DRAFT: 2009 06 12

SUPPORT Tools for Evidence-informed policymaking in health

15. Monitoring and evaluating policies and programmes

Atle Fretheim[1]
Andrew D Oxman[2]
John N Lavis[3]
Simon Lewin[4]

1. Norwegian Knowledge Centre for the Health Services, P.O. Box 7004, St. Olavs plass, N-0130 Oslo, Norway. Email: atle.fretheim@nokc.no
2. Norwegian Knowledge Centre for the Health Services, P.O. Box 7004, St. Olavs plass, N-0130 Oslo, Norway. Email: oxman@online.no
3. Centre for Health Economics and Policy Analysis, Department of Clinical Epidemiology and Biostatistics, and Department of Political Science, McMaster University, 1200 Main St. West, HSC-2D3, Hamilton, ON, Canada, L8N 3Z5. Email: lavisj@mcmaster.ca
4. Norwegian Knowledge Centre for the Health Services, P.O. Box 7004, St. Olavs plass, N-0130 Oslo, Norway. Email: simon.lewin@nokc.no

Corresponding author:
Dr Atle Fretheim
Norwegian Knowledge Centre for the Health Services
P.O. Box 7004, St. Olavs plass
N-0130 Oslo, Norway

Email: atle.fretheim@nokc.no

Abstract

Background: This is article number 15 in a series of 21 articles on tools for evidence-informed health policymaking. *Monitoring* is the term commonly used to describe the process of systematically collecting data that can inform policymakers, managers and other stakeholders as to whether a new policy or programme is progressing in accordance with their expectations. Data that are used for monitoring purposes are used as indicators to judge, for example, if objectives are being achieved, or if allocated funds are being spent appropriately. The term *evaluation* is sometimes used inter-changeably with *monitoring*. However, the term *impact evaluation* usually implies that there is a specific attempt to try to determine whether the observed changes in outcomes can be attributed to a particular policy or programme.

Objectives: To provide support to policymakers in their decision making related to monitoring and evaluation activities.

Key messages:

- The following questions can be used to guide the monitoring and evaluation of a policy or programme:
  1. Is monitoring necessary?
  2. What should be measured?
  3. How will the findings be utilised?
  4. Should an impact evaluation be conducted?
  5. How should the impact evaluation be done?
- Finding the right balance between the feasibility and quality of a monitoring and evaluation system is often a challenge when assessing a new policy or programme. Such a system may be highly resource demanding and may (or may not) as a consequence be worth implementing. Sometimes existing information sources may suffice
- The number of indicators used and collated for monitoring purposes should be limited in order to avoid putting too much strain on health services
- Monitoring is not worthwhile if the data collected are not used. But data are particularly useful if corrective action is taken when a gap between expected and actual performance is identified
- There is often significant uncertainty about whether a new programme is effective or not or, more significantly, whether it causes more harm than good. These issues are important to clarify both for policymakers implementing new programmes, and for those who could benefit from knowing about effective programmes in health policymaking. *Impact evaluations* are needed to address these issues because routine monitoring usually does not provide the data necessary for such assessments
- Conducting an impact evaluation can be costly. Whether such a study represents good value for money can be ascertained by comparing the consequences of undertaking an evaluation with the consequences of not doing so
- Attributing an observed change to a programme or policy requires a comparison between individuals or groups that are exposed to it, and others who are not exposed to it. The optimal approach to such effect evaluations is to conduct a randomised controlled trial. Interrupted time series analyses and controlled before-after studies are alternatives, but these are less reliable methods for estimating the effects of an intervention

Background

This is article number 15 in a series of 21 articles on tools for evidence-informed health policymaking. It is also the third of three articles in the series about planning implementation, scaling up, and monitoring and evaluation strategies. The purpose of this article is to suggest how to monitor the implementation of policies and programmes and evaluate their impacts.

Policymakers, managers and other stakeholders will often need to know whether the implementation of a new policy or programme has been done in accordance with their expectations. Is the programme rollout progressing as planned? Are the objectives being achieved, and are the allocated funds being spent appropriately? *Monitoring* is the term commonly used to describe the process of systematically collecting data to provide answers to such questions [1]. The term *performance monitoring* is often used when the main focus of an evaluation is on comparing "how well a project, program, or policy is being implemented against expected results" [1].

Data are frequently used as *indicators* as part of the monitoring process, i.e. a "quantitative or qualitative factor or variable that provides a simple and reliable means to measure achievement, to reflect the changes connected to an intervention, or to help assess the performance" [1].

The term *evaluation* is sometimes used interchangeably with *monitoring*, but the former usually suggests a stronger focus on the achievement of results. The term *impact evaluation* is frequently used when an attempt is made to evaluate whether observed changes in performance can be attributed to a particular policy or programme.

