Power Calculations and Statistical Significance vs Size of Impact in Relation to Intervention Context

Statistical power is the likelihood of finding an effect (i.e. detecting a difference between the treatment and comparison groups) when in fact one exists. Low power makes a false negative (i.e. an effect is erroneously not found when it truly exists) more likely. On the other hand, in relation to false positives, when the p-value, or calculated probability, approaches zero, this suggests a low chance of incorrectly rejecting the null hypothesis of no difference between the treatment and comparison groups (i.e. that a difference is found by chance, when there should be no difference).

Power calculations are used in the design of evaluation studies to show the smallest sample size required to detect a meaningful difference in outcomes between the treatment and comparison groups. Although conducting and reporting power calculations among economists has not been historically common practice (Carden,

2009), it is increasingly becoming the norm in the discipline.[5] This may be the result of the increasing popularity of experimental approaches, and citing a particularly influential paper as an example, Bandiera *et al.* (2014), evaluated the economic impact of the *Empowerment and Livelihood for Adolescents* (ELA) programme on improving the lives of adolescent girls through girls' clubs offering vocational and life skills. In the section discussing their study design, the authors explain that girls' clubs were randomly assigned across communities and that participation in clubs was voluntary. They also explain that a representative sample of communities of adolescent girls was randomly assigned to treatment and control at baseline and two years later, which generated a panel of 4,800 adolescent girls. The paper, however, does not report the use of power calculations to establish the sample size.

In contrast, it is probably fair to say that power calculations are still found more commonly in epidemiologic studies. For example, Wagman *et al.* (2015) was cited in the CEDIL workshop to illustrate the point. The study assessed whether the *Safe Homes and Respect for Everyone* (SHARE) project reduced physical and sexual IPV and HIV incidence among individuals enrolled in the Uganda Rakai Community Cohort Study (RCCS). The paper explains in detail how the study was powered on the basis of statistical calculations for the reductions of sexual and physical IPV prevalence rates expected from exposure to the programme.

Extrapolating from single studies, participants of CEDIL workshop noted the slightly imbalanced trend between economics and epidemiology in reporting power calculation referring to work by Ioannidis, Stanley and Doucouliagos (2017). The authors reviewed 159 economics' meta-analyses, summarising results from 6,700 studies and 64,000 estimates. They subdivided the studies and meta-analyses into the following categories: development economics (27%), microeconomics (23%), labour economics (18%), macroeconomics (17%), international economics (10%) and other (5%). They defined a study as adequately powered using the standard 80% cut-off, and classified as "powered" those studies for which the standard error was less than the ratio between effect size and 2.8, where 2.8 is the sum of 1.96 (for a significance level of 5%) and 0.84 (which is the standard normal of a 20/80% split in its cumulative distribution). The review showed that: the median proportion of adequately powered studies across the 159 meta-analyses was 10%, while the mean proportion was 22%; about 20% of meta-analysis were entirely composed of under-powered studies; the median power in the 159 meta-analyses was between 10-18%, while the mean power was between 30-33%; and the average effect size across all studies was 0.17 standard deviations.

As noted, epidemiologists have routinely performed power calculations, though interestingly, this has not prevented published studies from being underpowered. Turner, Bird and Higgins (2013) reviewed all meta-analyses from the *Cochrane Database of Systematic Reviews* with two or more studies. They found 14,886 meta-analyses including a total of 77,237 studies. The meta-analyses were divided across the following specialities: cancer (5%), cardiovascular (8%), central nervous system (8%), digestive

---

[5] McCloskey and Ziliak in a couple of papers (McCloskey and Ziliak, 1996; Ziliak and McCloskey, 2004) showed that only 4% of papers published in the *American Economic Review* mentioned conducting power calculations up to 1996, while only 8% reported the same up to 2004.

(10%), gynaecology (26%), infectious diseases (5%), mental health (13%), pathological conditions (3%), respiratory diseases (9%), urogenital (6%), and other (7%). Since the review looked at binary outcomes only, power was simply calculated as the ability of the studies to detect a relative risk reduction (RRR) of 10%, 20%, 30% and 50%, with a particular focus on the 30% RRR. The review showed that 70% of the studies are underpowered (at 50%) to detect this level of impact, while 83% of studies were found to be underpowered at the more standard 80% threshold. The overall median of all median power across meta-analyses is 13% for a 30% RRR, and while some specialities have higher median power (for example, cancer is 23% and pathological conditions is 17%), the distribution was quite even across the specialities. The reasons for so many (ex-post) underpowered studies were mostly attributed to researchers' over-enthusiasm in setting the likely effect sizes of interventions, problems with recruitment of participants/observations, and the inaccurate conduct of power calculations.

2.6 Publishing Protocols and Pre-Specified Outcomes

The publication of protocols and pre-registration of research has become the norm for randomised trials done by epidemiologists and is slowly becoming so for economists. This is also the case for systematic reviews in epidemiology (Booth *et al.*, 2011, 2013). When the evaluation method adopted for a study is not a randomised trial, then the practice of publishing a protocol is less frequent in both disciplines. This is likely to be because randomised trials have defined guidelines about pre-registration for publication of the final report in many journals, whereas the same is not true for other study designs.

Data Mining

Data mining is the practice of analysing and manipulating the data until it produces the desired signal and statistical significance level on the variable of interest. Correlated to the data mining problem is another practice called p-hacking, which results when researchers refine their analyses until they reach statistical significance to increase the probability of their statistically significant results being published. Many argue that current scientific publication practices create strong incentives to publish statistically significant (i.e. positive) results, and this in turn pushes researchers to selectively attempt to publish only statistically significant research findings. For example, Brodeur *et al.* (2016) look at all statistical tests reported by *American Economic Review*, *Journal of Political Economy* and *Quarterly Journal of Economics* between 2005 and 2011 and show evidence of 'inflation bias'. Many studies included in their review report coefficients with t-statistics clustered around and just below the standard 5% statistically significant level. They also show that p-hacking is more likely to be found in quasi-experimental studies than in randomised trials. The introduction of clear protocols and statistical analysis plans for randomised trials makes this practice less possible or, at least, more obvious if it is done. Also, the increased popularity of experimental approaches over quasi-experimental methods in economics is likely to reduce this source of publication bias in the future.

Instances of p-hacking are not unique to economics, although examples in epidemiology (especially for diagnostic and prognostics) are more difficult to find

(Albarqouni, López-López and Higgins, 2017). Another known practice in both fields is that of selective reporting, which appears to be evident in clinical epidemiology and results from pressure to report only the most promising results from trials (De Vries et al., 2018). A possible solution to the problems of p-hacking could be for funding organisations and ethics committees to require the publication of protocols and full data analysis from all the studies they fund or approve.

