

DOES IT WORK?

A simple guide to evaluation with an emphasis on injury prevention

March 2006



UNIVERSITY OF WALES SWANSEA

Ronan Lyons, MD.
Professor of Public Health
School of Medicine
University of Wales Swansea



Mariana Brussoni, Ph.D.
Research Fellow
Centre for Child & Adolescent Health
University of the West of England

Introduction

The purpose of this document is to provide simple guidance to practitioners and volunteers who have been asked to provide an evaluation of their project or programme. Whilst the emphasis is on injury prevention the general principles also apply to other interventions aimed at improving health.

Many interventions or services are set up with short term funding and at some stage the organisers are asked to evaluate the effectiveness of the intervention or service in order to continue in existence. This document aims to provide basic guidance on the nature of evidence, evaluation and how and when to undertake such work. The document covers topics such as

- how we know when something works,
- the different types of study design,
- what types of outcome measures can be used,
- how many people or participants need to be in a study, and provides
- links to documents which provide more detail.

We have tried to avoid as many technical terms as possible but some are unavoidable and where they are included they are in *italics* and are explained in the appendix.

How do we know that something works?

This depends on the level of evidence needed to convince whoever needs convincing. People, patients, practitioners, researchers, funders and other groups may vary in what they consider to be sufficient evidence. Healthcare appears to require a higher level of evidence than many other sectors e.g. education and justice policies.

When something is obviously effective – i.e. the change in outcome is so noticeable and follows closely on the intervention – then formal evaluation is unnecessary and may even be unethical. These obvious outcomes do not require complex evaluations. For example, no rigorous evaluations have ever taken place to demonstrate the effectiveness of immobilising broken bones in plaster of Paris. Nobody would ever suggest formally testing this intervention, although they might test different types of immobilisation. Likewise, the introduction of penicillin so dramatically changed survival in severe infections that it has not been tested in a scientific trial. However, most healthcare and public health interventions are not so dramatically obvious and require evaluation.

Evaluation of interventions aimed at determining effectiveness (change in outcome) involves a variety of steps. These may include basic observation and intuition that a given intervention may work; small-scale feasibility studies to get a sense of whether an intervention is doable; pilot studies to test out methods before launching major studies; and large-scale trials to understand whether an intervention works in practice. Sometimes, a societal impact evaluation is carried out to determine whether the predicted benefits from trial data are actually realised in different places or settings.

Evaluation can be carried out at different levels of complexity depending on the question it is intended to answer. The following levels are appropriate to the following questions.

Is what we are doing or propose to do evidence based (scientifically shown to work)?

What is required here is a literature search to determine whether other researchers have tested the intervention and shown it to work. Finding the relevant studies is the big issue. The scientific literature is huge and growing every day. Nobody can be said to have read all the studies in their specialist areas. To manage this, issue groups have come together to summarise the literature on various topics. The most famous of these is the Cochrane Collaboration that summarises evidence from randomised control trials of healthcare and the Campbell Collaboration that carries out similar work in non-healthcare settings. The crucial thing is to search first for [systematic reviews](#) or [meta-analyses](#) of the literature before searching for individual studies. These are high quality summaries of the best evidence carried out by skilled researchers. It makes sense not to try to do something when someone else has provided the answers. Finding all the evidence is a skilled task and librarians or knowledge management specialists can provide guidance on the most appropriate search strategies to use to discover the existing evidence. The CAPIC website has a searchable database of all the systematic reviews covering injury prevention as a service to injury prevention practitioners.

How can we find out if our proposed research is already known to work so that we do not duplicate existing work?

Before carrying out any evaluation it is necessary to find out what is already known on the topic (see above). This will avoid repeating evaluations that are already available and also learning about the types of design and outcome measures used by others (thus avoiding poor quality work). If a rigorous evaluation has already been carried out elsewhere on your proposed intervention, it may be more appropriate to conduct an evaluation that focuses on implementation issues relevant to your particular setting. The purpose of a [process evaluation](#) is to show how well the intervention was implemented in your local setting.

Do we need complex and expensive trial designs? Will a simple before and after comparison not do?