Questions to consider

1. Is monitoring necessary?
2. What should be measured?
3. How will the findings be utilised?
4. Should an impact evaluation be conducted?
5. How should the impact evaluation be done?

1. Is monitoring necessary?

The importance of monitoring depends on the perceived need among relevant stakeholders to know more about what is happening 'on the ground'.

Determining whether a system for monitoring a policy or programme should be established may depend on several factors, including:

- Whether a monitoring system is already in place that includes the required indicators, or if a new evaluation system is required
- The likely costs of establishing the system required. For example, could a few new indicators be added to existing data collection procedures already in place, or would additional large-scale household surveys be needed?

- Whether the findings are likely to be useful. What actions should be taken if monitoring reveals that things are not going as planned?

Illustrative examples of monitoring systems that have been implemented are given in Box 1 [2, 3].

2. What should be measured?

A number of factors need to be considered when selecting the data indicator(s) used for monitoring [4, 5]:

- Validity: the extent to which the indicator accurately measures what it purports to measure
- Reproducibility: the extent to which the indicator would be the same if the method by which it was produced was repeated
- Acceptability: the extent to which the indicator is acceptable to those being assessed and those undertaking the assessment
- Feasibility: the extent to which valid, reliable and consistent data are available for collection
- Reliability: the extent to which there is minimal measurement error or the extent to which findings are reproducible should they be collected again by another organisation
- Sensitivity to change: the extent to which the indicator has the ability to detect changes in the unit of measurement
- Predictive validity: the extent to which the indicator has the ability to accurately predict relevant outcomes

A trade-off is often apparent between wanting to use desired and optimal indicators on one hand, and having to use those indicators which can be measured using existing data on the other. Good reasons may exist not to select more indicators than are absolutely essential. These reasons include the need to: limit the burden of data collection within a health system, avoid the collection of data that are not utilised, and focus on collecting data of higher quality, even if this means collecting less data [6].

The appropriate data can be collected routinely within a health service, for example, through surveys conducted at regular intervals, or through interviews. Consideration should also be given to the level of motivation of those expected to collect data. In many instances, health personnel will need to integrate data collection into a busy daily schedule. Therefore if the information being collected has little or no local obvious value to them, motivation levels for undertaking such tasks might be low. Similarly, if incentives or penalties are associated with the findings from the monitoring process (e.g. pay for performance schemes), the risk of data manipulation or system gaming may also need to be considered.

See Box 2 for illustrative examples of selected indicators [2, 3].

3. How will the findings be utilised?

Monitoring is not worthwhile if data remain unused. But data are particularly useful if corrective action is undertaken when a gap between expected and actual results is identified. Such findings may result in expectations being reconsidered. These may include assessments, for example, of whether the initial plans were too ambitious, and whether a new policy has failed to work as effectively as expected.

See Box 3 for illustrative examples of how findings can be utilised [2, 3].

4. Should an impact evaluation be conducted?

One of the limitations of monitoring activities, as described above, is the fact that they do not necessarily indicate whether a policy or programme has led to improved performance. This is because monitoring indicators will almost always be influenced by factors other than those related to particular interventions. This makes it extremely difficult to determine which factors caused observed changes. If monitoring reveals that performance is improving, this does not necessarily mean that the intervention is the (only) causal factor – it is conceivable that things may have improved anyway without the intervention (see Fig 1 [7]).

The establishment of a causal relationship between a programme or policy and changes in outcomes is at the core of what impact evaluation is about. As the World Bank has stated: "The central impact evaluation question is what would have happened to those receiving the intervention if they had not in fact received the program" [8].

While there may be strong reasons to expect positive results based on solid documentation from, for example, previous evaluations, very often such evidence is lacking, or the evidence available may not be applicable to the current setting. Thus, there is a real risk that a new programme may be ineffective or, even worse, cause more harm than good. This issue is important for policymakers to clarify when implementing new programmes. It is also important because of the benefit that such knowledge about effective programmes could bring to health policymaking.

Illustrative examples of conducted impact evaluations are provided in Box 4 [9-11].

Conducting impact evaluations can be costly. Whether such studies represent good value for money can be ascertained by comparing the consequences of undertaking an evaluation with the consequences of not undertaking an evaluation. An example of a comparative outline is provided in Table 1.

Embedding an impact evaluation into plans for rolling out an intervention is generally more likely to represent value for money when results can be obtained as the intervention is being rolled out. In this scenario, there is an opportunity to improve or stop the rollout based on the results of the impact evaluation. This would be most likely to provide value for money when a pilot study is not possible and when it would be possible and practical to modify or stop the rollout, if needed, based on the results. An impact evaluation may also be useful after the programme has been fully implemented, for example if there is uncertainty about continuing the programme. Finally, the findings from an impact evaluation can – independent of timing – be useful for other policymakers who consider implementing a similar programme.