2.7 Reporting

The workshop discussion also raised additional points on the different reporting practices of economics and epidemiology: 1) in economics, for example, findings are typically published in a single journal article which is preceded by a series of working papers, while those in epidemiology are typically spread across one or more journal articles, each describing impact on different outcome measures (e.g. Nga *et al.*, 2009, 2011), and sometimes showing subgroup analyses in secondary publications (e.g. Azman *et al.*, 2013); and 2) the average length of journal articles reporting economic and epidemiology studies is also quite different. The former often reach 40 pages and do not follow a standardised reporting structure; whereas published papers in epidemiology rarely exceed 15 pages due to strict limits by journals (often with word limits of no more than 3,000 words) and follow a standardised reporting structure (Card and DellaVigna, 2013). There are specific reasons for these observed differences, and the section below overviews those that emerged from the workshop discussion.

Structure of Publications

The more standardised structure of epidemiological papers is making these publications easy to navigate because they have a clear template for the article, with dedicated sections to describe study design and methods, results and discussion. Epidemiology studies also tend to adhere to formal reporting guidelines, such as the Consolidated Standards of Reporting Trials (CONSORT) series (Consort:Transparent Reporting of Trials, 2015), which dictate best practice in reporting the results of randomised trials.

Peer reviewed economic papers, on the other hand, are structured very differently, with a large amount of information communicated in long documents running over approximately 40 pages, where the authors set out a conceptual framework, conduct extensive robustness analysis and study mechanisms of impact either through formal models or through heterogeneity analysis. In the workshop, this point was illustrated by discussing two papers in the WASH theme. The Duflo *et al.* (2015) evaluation of the economic impact of an integrated water and sanitation improvement programme in rural India is 40 pages long, featuring a six-page opening section that describes in detail the background of the intervention and the socio-economic context of the targeted area. The paper also includes an eleven-page technical annex which explains all the robustness checks carried out, the data cleaning process, and shows the costing exercise of the program. On the other hand, Dreibelbis' epidemiology study (Dreibelbis *et al.*, 2014), which looks at the impact of a WASH intervention on diarrhoea-related outcomes among younger siblings of school-going children, is a seven-page paper organised around a standard structure of four brief sections (i.e. objectives, methods,

15

results, and conclusions) with a clear visual depiction of the study design, and a brief section at the end outlining the limitations of the study.

Timeline to Final Publication

In economics, the final publication of an evaluation study as a paper in a peer-reviewed economic journal can take many years from the time that the analysis is finished. Findings, however, are publicly disseminated as they emerge through various platforms such as seminars and conferences, working papers, study reports and policy briefs. Working papers are considered to have some academic and policy value – especially given the ongoing practice to shift towards open science. Disclosure of early findings among peers and colleagues encourages critical review while generating early discussion on policy implications. However, this 'working papers' approach brings with it the problem that these papers might present findings that can change or need to be refined, so that outcomes reported and discussed in earlier versions may be absent from later versions or might have become irrelevant.

The timeline observed in economics between conducting a study, arriving at its early findings and final publication may not always fit the needs of timely policy making. However, it can also be the case that the policy discussions take place once the analysis is done and do not need to wait for the final journal paper. The International Initiative for Impact Evaluation (3ie) monitors policy influence stories. In the case of individual studies, rather than systematic reviews, the policy influence most usually comes from engagement between authors and intended users of the evidence, not from the publication of the paper in an academic journal.

The longer time taken to publish an impact evaluation in economics is often warranted by the depth of the analysis and by the complexity of the models it presents. This is not necessarily a waste of time, because the contribution to knowledge and policy of some papers would not necessarily be improved if a journal article was written and appeared rapidly in, say, three months. For example, the paper by Attanasio and co-authors on the evaluation of Mexico's Progresa CCT (Attanasio *et al.*, 2012) was submitted in 2001, but the final peer reviewed publication of the study appeared in 2012. This paper presents a complex structural model and took many years to write, although the findings had extensive policy impact in a pilot of a new design of the Progresa programme with the Mexico government well before its final publication date.

The timeline for publication appears to be much reduced in Epidemiology. Clinical trials, for example, take on average two years between completion and publication time, which in turn signals the pressure of timely dissemination of medical knowledge and clinical practice (Ross *et al.*, 2013; Welsh *et al.,* 2018).

Publication of Data and Code

In general, in economics, peer reviewers have the right to ask for data and code from a study whose manuscript they are reviewing and, normally, these would be made available to them as part of the peer reviewing process. Post-publication data sharing has become a formal requirement for peer-reviewed publication in economics, with the top economic journals (see *American Economic Association* guidelines) requiring this to

ensure that the analyses are replicable by other scholars. Authors are expected to document their data sources, models and estimation procedures as thoroughly as possible. This is to allow other researchers to replicate the data analysis presented in the published paper and to allow the building of a repository of data. It can also increase the number of citations of the article, enhancing the reputation of both the authors and the journal.

In Epidemiology, although similar practices are rarely found, there are initiatives seeking to change this (see *www.alltrials.net,* and Lo, 2015). Naudet *et al.* (2018), for example, discuss the effectiveness of data sharing requirements for randomised trials in two leading medical journals, *BMJ* and *PLOS Medicine*, suggesting the need to use specific incentives such as dedicated funding to make the data available along with all steps of the statistical analysis to overcome a general diffidence towards the reanalyses and replication of clinical trials. Interestingly, the authors wrote that "ensuring patient privacy and lack of explicit consent for sharing" (the two issues typically cited by authors as major barriers to sharing trial data (and generally accepted as valid exemptions), were not mentioned by any of the investigators they approached. Beyond providing incentives to authors, epidemiology journals adopting such practices could also be encouraged to make data and code sharing a requirement for reproducible research to enhance their reputation and credibility.

One reason why data sharing has been more acceptable in economics might be the aforementioned time lag between the first version of the evaluation findings (typically in a working paper format) and the final version in a peer reviewed journal. Another reason is that economists generally prepare a single paper that reports all findings. This practice would suggest that by the time the study is published, the authors will have exhausted their first-user rights to analyse the data, satisfying their priority to conduct original publishable work. Conversely, in epidemiology, the primary findings of an evaluation study are usually published in a stand-alone paper, which might then be followed by a series of other publications with additional results, giving the original researchers the opportunity to be the first to undertake these further analyses on the data and to publish the outputs.

2.8 Replicability of Impact Evaluations
Replication is an important component of impact evaluation work. It is standard practice that when a new result is found, other scientists around the world may attempt to independently replicate the results. The replication of results with new data is referred to as external replication, whereas internal replication is the attempt to reproduce the results of the original study using the same data.[6]

In economics, replication of impact evaluation studies has been the basis for greater research transparency and for making data available; in this respect, 3ie has gone a step

---

[6] Hamermesh, (2007) calls these scientific and pure replications respectively.

further in funding internal replications of high profile studies in international development.[7]

In epidemiology, the practice of replication aims at building bodies of evidence, even when this requires paying attention to conflicting findings.

CEDIL workshop participants therefore agreed on the following points: (1) it is necessary to test the same intervention more than once, since there are inevitable variations in context that may well change that degree of impact of any given intervention; and (2) researchers should ensure that different results do not emerge from using different data sets from the same study. If researchers use both different data and different models, there is little basis for comparing their results.