Almost everything changes over time. A change over time will not convince many that the change was due to the intervention. Other obvious causes for change include time trends (the rate was falling or rising anyway), the effects of other interventions happening simultaneously, and the phenomenon called [regression towards the mean](#). The latter is a very common phenomenon which is largely due to time dependant variation in small numbers or in measurement of a biological variable. For example, if we plotted the location of car crashes over a three year period we would find many areas in which there was a larger than expected number – a cluster. If we then did nothing but mapped crash locations for a second three year period we would find many of the clusters had

disappeared and many more would have appeared in different areas – all due to [*regression towards the mean*](#). Areas chosen for having a high initial rate tend to have lower rates at a second measurement and visa versa. In some areas the cluster would have remained and these areas are likely to have a characteristic which results in a constantly high injury rate. But if an intervention had occurred, would many people be convinced of its effectiveness? Those who are aware of this [*regression towards the mean*](#) phenomenon are unlikely to accept a simple [*before and after*](#) comparison as providing sufficient evidence in many instances. The way around the [*regression towards the mean*](#) problem is to have multiple before and after measures so that one is confident of choosing a group or area with persistently high levels.

Having a study design with multiple before and after measures will not get around other potential problems such as other non-measured influences on the outcome that are happening at the same time and could be responsible for the change. The way around this issue is to use control groups or populations which are closely matched with the intervention group but do not have the intervention. The change (after minus before) in the intervention group is compared with the change in the control group. It is important to think about other changes which are happening at the same time as the intervention and could affect the results. For example, if you were trying to measure the effect of pedestrian training on road injuries by comparing two areas, one having the intervention and the other not, and at the same time additional traffic calming was placed in one of the areas it might be difficult to determine how much of any change was due to the pedestrian training.

The [*before and after comparison with a control area or group*](#) is a better research design than one which only measures these factors in the intervention area ([*before and after study*](#)). However, it may still mislead if the researchers choose intervention subjects because they think a better outcome is more likely in this group or because the intervention group know that they are in that group and this might make a positive response more likely. This is related to wishful thinking or something called the [*placebo effect*](#). The [*placebo effect*](#) is very common and is best described in studies of new medicines. It occurs in around 30% of people who improve after receiving dummy or pretend treatments. People who are placebo responders are very lucky, as they tend to get better whatever form of treatment is used. It also occurs in many other types of study.

To get around these more subtle problems or ([*biases*](#)), a research design called the [*double blind randomised control trial*](#) is used where neither the investigators nor the participants know who is in the intervention or control group until the study is over. This type of trial is the gold standard for evaluation but is not always practical. In most injury prevention situations it is not possible to hide who gets the intervention and a [*single blind*](#) or [*unblinded \(open\)*](#) study may all that can be achieved.

Sometimes when not having an intervention is unacceptable and several areas or groups are involved it is possible to randomise them into early or late intervention groups with the late group acting as a control for the early group ([*randomisation by time*](#)). Another variant of this involves different areas all receiving one of two interventions with those receiving one

intervention acting as controls for the other intervention and visa versa ([randomisation by topic](#)).

What type of evaluation should be chosen?

This depends on the purpose of the evaluation. A flow-chart at the end of this section explains some of the options available.

If the intervention is novel and we do not know whether it works or is better than the current alternatives, then the best type of design is one of the [controlled trial designs](#). These will need to be preceded by simple [case series](#) with the use of *before and after* measures to be able to design the trial appropriately and make sure that enough people (to make sure an important effect can be detected) but not too many (to reduce costs) are included.

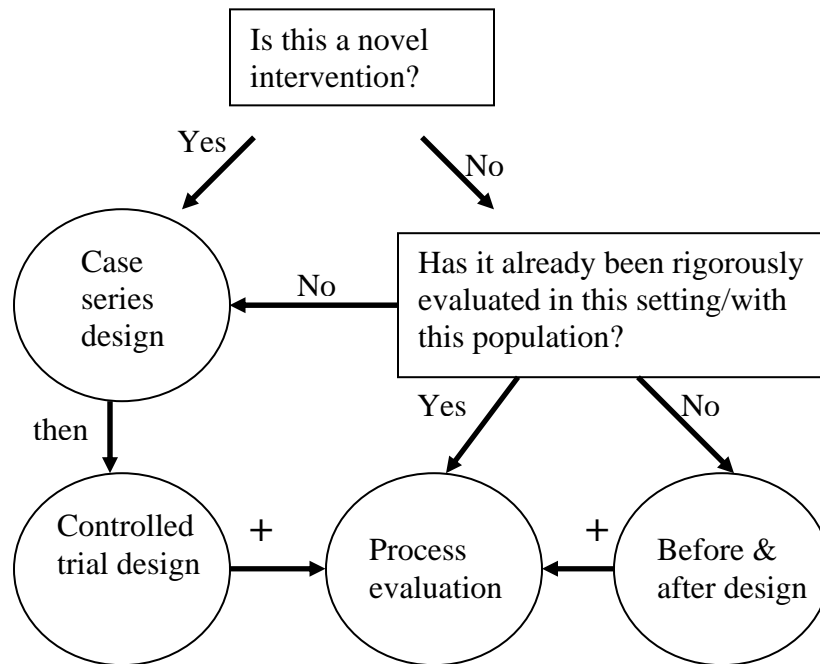
If the intervention is already known to work then it could be argued that as we know it works further outcome evaluation (i.e., to see if it works or not) could be unnecessary, and a [process evaluation](#) is sufficient. If an outcome evaluation is still thought necessary, perhaps in the case of implementing an intervention in a completely novel setting, then a [before and after design](#) is commonly used. Such ‘evaluations’ may, or may not, persuade funders and policy makers, depending on their level of scientific training or desire to accept a ‘positive’ result.