An example of an impact evaluation that was embedded in the roll-out of a programme is shown in Box 5 [12-14].

Impact evaluation, as is apparent from the discussion above, will most often be desirable whenever there is insufficient evidence, and this is commonly the case for changes that are implemented in health systems. However, resources for impact evaluations are limited and

rigorous evaluation may not always be possible. Therefore, just as it is important to decide how best to use scarce resources for healthcare, it is equally important to determine how best to use scarce resources for both impact evaluations and research generally.

5. How should the impact evaluation be done?

Attributing an observed change to a programme or policy requires a comparison to be made between the individuals or groups exposed to it, and others who are not. It is also important that the groups that are compared are as similar as possible, in order to rule out influences other than the programme itself. This can effectively be done by randomly allocating individuals or groups of people (e.g. within geographic areas) to both receive the programme and not to receive it, in what is termed a *randomised trial*. Usually such trials are conducted as pilot projects before a programme is introduced at a national level, but they can also be undertaken in parallel with full scale implementation. See Table 2 [15] for an overview of a number of evaluation designs. The weaknesses and strengths of each method mentioned in Table 2 are outlined in Table 3.

Randomised controlled trials may, however, not always be feasible. Alternative approaches include the comparison of changes, from before to after programme implementation, with changes during the same time period in areas where the programme was not implemented (e.g. in neighbouring districts or countries) in a process known as a *controlled before-after evaluation*. Alternatively, an *interrupted time-series* may be used in which data are collected from multiple time points before, during, and after programme implementation (Figure 2 provides an example of such a time series [11]).

Simply comparing the value of an indicator before and after programme implementation is not generally recommended since the risk of misleading findings is considered to be high – observed changes may be caused known and unknown factors other than the programme itself [16, 17].

Impact evaluations should be planned well ahead of programme implementation. Also, they are likely to be most informative if a qualitative component is included e.g. by conducting interviews or group discussions in an attempt to understand why the results came out the way they did [18].

See Box 6 for illustrative examples of methods for conducting an impact evaluation.

Rigorous evaluations can be expensive to conduct, and budget, time or data constraints may act as a disincentive to ensure rigorous implementation. Such constraints can impact on the reliability of impact evaluations in a number of ways:

- By compromising the overall validity of the results, for example, due to insufficient planning or follow-up, or through a paucity of baseline data, a reliance on inadequate data sources, and the selection of inappropriate comparison groups, and
- Through the use of inadequate samples, e.g. due to the selection of samples that are convenient to sample but may not be representative, through the use of sample sizes that are too small, and inadequate attention being given to contextual factors

Budget, time and data constraints can be addressed by starting the planning process early or finding ways to reduce the cost of data collection. However, it is important to ensure that neither the threats to the validity of the results, or the limitations of the sample, are such that

the results of the evaluation will be unable to provide reliable information. Before implementing an evaluation, an assessment should therefore be made as to whether an adequate evaluation is possible. If it is not, an assessment needs to be undertaken as to whether a programme should be implemented without prior evaluation, in the face of uncertainty about its potential impacts.

Impact evaluations are not worthwhile when findings are not used. Results should be monitored in order to inform decisions about whether to continue, change or stop existing programmes. Clearly, other interests will also need be taken into consideration, and decision makers may elect not to emphasise particular findings from certain evaluations when such findings, for instance, conflict with other interests that are perceived as more important [19].

Resources

Useful documents and further reading

MacKay K. How to Build M&E Systems to Support Better Government. 2007. Washington DC, The World Bank.
Available at: www.worldbank.org/ieg/ecd/docs/How_to_build_ME_gov.pdf
(Accessed Feb 25th 2009)

Barber S. Health system strengthening interventions: Making the case for impact evaluation. 2007. Geneva, Alliance for Health Policy and Systems Research.
Available at: www.who.int/alliance-hpsr/resources/Alliance%20%20HPSR%20-%20Briefing%20Note%202.pdf
(Accessed Feb 25th 2009)

Savedoff WD, Levine R, Birdsall N. When will we ever learn? Improving lives through impact evaluation. Report of the Evaluation Gap Working Group. 2006. Washington DC, Center for Global Development.
Available at: www.cgdev.org/content/publications/detail/7973/
(Accessed Feb 25th 2009)

Segone M (ed). Bridging the gap: The role of monitoring and evaluation in evidence-based policy making. UNICEF, the World Bank and the International Development Evaluation Association.
Available at: www.unicef.org/ceecis/evidence_based_policy_making.pdf
(Accessed March 2nd, 2009)