---

[7] http://www.3ieimpact.org/en/evaluation/impact-evaluation-replication-programme/

Evidence of Differences between Economists and Epidemiologists in Systematic Reviews

Systematic reviews are not conducted routinely in economics where the norm is to critically review a body of literature in great theoretical and empirical detail but rarely with the goal of informing policy making. It is also more common among economists to inform policy from the evidence from single studies. Epidemiologists on the contrary use systematic reviews as a means to generate a body of evidence from findings across different studies, even when this may include examining contradictory findings. Another observation made during the workshop was the general perception that in economics there is less interest in addressing why a study contradicts previous studies. For example, two systematic reviews of conditional cash transfers (CCT) interventions led by epidemiologists have been published in the *Cochrane Database of Systematic Reviews* that appear to have gone largely unnoticed among the development economists' community (Lagarde, Haines and Palmer, 2009, and Pega *et al.*, 2017).

One systematic review of CCT was published in the Campbell Collaboration library of systematic reviews by economists (Baird et al., 2013) and a critical review of studies of PROGRESA was published by the *Journal of Economic Literature* (*JEL*) (Parker and Todd, 2017). The first review published by Campbell is rather unusual for a team of economists, in that it follows systematic methods of search, appraisal and analysis, and it is aligned with the organisation's effort to promote systematic reviews across disciplines. The *JEL* review, on the other hand, is more typical of reviews in economics and in line with the reviews that economists normally read: it lacks a systematic scan of the literature, it does not critically asses the quality of the studies reviewed, and does not aim to summarise the evidence on the effectiveness of the interventions. In addition, it contains evidence relating to just one programme (PROGRESA) rather than the many similar CCTs around the world.

The sections below make a number of observations of the differences between individual studies and systematic reviews conducted by epidemiologists and economists for two popular types of interventions: conditional cash transfers and school feeding.

3.1 Conditional Cash Transfers (CCTs)

The reviews considered are: (1) Lagarde, Haines and Palmer (2009), (2) Pega *et al.* (2017), (3) Baird *et al.* (2013) and (4) Parker and Todd (2017). The reviews published in the *Cochrane Database of Systematic Reviews* exclusively selected randomised trials and controlled before-after studies, pointing to strenuous efforts by economists to reduce bias through the design and conduct of the study (1,2). Reviews by economists were rich in details on the methods used for the analyses of the data and included studies a wider

methodological variety of studies: matching, RDD, IV etc. (3,4) The three systematic reviews included quasi-experimental designs in the following proportions: 2/10 (20%) (1), 5/16 (31%) (2), and 40/75 (53%) (3). Quasi-experimental studies appeared more prominently in the economists' review while epidemiologists tended to focus on experimental studies. The *JEL* review included a study that would probably not be included in a review by epidemiologists: a study estimating the impact of *Oportunidades* on infant mortality using municipal level data and the percentage of rural households receiving the transfer.

3.2 School Feeding Programmes

We looked at differences between evaluations studies by epidemiologists and economists through the Cochrane Review by Kristjansson *et al.* (2007). The review investigated school feeding programmes for improving the physical and psychosocial health of disadvantaged students. The authors identified 30 potentially eligible studies, of which 12 were excluded for reasons other than the methods of the study (i.e. the intervention or the population were judged not to be relevant or there was a lack of relevant outcome data). The review also excluded a large number of studies on methodological grounds, such as cross-sectional comparisons of participants and non-participants and longitudinal studies with no control group, but the information in the review does not indicate if these had been conducted by economists or epidemiologists. Of the 18 included studies: 7 were classified as randomised trials, 9 as controlled before-after studies (CBA), and 2 as interrupted time series (ITS).

With one exception (Jacoby, Cueto and Pollitt, 1996), all included studies had been conducted by non-economists - either epidemiologists, educationalists or nutritionists. The review did not include details of the CBA studies and it is therefore not possible to comment on the methodologies of these quasi-experimental studies. Some epidemiologists include interrupted time series in their systematic reviews, but these studies do not rank highly in epidemiological hierarchies of evidence for studying the effects of interventions. Economists, however, typically do not consider this type of study as a valid method for causal inference, despite time series being a dominant topic in econometrics and the analysis of structural time breaks being a primer in beginners' econometrics courses.

An economic study by Jacoby (2002) appears to have been omitted from the review, while another economic study by Ahmed (2004) was identified but excluded. Jacoby conducted a natural experiment exploiting the fact that children's interviews were conducted during both school and non-school days (hence with and without a school meal) and correcting for selection bias using instrumental variables. The study by Ahmed replicated the study by Jacoby, generating the variation in school and out-of-school interviews by random assignment. The exclusion of the Ahmed study reminds us that natural experiments and instrumental variable approaches do not rank highly in the hierarchy of evidence employed by epidemiologists, coming after controlled before-after studies and interrupted time series, and they are rarely included in systematic reviews of evidence. It should also be noted that the economists' reliance on IV methods

is so widespread that it may have become excessive and the broad use of weak and unconvincing instruments has been criticised within the discipline (Deaton, 2010).

Econ-Epi Evaluation Glossary

Despite the existence of a wide range of evaluation glossaries in economics and many epidemiology glossaries, we have not been able to find a platform where evaluation terms used across these two disciplines are defined and explained. The existence of a common space to reconcile terms used by evaluation practitioners would facilitate communication in the understanding of each discipline's evaluation work. To help fill this gap, and to illustrate the value of such a platform, the annex to this paper includes an annotated evaluation glossary for epidemiology and economics. The next two sections explain how the glossary was collated and provide additional insights on the extent of overlapping terms. The unified glossary is intended to become an online tool that will be regularly updated and freely accessible to interested users.

Methods

We undertook a targeted search in Google and Google Scholar using the term 'impact evaluation glossary'. This identified two impact evaluation glossaries both published by economic organisations; 3ie and the Centre for Effective Global Action (CEGA) at the University of California (which was based on the World Bank's Impact Evaluation in Practice glossary; (Gertler *et al.*, 2017)). We did not identify any epidemiological glossary specifically for impact evaluations and selected the more general glossary from Cochrane as our starting point.

The complete list of terms from the three glossaries, the workshop and case studies were compiled into a single table in Microsoft Excel and definitions were added: 3ie and CEGA glossaries provide the economic definition, while Cochrane provides the epidemiological definition. Terms not present in the Cochrane glossary were cross checked in an epidemiological dictionary[8] and, if present, the definition was added to the table. Following this, terms were categorised as either:

- Unique to epidemiology;
- Unique to economics;
- Common to both;
- Common to both but different definition used by the two disciplines; or
- Different term used to describe the same concept.

The terms were classified into five themes to group similar terms together: study designs and approaches (e.g. cohort study, difference-in-difference), data and data collection methods (e.g. census data, survey instrument), trial procedures (e.g. run-in period, blinding), analysis (e.g. treatment of the treated, regression), and outcome indicators (e.g. effect size, relative risk).

---

[8] Porta M. A dictionary of epidemiology. Oxford University Press; 2008.

From the complete list of terms identified, we selected a subset that are presented in the annex to this paper, while the complete list is available as a supplementary appendix.