Sometimes evaluation is tagged on as an afterthought to a policy or investment which has cost a considerable amount of money, but is not grounded in scientific evidence, and requires a retrospective evaluation to justify the original spend. In these circumstances it is nearly always impossible to carry out an outcome evaluation due to the intervention but an evaluation of process is possible and is all that should be attempted.

If the purpose behind the funded intervention is to promote enhanced uptake in a particular group e.g. deprived areas, then a [before and after comparison with a control area](#) is quite reasonable. Enhanced uptake of an effective intervention in deprived groups could be deemed to be a sufficiently convincing outcome. Evaluations using an [area deprivation score](#) do have a drawback. They assume that everyone in a deprived area is equally deprived. This may not necessarily be the case and perhaps more affluent groups living within these areas are more likely to take up the intervention. The only way around this is to measure ‘deprivation’ at the individual level using a surrogate such as social class, income, educational level, or uptake of benefits. Such measures are more time consuming and costly to collect and analyse.

Evaluation of [process measures](#) is often all that is required if the question relates to the effectiveness of a service or intervention already based on evidence. Since an evaluation of outcomes needs an experimental design with control groups these will be missing in the new service where everybody gets the intervention. In such circumstances the most important things to evaluate are process measures. These include measurement of the types of people receiving the service (are they similar to those in whom the intervention was

shown to work in the scientific literature?), the extent to which the intervention happened (e.g. what proportion of children received practical pedestrian training?), and measures of target group, user and staff satisfaction with the new service or intervention. It may be possible to estimate the outcomes based on these issues. For example, if the scientific literature shows that for every ten people receiving the intervention two avoid injury then the numbers estimated to improve could be predicted (but not measured) from an analysis of those taking up the intervention.



What outcome measures should or could be used?

Again this depends on the perspective of the evaluation and evaluatees. There are many possibilities including:

- Delayed death or longer survival
- Fewer injuries
- Less severe injuries
- Less disability
- Lower costs to individuals, services, and society
- Greater uptake of services or programmes
- More choice for people
- Change in knowledge
- Change in attitudes
- Change in behaviour
- Enhanced public/user satisfaction
- More equitable uptake by population subgroups

All of the above are valid outcomes. No single evaluation measures all potential outcomes and choices are determined by preferences and costs. The important issue is to consider the range of relevant outcomes before deciding which are best or most affordable.

How many people or participants will be needed to find out if the intervention has worked?

This is quite a difficult question to answer and requires several pieces of information. If a scientific answer is required then you may be expected to have calculated something called the *sample size* or *power of a study* required to have a good chance of showing an effect. Normally, the following information is needed: how common is the event or estimated frequency of the desired answer before the intervention happens and how much success do you think you will realistically receive. For a formal calculation of sample size before a study starts you should seek advice from a statistician. The appendix contains examples of the calculation of sample size based on the table below.

The following table shows the average frequency per 1000 people of different types of injury happening in the home. It is derived from the 2002 (and final) Report of the Home Accident Surveillance System (HASS), previously run by the Department of Trade and Industry. The information in the HASS database is now held by the Royal Society for the Prevention of Accidents (RoSPA www.rosipa.com).

Table 1: Frequency different types of injury happening in the home per 1000 people in 2002 from Home Accident Surveillance System Database held by the Royal Society for the Prevention of Accidents (RoSPA www.rosipa.com).