Monitoring and Evaluation (M&E): Some Tools, Methods and Approaches. 2004. Washington DC. The World Bank.
Available at:
lnweb90.worldbank.org/oed/oeddoclib.nsf/24cc3bb1f94ae11c85256808006a0046/a5efbb5d776b67d285256b1e0079c9a3/$FILE/MandE_tools_methods_approaches.pdf
(Accessed March 2nd, 2009)

Grimshaw J, Campbell M, Eccles M and Steen N. Experimental and quasi-experimental designs for evaluating guideline implementation strategies. Family Practice 2000; 17: S11–S18.
Available at: http://fampra.oxfordjournals.org/cgi/reprint/17/suppl_1/S11
Accessed May 19th 2009.

Links to websites

Independent Evaluation Group (IEG) at the World Bank: www.worldbank.org/ieg

International Initiative for Impact Evaluation (3ie): www.3ieimpact.org

Trial Protocol Tool and Trial Management Tool: http://www.support-collaboration.org/researchers.htm

NorthStar: http://www.rebeqi.org/?pageID=34&ItemID=35

Health Metrics Network: http://www.who.int/healthmetrics/en/

Box 1. Illustrative examples: Is monitoring necessary?

Scaling up provision of antiretroviral therapy (ART) in Malawi [2]

When Malawian health authorities decided to make ART available to a large proportion of the population who were HIV-positive, a system was put in place to monitor the implementation of this new policy. The principles of the system were based on the WHO-approach used for the monitoring of national tuberculosis programmes. Each patient started on ART was given an identity card with a unique identity number. This card which contained the recorded information was kept at the clinic.

Lung cancer surgery in Denmark [3]

Danish authorities issued national clinical practice guidelines for the management of lung cancer prompted by poor outcomes for patients subjected to lung cancer surgery. To monitor the implementation of the guidelines, a register of lung cancer patients was established which included specific information about the patients undergoing surgery.

Box 2. Illustrative examples: What should be measured?

Scaling up the provision of antiretroviral therapy (ART) in Malawi [2]

As part of the Malawian government's ART rollout programme, basic information is collected for new patients, including their name, address, age, height, name of guardian, and the reason for starting ART. Patients are asked to attend on a monthly basis to collect their medication. During their visit, their weight is recorded and they are asked about their general health, ambulatory status, work, and any drug side effects. Pill counts are also undertaken and recorded as a measure of ensuring drug adherence. In addition, the following standardised monthly outcomes are recorded using the following categories:

- *Alive*: Patient is alive and has collected his/her own 30-day supply of drugs
- *Dead*: Patient has died while on ART
- *Defaulted*: Patient has not been seen at all during a period of 3 months
- *Stopped*: Patient has stopped treatment completely either due to side effects or for other reasons
- *Transfer-out*: Patient has transferred-out permanently to another treatment

Lung cancer surgery in Denmark [3]

Indicators collected by the Danish Lung Cancer Registry include the extent ('stage') of cancer in the body, the surgical procedure used, any complications that occurred, and the survival outcome.

Box 3. Illustrative examples: How will the findings be utilised?

Scaling up the provision of antiretroviral therapy (ART) in Malawi [2]

Data collected as part of the Malawian monitoring system of the ART rollout may be analysed and used in a variety of ways. Comparisons can be made of treatment outcomes for patients who were recruited at different times. If, for example, the rate of switching from first- to second-line regimens increases, or rates of mortality do likewise, an increase in drug resistance to the first-line regimen could be the cause. If the rate of deaths or defaulters declines, this could indicate that the management of the ART treatment programme is improving. If outcomes are particularly poor in certain geographic areas or clinics, action may need to be taken to address this.

Lung cancer surgery in Denmark [3]

Data from the Danish Lung Cancer Registry are used, among other reasons, to monitor whether national recommendations for lung cancer surgery are being followed. Local, regional, and national audits are performed with the purpose of identifying problems or barriers that may impede adherence to the national guidelines. Based on the findings, specific strategies for quality improvement are then proposed.

Box 4. Illustrative examples: Should an impact evaluation be conducted?

Home-based antiretroviral therapy (ART) in Uganda [9, 10]

A major obstacle to scaling up the delivery of ART in developing countries is the shortage of clinical staff and/or difficulties with accessing care due to transportation costs. One proposed solution is home-based HIV care, where drug delivery, the monitoring of health status and the support of patients is carried out at the home of the patient by non-clinically qualified staff. However, it is highly uncertain whether this strategy is able to provide care of sufficient quality, such as timely referrals for medical care, or whether such a system is cost-effective. Therefore, before implementing home-based care programmes widely it is important that they are evaluated for their (cost-) effectiveness.