Findings

In total, we identified 207 unique terms from the three glossaries, expert workshop and case studies; 42.5% (88/207) from epidemiology sources alone, 22% (46/207) from economics sources along and 33% (69/207) from both disciplines. For two of the included terms (participant and sample), the definition differed slightly between the two disciplines. In one case, the two disciplines used a different term to describe the same concept ('dummy variable' and 'indicator variable').

Of the 13 terms identified through the workshop and case studies; 10 were from economics and 3 from epidemiology. The Annex includes a table with an excerpt of the proposed glossary.

Conclusions

―――

Epidemiologists and economists often raise similar evaluation questions in the public health domain of international development but use different analytical tools to measure impact of same policy relevant interventions. In some cases, the tools are the same but are used in different ways. In other instances, the same tools are employed but are described using a different terminology. These differences lead to misunderstandings between the two disciplines. Moreover, as the results of evaluations by epidemiologists and economists differ without being strictly comparable, readers and decision makers are confused and unable to interpret the results or to reconcile the differences between evaluations by economists and epidemiologists. We believe that there is scope for harmonising the design, methods, and reporting practices across the disciplines and hope that this paper will contribute to that process.

This paper reviewed some of these differences building on work conducted separately by mixed teams of epidemiologists (clinical, social, and public health specialists) and development economists. This work was further developed through a workshop organised by CEDIL in November 2017 to identify possible opportunities to overcome some of existing differences and develop the potential for cross-learning; as well as for advancing joint innovative methods of evaluation on global health issues in international development in the following five areas.

First, the approach adopted in best practice economics to use theory to test and predict behavioural change in response to an intervention is an element of enrichment in evaluation research. This approach could be combined usefully with the epidemiological practice of using a general theory to frame the evaluation of a specific intervention to predict the outcome under different scenarios and enhance the transferability of the findings.

Second, the generalisability and transferability of evaluation outcomes is recognised as a slippery territory where both disciplines could: 1) benefit from a better description of the settings of the intervention (epidemiology) to assess the external validity of the findings, and 2) extrapolate from the specifics of the single intervention (and the evidence of success or lack of success it has in one particular setting) to guide and refine the design of interventions in other settings.

Third, the practice of setting the findings of evaluations in the totality of the evidence (e.g. in relevant systematic reviews) is something that both economists and epidemiologists could substantially improve on. According to a study published on the Lancet (Clarke, Hopewell and Chalmers, 2010) the lack of integration of results of new trials into existing systematic reviews greatly diminishes both the scientific and ethical value of the trials. The authors suggest no improvement in this practice in periodic assessments of reports of randomised trials published between 1997 and 2009.

Fourth, the paper shows the rich menu of evaluation methods available in economics and epidemiology. It is possible and desirable to build opportunities for experimenting beyond common practices in each discipline, with plenty of scope for cross learning and adaption (e.g. regression discontinuity design, and the use of systematic reviews to advance learning in the field).

Finally, and linked to the point above, this paper creates the first working version of a new Econ-Epi evaluation glossary. Drawing from existing dictionaries of evaluation in each discipline, from the papers reviewed for this study, and from the discussion during the CEDIL workshop, we collated a glossary that we recommend being posted online and regularly updated on the CEDIL website to foster a better understanding of the evaluation terms used in each discipline.

Implications for CEDIL and DFID

This paper has implications for many organisation involved in impact evaluations and systematic reviews but, in accordance with the CEDIL remit we conclude with a focus on CEDIL and DFID. The review presented in this paper suggests several research approaches to be considered by CEDIL and DFID to improve how impact evaluations by economists and epidemiologists in global health interventions explore opportunities for interdisciplinary work. Below is a summary of these recommendations.

- Developing a common impact evaluation glossary. While writing this paper, we began the compilation of a glossary that would serve each discipline by translating unfamiliar terms that are commonly used by the other discipline. This effort could be taken further to include descriptions of impact evaluation methods and approaches, rather than just terms, that could become a useful reference for researchers of both disciplines (and, potentially, other disciplines as well).

- Developing guidelines for best impact evaluation practices. One lesson from the paper is that epidemiologists and economists have much to learn from each other. Guidelines for best practices, for example in reporting randomised trials (CONSORT), exist in epidemiology but nothing similar has been developed by economists. CEDIL could promote an attempt to produce guidelines on designing, conducting, analysis and reporting experimental and quasi-experimental studies which draws from knowledge accumulated by both disciplines. These guidelines could be tested by CEDIL when commissioning impact evaluation studies.

- Both disciplines could usefully learn from each other as to how theory may be used in studies as a basis for generalisability. Economists have still some way to go with respect to replication and the use of systematic reviews to build bodies of evidence.

- Encouraging collaborations between epidemiologists and economists. A collaboration between researchers from the two disciplines within teams for specific projects is one obvious way to promote mutual learning. CEDIL is in a unique position to require teams conducting impact evaluation to be multidisciplinary and to include both economists and epidemiologists. In addition, CEDIL could design impact evaluation courses with inputs from the two disciplines as part of its capacity building work.

- Dealing with power. While there are areas in which economists and epidemiologists could learn from each other, there are also areas where both could learn from other disciplines. One of these areas is in study design, the definition of sample size and statistical power, which is of fundamental importance in the estimation of impact of interventions. Economists normally neglect power calculations, but epidemiology studies, which routinely conduct

power calculations, are nevertheless often underpowered. CEDIL could promote research on understanding the causes and remedies to underpowered studies that are common to the two disciplines.

More generally, an important conclusion emerging from this paper is that evaluations conducted in epidemiology and economics on similar topics offer unique opportunities for advancing interdisciplinary science work of policy relevance. But *what makes interdisciplinary science work*? The discussion in this paper suggests a few clear avenues.

First, the importance of identifying an important question in international development (something that CEDIL could encourage under one of its Programme of Works), for example in the public health domain, where both disciplines share a common research interest.

Second, awarding research grants to collaborative projects from across disciplines (of which a team of economists and epidemiologist is one example) which are underpinned by discussion and agreement on the best combination of methods from their respective disciplines to evaluate what works and how (and the possible adaptability to other contexts).

This kind of interdisciplinary science work can enable more effective generalization and transferability of knowledge from evaluations carried out by multiple disciplines, not necessarily or exclusively by economists and epidemiologists.

References

Acharya, A. *et al.* (2013) 'The Impact of Health Insurance Schemes for the Informal Sector in Low-and Middle-Income Countries: A Systematic Review', *The World Bank Research Observer*, 28(2), pp. 236–266. doi: 10.1093/wbro/lks009.

Ahmed, A. U. (2004) 'Impact of feeding children in school: Evidence from Bangladesh', *International Food Policy Research Institute*, 2033, pp. 20006–1002. Available at: http://www.lcgbangladesh.org/FSN/reports/IFPRI Final Report_School Feeding in Bangladesh.pdf.

Albarqouni, L. N., López-López, J. A. and Higgins, J. P. T. (2017) 'Indirect evidence of reporting biases was found in a survey of medical research studies', *Journal of Clinical Epidemiology*. doi: 10.1016/j.jclinepi.2016.11.013.