Mandatory use of thiazides for hypertension in Norway [11]

Policymakers in Norway decided that the prescription of thiazides as anti-hypertensive drugs would be mandatory for physicians instead of more costly alternatives, in instances where drug expenses were to be reimbursed. The policy was implemented nationally a few months after the decision was made. However, critics continued to argue that the new policy was unlikely to lead to the expected results. The Ministry of Health therefore decided to sponsor a study to assess the impact of the policy they had implemented.

Box 5. Illustrative example: Evaluation of a health system reform in Mexico [12-14]

In 2001, the Mexican government rolled out a new system of insurance called the Seguro Popular (or Popular Health Insurance scheme), to extend coverage to roughly 50 million Mexicans who were not yet covered by existing programmes. Taking advantage of the timetable of the progressive rollout, the government set up a randomised evaluation comparing the outcomes for those communities receiving the scheme with those still waiting for it. In addition to evaluating whether the reform achieved the outcomes intended and did not have unintended adverse effects, the evaluation also provides for shared learning. Information can be used to guide adjustments to the scheme.

Box 6. Illustrative examples: How should the impact evaluation be undertaken?

Home-based antiretroviral therapy (ART) in Uganda [9, 10]

To ensure a fair comparison between home-based and facility-based ART, researchers in Uganda conducted a randomised trial. The study area was divided into 44 distinct geographical sub-areas. In some of these, home-care was implemented, while in others a conventional facility-based system continued to be used. The selection and allocation of areas to receive, and not to receive, the home-based care system, was randomly determined. This reduced the likelihood of there being important differences between the comparisons groups, which might otherwise have influenced the study if the districts themselves had decided whether to implement home-based care, or if decisions were made based on, for example, their existing preparedness to implement home-based care. The random allocation system used was also the fairest way of deciding where to start home-based care since each district had an equal chance of being chosen.

Mandatory use of thiazides for hypertension in Norway [11]

The mandatory prescription of thiazides for treating hypertension was implemented right across Norway, and with an urgency that made a planned, rigorous impact evaluation impossible to conduct. However, by accessing the electronic medical records of 61 clinics at a later stage, researchers extracted prescription data from one year before to one year after the new policy was introduced. They analysed the data in using an interrupted time-series. Monthly rates of thiazide prescribing and other outcomes of interest were analysed over time to see if any significant changes could be attributed to the implemented policy. Analysis indicated that there was a sharp increase in the use of thiazides (from 10 to 25% over a pre-specified 3 month transition period), following which the use of thiazides levelled off (see Figure 2).

Table 1. Advantages and disadvantages of impact evaluations in relation to when the results become available

Timing of the evaluation	**Findings of the evaluation:** The balance between the benefits, harms and costs			
	Favour the intervention:		**Do not favour the intervention:**	
	Evaluation	No evaluation	Evaluation	No evaluation
Pilot study prior to rolling out the intervention	• Delay in rollout • Potential for improvements prior to rollout	• No delay • No potential for improvements prior to rollout	• Possible to stop rollout • Potential to reconsider options	• Not possible to stop rollout • No opportunity to reconsider options
Results available as the intervention is rolled out	• Potential for improvements	• No potential for improvements	• Possible to stop rollout or make modifications	• Not possible to stop rollout or make modifications
Results not available until after the intervention has been rolled out	• Support for continuation of the intervention	• No support for continuation of the intervention	• Would stimulate and inform reassessment of options and modification or withdrawal of the intervention	• Would stimulate and inform reassessment of options and modification or withdrawal of the intervention
Independent of timing	• Potential for others to learn	• No potential for others to learn	• Potential for others to learn	• No potential for others to learn

Table 2. Evaluation designs (adapted from the Cochrane Handbook for Systematic Reviews of Interventions [15])