Angrist, Joshua, Eric Bettinger, Erik Bloom, Elizabeth King, and Michael Kremer. (2002) 'Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment'. *American Economic Review*, 92 (5): 1535–58.

Ashraf, N., Bandiera, O. and Jack, B. K. (2014) 'No margin, no mission? A field experiment on incentives for public service delivery', *Journal of Public Economics*, 120, pp. 1–17. doi: 10.1016/j.jpubeco.2014.06.014.

Attanasio, O. *et al.* (2015) *Estimating the Production Function for Human Capital: Results from a Randomized Control Trial in Colombia, NBER Working Paper No. 20965*. doi: 10.3386/w20965.

Attanasio, O. P., Meghir, C. and Santiago, A. (2012) 'Education choices in Mexico: Using a structural model and a randomized experiment to evaluate PROGRESA', *Review of Economic Studies*, 79(1), pp. 37–66. doi: 10.1093/restud/rdr015.

Awasthi, S. *et al.* (2013) 'Population deworming every 6 months with albendazole in 1 million pre-school children in north India: DEVTA, a cluster-randomised trial', *The Lancet*. doi: 10.1016/S0140-6736(12)62126-6.

Azman, A. S. *et al.* (2013) 'Household transmission of influenza A and B in a school-based study of non-pharmaceutical interventions', *Epidemics*, 5(4), pp. 181–186. doi: 10.1016/j.epidem.2013.09.001.

Baird, S. *et al.* (2014) 'Conditional, unconditional and everything in between: a systematic review of the effects of cash transfer programmes on schooling outcomes', *Journal of Development Effectiveness*. doi: 10.1080/19439342.2014.890362.

Bandiera, O. *et al.* (2014) 'Women's Empowerment in Action: Evidence from a Randomized Control Trial in Africa'.

Banerjee, A. V., Chassang, S. and Snowberg, E. (2016) *Decision Theoretic Approaches to Experiment Design and External Validity, SSRN*. doi: 10.2139/ssrn.2770140.

Barratt, A. *et al.* (2004) 'Tips for learners of evidence-based medicine: 1. Relative risk

reduction, absolute risk reduction and number needed to treat', *CMAJ*, pp. 353–358. doi: 10.1503/cmaj.1021197.

Booth, A. *et al.* (2011) 'Establishing a minimum dataset for prospective registration of systematic reviews: An international consultation', *PLoS ONE*. doi: 10.1371/journal.pone.0027319.

Booth, A. *et al.* (2013) 'PROSPERO at one year: an evaluation of its utility.', *Systematic reviews*. doi: 10.1186/2046-4053-2-4.

Bor, J. *et al.* (2014) 'Regression discontinuity designs in epidemiology: Causal inference without randomized trials', *Epidemiology*. doi: 10.1097/EDE.0000000000000138.

Brodeur, A. *et al.* (2016) 'Star wars: The empirics strike back', *American Economic Journal: Applied Economics*, 8(1), pp. 1–32. doi: 10.1257/app.20150044.

Card, D. and DellaVigna, S. (2013) 'Nine Facts about Top Journals in Economics', *Journal of Economic Literature*, 51(1), pp. 144–161. doi: 10.1257/jel.51.1.144.

Card, D., DellaVigna, S. and Malmendier, U. (2011) 'The Role of Theory in Field Experiments', *Journal of Economic Perspectives*. doi: 10.1257/jep.25.3.39.

Carden, A. (2009) 'The Cult of Statistical Significance: How the Standard Error Costs us Jobs, Justice, and Lives - by Stephen T. Ziliak and Deirdre N. McCloskey', *Economic Affairs*. doi: 10.1111/j.1468-0270.2009.1884_3.x.

Cassidy, J. *et al.* (2011) *Textbook of Pediatric Rheumatology*, *Textbook of Pediatric Rheumatology*. doi: 10.1016/C2009-0-42830-7.

Clarke, M., Hopewell, S. and Chalmers, I. (2010) 'Clinical trials should begin and end with systematic reviews of relevant evidence: 12 years and waiting', *The Lancet*, pp. 20–21. doi: 10.1016/S0140-6736(10)61045-8.

Consort:Transparent Reporting of Trials (2015) *Consort - Welcome to the CONSORT Website*, *Medical Research Council*. Available at: http://www.consort-statement.org/.

Davey C, Hargreaves J, Hassan S, Cartwright N, Humphreys M, Masset E, Prost A, Gough D, Oliver S, Bonell C, (2018) 'Designing Evaluations to Provide Evidence to Inform Action in New Settings', CEDIL Inception Paper No 2: London

Davis, R. *et al.* (2015) 'Theories of behaviour and behaviour change across the social and behavioural sciences: a scoping review', *Health Psychology Review*, 9(3), pp. 323–344. doi: 10.1080/17437199.2014.941722.

Davies, R. (2018) 'Representing Theories of Change: A Technical Challenge with Evaluation Consequences', CEDIL Inception Paper 15: London

Dawood, F. S. *et al.* (2012) 'Estimated global mortality associated with the first 12 months of 2009 pandemic influenza A H1N1 virus circulation: A modelling study', *The Lancet Infectious Diseases*. doi: 10.1016/S1473-3099(12)70121-4.

Deaton, A. (2010) 'Instruments, Randomization, and Learning about Development', *Journal of Economic Literature*, 48(2), pp. 424–455. doi: 10.1257/jel.48.2.424.

De Vries, Y., Roest, A., De Jonge, P., Cuijpers, P., Munafò, M., & Bastiaansen, J. (2018). 'The cumulative effect of reporting and citation biases on the apparent efficacy of

treatments: The case of depression'. *Psychological Medicine, 48*(15), 2453-2455. doi:10.1017/S0033291718001873

Dreibelbis, R. *et al.* (2014) 'The impact of school water, sanitation, and hygiene interventions on the health of younger siblings of pupils: A cluster-randomized trial in Kenya', *American Journal of Public Health*, 104(1). doi: 10.2105/AJPH.2013.301412.

Duflo, E. *et al.* (2015) *Toilets Can Work: Short and Medium Run Health Impacts of Addressing Complementarities and Externalities in Water and Sanitation*. doi: 10.3386/w21521.

Duflo, E., Hanna, R. and Ryan, S. P. (2012) 'Incentives work: Getting teachers to come to school', *American Economic Review*, pp. 1241–1278. doi: 10.1257/aer.102.4.1241.

Dupas, P. (2009) 'What matters (and what does not) in households' decision to invest in malaria prevention?', in *American Economic Review*. doi: 10.1257/aer.99.2.224.

Eccles, M. *et al.* (2005) 'Changing the behavior of healthcare professionals: The use of theory in promoting the uptake of research findings', *Journal of Clinical Epidemiology*. doi: 10.1016/j.jclinepi.2004.09.002.

Feinstein, A. R. (1995) 'Clinical epidemiology: The architecture of clinical research', *Statistics in Medicine*, pp. 1263–1263. doi: 10.1002/sim.4780141110.