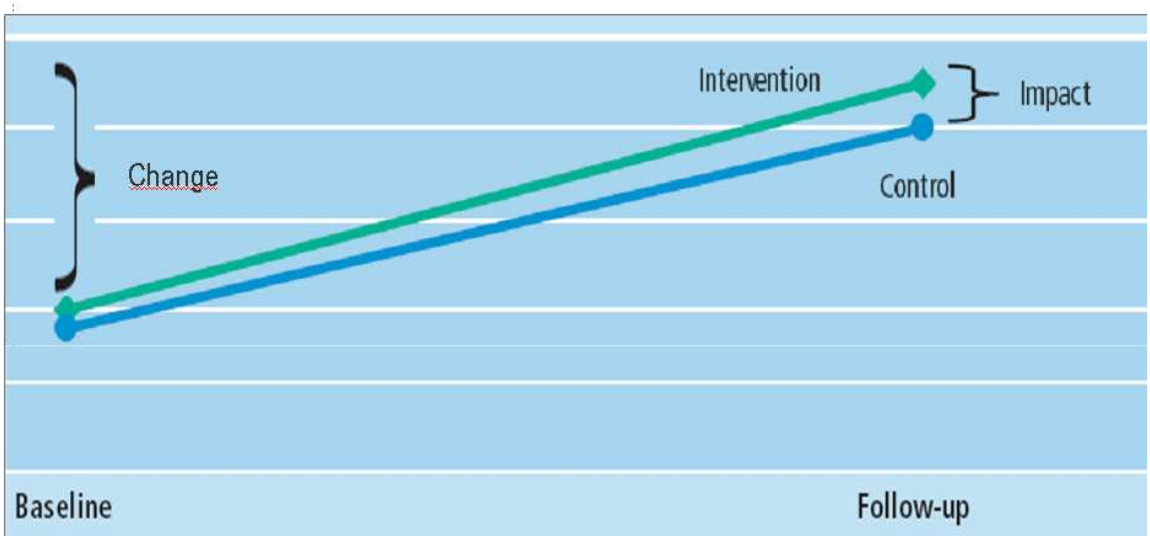
Randomised controlled trial	• An experimental study in which individuals are randomly allocated to receive different interventions (e.g. by the toss of a coin or using a list of random numbers generated by a computer)
Cluster randomised trial	• An experimental study in which groups of people (e.g. school classes or hospitals) are randomly allocated to receive different interventions
Non-randomised controlled trial	• An experimental study in which people are allocated to different interventions using methods that are not random (e.g. patients admitted during Week 1 week receive intervention A, those admitted in Week 2 receive intervention B, those in Week 3 receive intervention A again, etc.)
Controlled before-and-after study	• A study in which observations are made before and after the implementation of an intervention, both in a group that receives the intervention and in a control group that does not. Data collection should be done concurrently in the two groups
Interrupted-time-series study	• A study that uses observations at multiple time points before and after an intervention (the measurements are *interrupted* by the intervention). The design attempts to detect whether the intervention has had an effect significantly greater than any underlying trend over time
Historically controlled study	• A study that compares a group of participants receiving an intervention with a similar group from the past who did not
Cohort study	• A study in which a defined group of people (the *cohort*) is followed over time, to examine associations between different interventions received and subsequent outcomes. A *prospective* cohort study recruits participants before any intervention and follows them into the future. A *retrospective* cohort study identifies subjects from past records, describing the interventions received and follows them from the time of those records
Case-control study	• A study that compares people with a specific outcome of interest (*cases*) with people from the same source population but without that outcome (*controls*), to examine the association between the outcome and prior exposure (e.g. having an intervention). This design is particularly useful when the outcome is rare
Cross-sectional study	• A study that collects information on interventions (past or present) and current health outcomes for a group of people at a particular point in time, in order to examine associations between the outcomes and exposure to interventions
Qualitative study	• A study conducted in a natural setting which is usually designed to interpret or make sense of phenomena in terms of the meanings people bring to them. Typically, in such a study, narrative data are collected from individuals or groups of 'informants' or from documents, and then interpreted by the researcher(s)

Table 3. Selected strengths and weaknesses of evaluation designs

	Strengths	Weaknesses
Randomised controlled trial	• Widely considered to be the strongest design for establishing cause-effect relationships, which is the key focus of impact evaluation	• May be time consuming and represent logistical challenges • The results are not necessarily transferable to settings outside the study setting
Cluster randomised trial	• Same strengths as for ordinary randomised trials. In addition, the risk of 'contamination' is reduced e.g. that intervention A may be received by, or affect, individuals allocated to receive intervention B only. For example, if nurses are allocated randomly to implement a new routine, other nurses may be influenced by these changes and may start undertaking the same activities. It may therefore be better to randomise wards, and all of the staff within them, rather than individual nurses.	• Baseline differences may be a problem as the number units (or *clusters*) that are randomised would usually be lower than in a trial where individuals are randomised. May be time consuming and logistically challenging, but less so than an ordinary randomised trial
Non-randomised controlled trial	• May be easier and more practical to conduct than a randomised controlled trial	• When allocation is not done using random methods, selection biases may occur, e.g. because patients and health workers adjust their behaviour to the allocation procedure if they prefer one intervention to another
Controlled before-and-after study	• May be the only practical option, e.g. for large-scale interventions where randomisation is not feasible for practical or political reasons	• Known or unknown differences between the groups that are compared may exert more influence on the findings than the fact that they received different interventions. Consequently, drawing conclusions about cause-effect relationships may be risky • Requires the availability of baseline data
Interrupted-time-series study	• May be feasible and relatively easy to conduct, if the necessary data are made available. No	• The effect size is always difficult to estimate in such analyses because influences other than the intervention under investigation may impact on the