Gertler, P. J. *et al.* (2017) *Impact Evaluation in Practice*, *The World Bank Publications*. doi: 10.1596/978-0-8213-8541-8.

Hamermesh, D. S. (2007) 'Viewpoint: Replication in economics', *Canadian Journal of Economics*, pp. 715–733. doi: 10.1111/j.1365-2966.2007.00428.x.

Hoffmann, T. C. *et al.* (2014) 'Better reporting of interventions: Template for intervention description and replication (TIDieR) checklist and guide', *BMJ (Online)*. doi: 10.1136/bmj.g1687.

Ioannidis, J. P. A., Stanley, T. D. and Doucouliagos, H. (2017) 'The Power of Bias in Economics Research', *Economic Journal*, 127(605), pp. F236–F265. doi: 10.1111/ecoj.12461.

Jacoby, E., Cueto, S. and Pollitt, E. (1996) 'Benefits of a school breakfast programme among Andean children in Huaraz, Peru', *Food and Nutrition Bulletin*, 17(1), pp. 54–64.

Jacoby, H. G. (2002) 'Is there an intrahousehold "flypaper effect"? Evidence from a school feeding programme', *Economic Journal*, 112(476), pp. 196–221. doi: 10.1111/1468-0297.0j679.

Krezanoski, P. J., Comfort, A. B. and Hamer, D. H. (2010) 'Effect of incentives on insecticide-treated bed net use in sub-Saharan Africa: a cluster randomized trial in Madagascar', *Malaria Journal*, 9. Available at: http://www.malariajournal.com/content/9/1/186.

Kristjansson, E. A. *et al.* (2007) 'School feeding for improving the physical and psychosocial health of disadvantaged elementary school children.', *The Cochrane database of systematic reviews*, 24(1), pp. 1–139. doi: 10.1002/14651858.CD004676.pub2.

Lagarde, M., Haines, A. and Palmer, N. (2009) 'The impact of conditional cash transfers on health outcomes and use of health services in low and middle income countries',

Cochrane Database of Systematic Reviews. doi: 10.1002/14651858.CD008137.

Lo, B. (2015) 'Sharing clinical trial data: Maximizing benefits, minimizing risk', *JAMA - Journal of the American Medical Association*, 313(8), pp. 793–794. doi: 10.1001/jama.2015.292.

Martens, E. P. *et al.* (2006) 'Instrumental variables: Application and limitations', *Epidemiology*, 17(3), pp. 260–267. doi: 10.1097/01.ede.0000215160.88317.cb.

McCloskey, D. N. and Ziliak, S. T. (1996) 'The Standard Error of Regressions', *Journal of Economic Literature*, pp. 97–114. doi: 10.1257/jep.22.4.199.

Miguel, E. and Kremer, M. (2004) 'Worms: Identifying Impacts on Education and Health in the Presence of Treatment Externalities', *Econometrica*, 72(1), pp. 159–217. doi: 10.1111/j.1468-0262.2004.00481.x.

Morris, S. S. *et al.* (2004) 'Conditional Cash Transfers Are Associated with a Small Reduction in the Rate of Weight Gain of Preschool Children in Northeast Brazil', *The Journal of Nutrition*. doi: 10.1093/jn/134.9.2336.

Naudet, F. *et al.* (2018) 'Data sharing and reanalysis of randomized controlled trials in leading biomedical journals with a full data sharing policy: survey of studies published in The BMJ and PLOS Medicine', *BMJ*, 360, p. k400. doi: 10.1136/bmj.k400.

Nga, T. T. *et al.* (2009) 'Multi-Micronutrient-Fortified Biscuits Decreased Prevalence of Anemia and Improved Micronutrient Status and Effectiveness of Deworming in Rural Vietnamese School Children', *Journal of Nutrition*, 139(5), pp. 1013–1021. doi: 10.3945/jn.108.099754.

Nga, T. T. *et al.* (2011) 'Decreased parasite load and improved cognitive outcomes caused by deworming and consumption of multi-micronutrient fortified biscuits in rural Vietnamese schoolchildren', *American Journal of Tropical Medicine and Hygiene*, 85(2), pp. 333–340. doi: 10.4269/ajtmh.2011.10-0651.

Parker, S. W. and Todd, P. E. (2017) 'Conditional Cash Transfers: The Case of Progresa/Oportunidades', *Journal of Economic Literature*, 55(3), pp. 866–915. doi: 10.1257/jel.20151233.

Pega, F. *et al.* (2017) 'Unconditional cash transfers for reducing poverty and vulnerabilities: Effect on use of health services and health outcomes in low- and middle-income countries', *Cochrane Database of Systematic Reviews*. doi: 10.1002/14651858.CD011135.pub2.

Piketty, T. and Deaton, A. (2014) 'Instruments of development: Randomization in the tropics, and the search for the elusive keys to economic development', *The Quarterly Journal of Economics*, 4(3), pp. 1071–1131. doi: 10.1093/qje/qjr020.

PROGRESA Author, E. *et al.* (2012) 'Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to Evaluate PROGRESA', *Source: The Review of Economic Studies Review of Economic Studies*, 79(79), pp. 37–66. doi: 10.1093/restud/rdr015.

Ross, J. S., Mocanu, M., Lampropulos J.F., Tse, T., Krumholz, H. M. . (2013) 'Time to

publication among completed clinical trials', *JAMA Internal Medicine*. doi: 10.1001/jamainternmed.2013.136.

Rothwell, P. M. (2005) 'External validity of randomised controlled trials: "to whom do the results of this trial apply?"', *Lancet*. doi: 10.1016/S0140-6736(04)17670-8.

Sackett, D. L. (2002) 'Clinical epidemiology: What, who, and whither', *Journal of Clinical Epidemiology*, pp. 1161–1166. doi: 10.1016/S0895-4356(02)00521-8.

Spiegelman, D. (2016) 'Evaluating public health interventions: 1. examples, definitions, and a personal note', *American Journal of Public Health*, 106(1), pp. 70–73. doi: 10.2105/AJPH.2015.302923.

Turner, R. M., Bird, S. M. and Higgins, J. P. T. (2013) 'The Impact of Study Size on Meta-analyses: Examination of Underpowered Studies in Cochrane Reviews', *PLoS ONE*, 8(3). doi: 10.1371/journal.pone.0059202.

Wagman, J. A. *et al.* (2015) 'Effectiveness of an integrated intimate partner violence and HIV prevention intervention in Rakai, Uganda: Analysis of an intervention in an existing cluster randomised cohort', *The Lancet Global Health*. doi: 10.1016/S2214-109X(14)70344-4.

Welsh J, Lu Y, Dhruva SS, Tse, T., Krumholz, H. M. (2018) 'Age of Data at the Time of Publication of Contemporary Clinical Trials'. *JAMA Netw Open.* 1(4):e181065. doi:10.1001/jamanetworkopen.2018.1065

Welch, V. Ghogomu, E., Hossain, A., Awasthi, S., Bhutta, Z., Cumberbatch, C., Fletcher, R., McGowan, J., Krishnaratne, S., Kristjansson, E., Sohani, S., Suresh, S., Tugwell, P., White, H., and Wells, G. (2013) 'Deworming and Adjuvant Interventions for Improving the Developmental Health and Well-being of Children in Low- and Middle-income Countries: A Systematic Review and Network Meta-analysis', *Campbell Systematic Reviews*. doi: 10.4073/csr.2016.7.

White, H. (2010) 'A contribution to current debates in impact evaluation', *Evaluation*. doi: 10.1177/1356389010361562.

Ziliak, S. T. and McCloskey, D. N. (2004) 'Size matters: The standard error of regressions in the American Economic Review', *Journal of Socio-Economics*, 33(5), pp. 527–546. doi: 10.1016/j.socec.2004.09.024.

3ie impact evaluation glossary 2012 New Delhi, India: International Initiative for Impact Evaluation
Available at:
http://www.3ieimpact.org/media/filer_public/2012/07/11/impact_evaluation_glossary_-_july_2012_3.pdf (accessed 23 October 2017)

The Cochrane Reviewers' Handbook Glossary Cochrane Collaboration Report No.: Version 4.1.5.
Available at: http://www.epidemiologia.anm.edu.ar/cochrane/pdf/glossary.pdf
(accessed 21 Nov 2017)

Table A1: Economic and Epidemiology Papers Consulted

	CASE STUDIES									
	BEDNETS		WASH		GENDER: IPV/Sexual behavior		ART/HIV		CCT	
	Economics	Epidemiology	Economics	Epidemiology	Economics	Epidemiology	Economics	Epidemiology	Economics	Epidemiology
Title and Pub. Source	Dupas, P. (2009) What **Matters (and What Does Not) in Households' Decision to Invest in Malaria Prevention?** American Economic Review 99(2):224-30	Krezanoski PJ, Comfort AB, Hamer DH. Malar J. (2010) **Effect of incentives on insecticide-treated bed net use in sub-Saharan Africa: a cluster randomized trial in Madagascar.** Malaria Journal 9:186	Duflo E., et al. **Toilets Can Work: Short and Medium Run Health Impacts of Addressing Complementarities and Externalities in Water and Sanitation.** NBER Working Paper No. 21521, 2015	Dreibelbis, R. et al. (2014) **The Impact of School Water, Sanitation, and Hygiene Interventions on the Health of Younger Siblings of Pupils: a Cluster-Randomized Trial in Kenya.** American Journal of Public Health; 104 (1). doi: 10.2105/AJPH.2013.301412.	Bandiera, O. et al. **Women's Empowerment in Action: evidence from a randomized control trial in Africa.** EnGender Impact: the World Bank's Gender Impact Evaluation Database. Washington, DC World Bank Group. 2014	Wagman G. et al. (2015) **Effectiveness of an integrated intimate partner violence and HIV prevention intervention in Rakai, Uganda: analysis of an intervention in an existing cluster randomised cohort** Lancet Global Health (14) 70344-4. doi: 10.1016/S2214-109X(14)70344-4	Baranov, V., Bennet, D., Kohler, H-P. (2015) **The indirect impact of antiretroviral therapy: Mortality risk, mental health, and HIV-negative labor supply** Journal of Health Economics 44; 195–211	Bor, J., Moscoe, E., Mutevedzi, P., Newell, M. L., and Bärnighausen, T. (2014) **Regression Discontinuity Designs in Epidemiology Causal Inference Without Randomized Trials** Epidemiology 25: 729–737.	Attanasio, O., Meghir, C., and Santiago A. (2012) **Education Choices in Mexico: Using a Structural Model and a Randomized Experiment to evaluate Progresa** The Review of Economic Studies, Vol. 79, No. 1, pp. 37-66	Morris, SS, Olinto, P, Flores R, Nilson, AFN, and Figueiró AC (2004) **Conditional Cash Transfers Are Associated with a Small Reduction in the Rate of Weight Gain of Preschool Children in Northeast Brazil** The Journal of Nutrition, 134(9): 2336–2341
Research Question	What is the effect on the take-up of a preventative health product (LL-ITNs) of two interventions: varying the framing of the perceived benefits; and having people verbally commit to purchase the product.	What is the net the effect of incentives on the use of ITNs?	What is the impact of an integrated water and sanitation improvement programme in rural India on the number of treated diarrhoea episodes (impact of providing universal access within a village to hygienic latrines and in-home piped water)	What is the impact of WASH interventions on diarrhoea-related outcomes among younger siblings of school-going children?	What is the impact on adolescent girls' ability to make informed choices about sex, reproduction and marriage of a programme delivering two interventions: (1) 'hard' vocational skills to enable adolescent girls to start small-scale income generating activities; (i2) 'soft' life skills to build knowledge enabling girls to make informed choices about sex, reproduction and marriage.	Does the provision of a combination of IPV prevention and HIV services reduce IPV and HIV incidence in individuals enrolled in the Rakai Community Cohort Study (RCCS), Rakai, Uganda?	Does ART availability increase work time for HIV-negative people without caretaker obligations, who do not directly benefit from the medicine?	What is the causal effect of early versus deferred treatment eligibility on mortality for patients whose first CD4 count was just below the 200 cells/μL CD4 count threshold in a large South African cohort (2007–2011)?	What is the impact of the PROGRESA programme on school participation using experimental data to estimate a structural economic model of education choices?	What is the impact of the Brazilian federal government programme, Bolsa Alimentacao on the growth of children from beneficiary households relative to that of similar children accidentally prevented from receiving the benefit?

33

Annex 1: Econ–Epi Unified Evaluation Glossary

Classification	Econ-Epi	Term	Economic definition	Epidemiological definition
ANALYSIS	**Both – different terminology**	**Average treatment effect**	**The average treatment effect calculated across the whole treatment group, regardless of whether they actually participated in the intervention or not. Compare to treatment of the treated.**	**See intention to treat estimate**
	Both	Confounding factors/variables (a confounder) [econ: see also IV]	Confounding is a situation in which a measure of the effect of an intervention or exposure is distorted because of the association of exposure with other factor(s) ('confounders') that influence the outcome under investigation.	
	Both	Instrumental variable [epi: See also confounding bias, residual confounding]	An instrumental variable is a variable that helps identify the causal impact of a programme when participation in the programme is partly determined by the potential beneficiaries. A variable must have two characteristics to qualify as a good instrumental variable: (1) it must be correlated with programme participation, and (2) it must not be correlated with outcomes Y (apart from through programme participation) or with unobserved variables.	
	Both – different terminology	**Intention to treat estimate**	**See average treatment effect**	**An intention-to-treat analysis is one in which all the participants in a trial are analysed according to the intervention to which they were allocated, whether they received it or not. Intention-to-treat analyses are favoured in assessments of effectiveness as they mirror the noncompliance and treatment changes that are likely to occur when the intervention is used in practice, and avoid the risk of attrition bias that would arise when participants are excluded from the analysis.**
	Epi	Sensitivity analysis		An analysis used to determine how sensitive the results of a study or systematic review are to changes in how it was done. Sensitivity analyses are used to assess how robust the results are to decisions about which there is uncertainty or assumptions about the data and the methods that were used.
	Econ	Unobservable / Unobservable variables	Characteristics which cannot be observed or measured. The presence of unobservables can cause selection bias in quasi-experimental designs, if these unobservables are correlated with both participation in the programme and the outcome(s) of interest.	