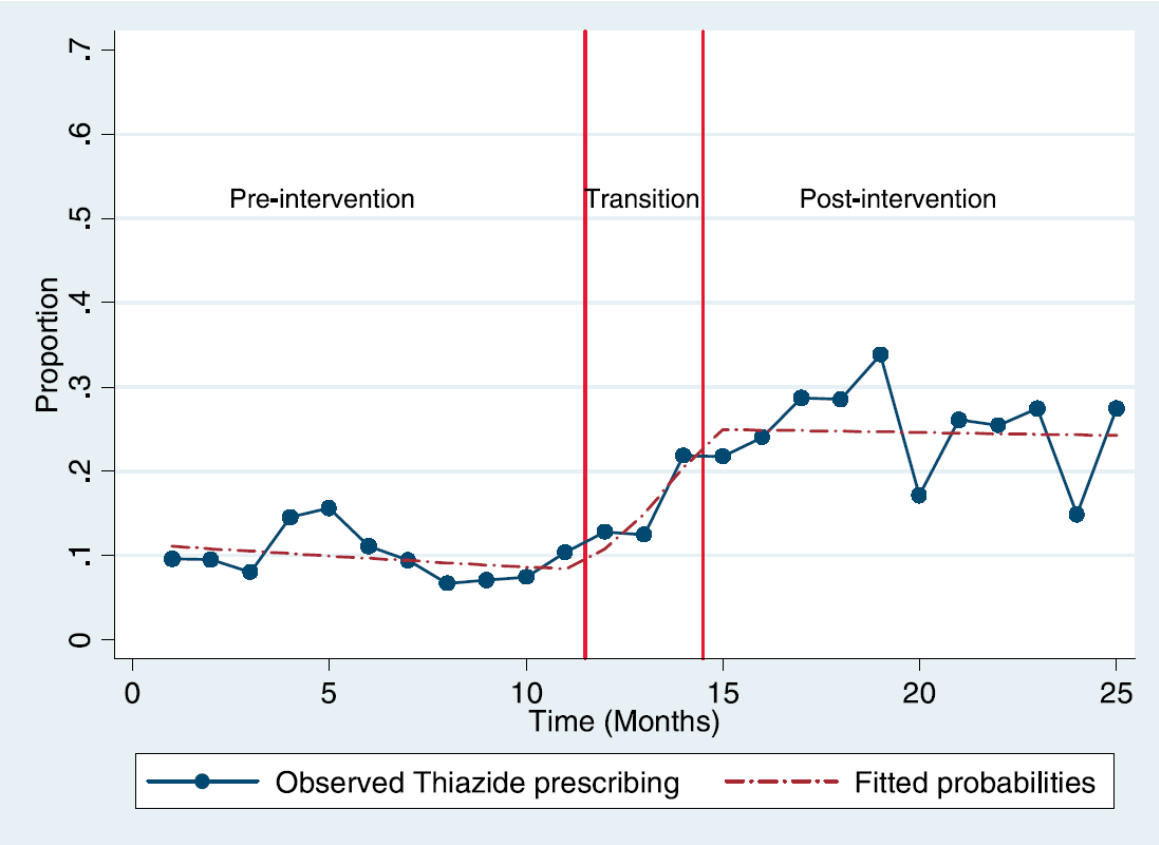
	Strengths	Weaknesses
	control group required	observed changes
Historically controlled study	• May be quickly and easily done if the necessary data are available	• Known or unknown differences between the groups that are compared may exert more influence on the findings than the fact that they received different interventions. Consequently, drawing conclusions about cause-effect relationships is risky
Cohort study	• Often large studies with high degree of external validity (i.e. the findings can be generalised). Often conducted over several years, which makes it possible to detect the long-term effects of an intervention	• Cohort studies are typically lengthy and costly, mainly due to the need for following up the – usually – high number of participants • Known or unknown differences between the groups that are compared may exert more influence on the findings than the fact that they were exposed to different interventions. Consequently, drawing conclusions about cause-effect relationships is risky
Case-control study	• More quickly and easily done than cohort studies	• The retrospective nature of such studies entails collecting information about events that have taken place earlier, which may be a source of error • Known or unknown differences between the groups that are compared may exert more influence on the findings than the fact that they received different interventions. Consequently, drawing conclusions about cause-effect relationships is risky
Cross-sectional study	• Requires no follow-up time and can therefore be conducted quickly and often at a low cost	• Known or unknown differences between the groups that are compared may exert more influence on the findings than the fact that they received different interventions. Consequently, drawing conclusions about cause-effect relationships is risky
Qualitative study	• Allows for the collection of more in-depth information than other quantitative designs. This enables an understanding of how interventions and programmes are, or are not, working	• Does not generate data that can be used to estimate the effect of an intervention, beyond the perception of those who are interviewed or surveyed

Figure 1. Comparing change in performance in two areas, one with and one without intervention (adapted from Barber [7])



Key:
With intervention: Green line
Without intervention: Blue line

Figure 2. Example of interrupted time-series analysis (taken from Fretheim et al [11])



References

1. Development Assistance Committee Working Party on Aid Evaluation. Glossary of Key Terms in Evaluation and Results Based Management. 2002. Paris, OECD Publications.

2. Harries AD, Gomani P, Teck R, de Teck OA, Bakali E, Zachariah R *et al*.: **Monitoring the response to antiretroviral therapy in resource-poor settings: the Malawi model.** *Trans R Soc Trop Med Hyg* 2004, **98:** 695-701.

3. Jakobsen E, Palshof T, Osterlind K, Pilegaard H: **Data from a national lung cancer registry contributes to improve outcome and quality of surgery: Danish results.** *Eur J Cardiothorac Surg* 2009, **35:** 348-352.

4. Smith PC, Mossialos E, Papanicolas I. Performance measurement for health system improvement: experiences, challenges and prospects. Background Document for WHO European Ministerial Conference on Health Systems: "Health Systems, Health and Wealth". 2008. Copenhagen, World Health Organization, Europe.

5. Campbell SM, Braspenning J, Hutchinson A, Marshall M: **Research methods used in developing and applying quality indicators in primary care.** *Qual Saf Health Care* 2002, **11:** 358-364.

6. MacKay K. How to Build M&E Systems to Support Better Government. 2007. Washington DC, The World Bank.

7. Barber S. Health system strengthening interventions: Making the case for impact evaluation. 2007. Geneva, Alliance for Health Policy and Systems Research.

8. The World Bank. Impact evaluation: Overview. http://go.worldbank.org/2DHMCRFFT2 . 2009. The World Bank.

9. Amuron B, Coutinho A, Grosskurth H, Nabiryo C, Birungi J, Namara G *et al*.: **A cluster-randomised trial to compare home-based with health facility-based antiretroviral treatment in Uganda: study design and baseline findings.** *Open AIDS J* 2007, **1:** 21-27.

10. Jaffar S, Amuron B, Birungi J, Namara G, Nabiryo C, Coutinho A *et al*.: **Integrating research into routine service delivery in an antiretroviral treatment programme: lessons learnt from a cluster randomized trial comparing strategies of HIV care in Jinja, Uganda.** *Trop Med Int Health* 2008, **13:** 795-800.

11. Fretheim A, Havelsrud K, MacLennan G, Kristoffersen DT, Oxman AD: **The effects of mandatory prescribing of thiazides for newly treated, uncomplicated hypertension: interrupted time-series analysis.** *PLoS Med* 2007, **4:** e232.

12. Moynihan R, Oxman A, Lavis JN, Paulsen E. Evidence-Informed Health Policy: Using Research to Make Health Systems Healthier. Rapport nr. 1-2008. 2008. Oslo, Nasjonalt kunnskapssenter for helsetjenesten.

13. Frenk J, Gonzalez-Pier E, Gomez-Dantes O, Lezana MA, Knaul FM: **Comprehensive reform to improve health system performance in Mexico.** *Lancet* 2006, **368:** 1524-1534.

14. Gakidou E, Lozano R, Gonzalez-Pier E, Abbott-Klafter J, Barofsky JT, Bryson-Cahn C *et al*.: **Assessing the effect of the 2001-06 Mexican health reform: an interim report card.** *Lancet* 2006, **368:** 1920-1935.

15. Cochrane Collaboration. *Cochrane Handbook for Systematic Reviews of Interventions*. Chichester: The Cochrane Collaboration and John Wiley & Sons Ltd.; 2008.

16. Savedoff WD, Levine R, Birdsall N. When will we ever learn? Improving lives through impact evaluation. 2006. Washington DC, Center for Global Development.

17. Shadish WR, Cook TD, Campbell DT: *Experimental and Quasi-Experimental Desings for Generalized Causal Inference* . Houghton Mifflin; 2002.

18. Lewin S, Glenton C, Oxman AD. **How are qualitative methods being used alongside complex health service RCTs? A systematic review**. BMJ, In Press

19. Scheel IB, Hagen KB, Oxman AD: **The unbearable lightness of healthcare policy making: a description of a process aimed at giving it some weight.** *J Epidemiol Community Health* 2003, **57:** 483-487.