Classification	Econ-Epi	Term	Economic definition	Epidemiological definition
DESIGN	Epi	Case-control study (synonyms: case referent study, retrospective study)		A study that starts with identification of people with the disease or outcome of interest (cases) and a suitable control group without the disease or outcome. The relationship of an attribute (intervention, exposure or risk factor) to the outcome of interest is examined by comparing the frequency or level of the attribute in the cases and controls. For example, to determine whether thalidomide caused birth defects, a group of children with birth defects (cases) could be compared to a group of children without birth defects (controls). The groups would be compared with respect to the proportion exposed to thalidomide through their mothers taking the tablets. Case-control studies are sometimes described as being retrospective, as they are performed by looking back in time.
	Epi	Cohort study (synonyms: follow-up, incidence, longitudinal, prospective study)		An observational study in which a defined group of people (the cohort) is followed over time. The outcomes of people in subsets of this cohort are compared to examine, for example, people who were exposed or not exposed (or exposed at different levels) to a particular intervention or other factor of interest. A cohort can be assembled in the present and followed into the future (i.e. a prospective study or a "concurrent cohort study"), or the cohort could be identified from past records and followed from the time of those records to the present (i.e. a retrospective study or a "historical cohort study"). Because random allocation is not used, matching or statistical adjustment at the analysis stage must be used to minimise the influence of factors other than the intervention or factor of interest.
	Econ	Difference-in-difference (syn: double difference)	Double difference: The difference in the change in the outcome observed in the treatment group compared to the change observed in the comparison group; or, equivalently, the change in the difference in the outcome between treatment and comparison. Double differencing removes selection bias resulting from time-invariant unobservables. Also called Difference-in-difference. See also single difference and triple difference.	

Econ-Epi	Term	Economic definition	Epidemiological definition
Econ	Ex ante evaluation design	An impact evaluation design prepared before the intervention takes place. Ex ante designs are stronger than ex post evaluation designs because of the possibility of considering random assignment, and the collection of baseline data from both treatment and comparison groups. Also called prospective evaluation.	
Econ	Ex post evaluation design	An impact evaluation design prepared after the intervention has started (or, even, completed). Unless there was random assignment, then a quasi-experimental design has to be used.	
Both	Interrupted time series analysis	A non-experimental evaluation method in which data are collected at multiple instances over time before and after an intervention is introduced to detect whether the intervention has an effect significantly greater than the underlying secular trend.	
Both	Matching	A method to create comparison groups in which comparison groups or individuals are matched to those in the treatment group based on characteristics felt to be relevant to the outcome(s) of the intervention.	
Econ	Regression discontinuity design (RDD)	An impact evaluation design in which the treatment and comparison groups are identified as being those just either side of some threshold value of a variable. This variable may be a score or observed characteristic (e.g. age or land holding) used by programme staff in determining the eligible population, or it may be a variable found to distinguish participants from non-participants through data analysis. RDD is an example of a quasi-experimental design.	
Epi	Stepped wedge trial		Stepped-wedge cluster-randomised trials (SWTs) are used in a wide range of areas of public health, as well as other areas of public policy such as education and international development. SWTs can be thought of as a modified crossover design because each cluster is in both arms at different times. All clusters start in the control arm and the intervention is introduced by random allocation and at regular intervals either to one cluster at a time or to small groups of clusters, until all clusters are eventually receiving the intervention.

Classification	Econ-Epi	Term	Economic definition	Epidemiological definition
OUTCOME INDICATOR	Epi	Relative Risk (RR) (synonym: risk ratio)		The ratio of risk in the intervention group to the risk in the control group. The risk (proportion, probability or rate) is the ratio of people with an event in a group to the total in the group. A relative risk of 1.00 indicates no difference between comparison groups. For undesirable outcomes, an RR that is less than one indicates that the intervention was effective in reducing the risk of that outcome.
	Epi	Risk difference (RD) (synonym: absolute risk reduction)		The absolute difference in the risk between two comparison groups. A risk difference of 0.00 indicates no difference between comparison groups. A RD that is less than zero indicates that the intervention reduced the risk of that outcome.
	Epi	Surrogate endpoints (synonym: intermediary outcomes; surrogate outcomes)		Outcome measures that are not of direct practical importance but are believed to reflect outcomes that are important; for example, blood pressure is not directly important to patients, but it is often used as an outcome in clinical trials because it is a risk factor for stroke and heart attacks. Surrogate endpoints are often physiological or biochemical markers that can be relatively quickly and easily measured, and that are taken as being predictive of important clinical outcomes. They are often used when observation of clinical outcomes requires long follow-up.

Classification	Econ-Epi	Term	Economic definition	Epidemiological definition
TRIAL PROCEDURES	Both	Blinding (epi synonym: masking)	Keeping secret group assignment (e.g. to treatment or control) from the study participants or investigators. Blinding is used to protect against the possibility that knowledge of assignment may affect patient response to treatment, provider behaviours (performance bias) or outcome assessment (detection bias). Blinding is not always practical (e.g. when comparing surgery to drug treatment). The importance of blinding depends on how objective the outcome measure is; blinding is more important for less objective outcome measures such as pain or quality of life. See also single blind, double blind and triple blind.	
	Econ	Endline	The situation at the end of an intervention, in which progress can be assessed or comparisons made with baseline data. Endline data are collected at the end of a programme or policy is implemented to assess the "after" state. Source: CEGA	
	Epi	Protocol		The plan or set of steps to be followed in a study. A protocol for a systematic review should describe the rationale for the review; the objectives; and the methods that will be used to locate, select and critically appraise studies, and to collect and analyse data from the included studies.
	Epi	Run-in period		A period before a participant joins a trial fully when no treatment is given. The data from this stage of a trial are only occasionally of value but the run-in period can serve a valuable role in screening out ineligible or non-compliant participants, in ensuring that participants are in a stable condition, and in providing baseline observations. A run-in period is sometimes called a washout period if treatments that participants were using before entering the trial are discontinued.
	Both - different definitions	**Sample**	**A subset of the population being studied. The sample is drawn randomly from the sampling frame. In a simple random sample all elements in the frame have an equal probability of being selected, but usually more complex sampling designs are used, requiring the use of sample weights in analysis.**	**A selected subset of a population. A sample may be random or non-random and may be representative or non-representative. Types of sampling: area sampling; cluster sample; grab sample; probability (random) sample; simple random sample; stratified random sample; systematic sample.**

Contact Us

Centre of Excellence for Development Impact and Learning (CEDIL)

London International Development Centre

36 Gordon Square

London WC1H 0PD

United Kingdom

www.cedilprogramme.org

@CEDILProgramme