Age Group and Sex						
Mechanism of injury	0 to 4		5 to 14		75 and over	
	Male	Female	Male	Female	Male	Female
Typical number of injuries per 1,000 population						
Fall on same level (slip/trip/stumble)	26	20	8	7	17	36
Fall on/from stairs/steps	15	13	3	5	6	8
Fall from ladder/stepladder	0.3	0.2	0.3	0.2	1.1	0.3
Fall from buildings/structures	0.6	0.5	0.4	0.3	0.1	0.2
Other falls	42	34	10	10	19	38
All falls	84	67	24	23	43	83
Struck by injuries	32	17	20	15	6	9
Pinch or crush injuries	8	7	3	3	0.6	0.6
Cut/tear/puncture wounds	6	4	9	1	3	2
Foreign Body	10	10	3	3	1	0.5
Suffocation	1	0.7	0.2	0.2	0.4	0.3
Poisoning	9	8	0.7	0.4	0.2	0.1
Thermal injuries (burns and scalds)	9	8	2	2	8	0.7
Total	159	121.7	61.9	47.6	62.2	96.2

Car crash and collision injuries are very common. The likelihood of a person being injured each year in a car crash or collision in the UK is around 1 in 200. More detailed information on the frequency of different types of injury can be obtained from many of the analyses provided on the CAPIC website www.capic.org.uk or by linking from this to other websites.

Are there any more detailed sources of information that provide guidance on how to evaluate projects?

Yes, there are several sources available. The best one we have come across which deals with injuries is the *Injury Prevention Programme Evaluation Manual* produced by the British Columbia Injury Prevention Research Unit in Canada. This is available on the internet on the University of British Columbia Injury Research Unit website <http://www.injuryresearch.bc.ca/> under publications. There are many useful injury and behavioural scales to be found in the tools repository on the website. This is also saved as a PDF file on the CAPIC website www.capic.org.uk).

This is a very user-friendly manual, outlining in simple terms the evaluation process. It includes worksheets that you can work through to plan your evaluation. These worksheets include:

- Determining who will be involved and who is interested in your evaluation;
- Writing programme goals and objectives;
- List of objectives for each goal;
- Determining your target population and how to access them;
- The activities you will need to undertake to meet your objectives;
- Determining your evaluation timeline;
- Available resources to conduct the evaluation;
- How to collect necessary data;
- Determining the audience for your evaluation findings and report;
- Determining next steps.

The manual also works through a sample injury prevention programme to demonstrate the planning of an evaluation.

The National Centre for Injury Prevention and Control at the Centres for Disease Control in the US has also produced a report called *Demonstrating Your Programme's Worth*: www.cdc.gov/ncipc/pub-res/demonstr.htm .

Appendix

Definitions:

Area Deprivation Score - A score given to all people living within a certain area, usually enumeration districts or wards which are areas used to elect local politicians termed councillors. Deprivation scores are based on an index or score such as the Carstairs Score. Four variables from census data are used in the Carstairs Score calculation: overcrowding, male unemployment, low social class, and no access to a vehicle. The Townsend Index of Material Deprivation is similar to the Carstairs Score but is made up from slightly different census variables. There are also separate indices of multiple deprivation for each of the countries of the UK calculated from a mixture of census and non census variables e.g. Index of Multiple Deprivation, Welsh Index of Multiple Deprivation.

Biases – Anything which might lead the investigator to come to a wrong conclusion.

Before and after study,(also called Pre-Post Study/Design) – A type of study which measures something before and after an intervention and uses the difference to determine whether the intervention was successful.

Before and after study with a control area (or group) - A study which measures something before and after an intervention in both the intervention area (or group) and control area (or group) and uses the relative difference to determine whether the intervention was successful.

Case series – A descriptive account of a number of cases. Cases can refer to injury events, injured people, or interventions.

Controlled trial designs – A study design in which there is a control group and in which an experiment (=trial) has taken place. The experiment refers to the intervention.

Double blind randomised control trial – A study design in which neither the investigators nor the participants know who is in the intervention or control group until the study is over. This is achieved by one person randomly allocating participants to the intervention or control group. Others involved in giving or receiving the study intervention or control intervention do not know which group they belong to and the person analysing the data also does not know which is the intervention group until after the results are known.

Meta-analysis – A method of combining the results of a number of studies which produces an average or summary numerical finding.

Power of a study – Studies are designed with the ability to detect an important result but not to include too many people due to cost or ethical considerations. Most studies are designed to have a *power* of at least 80%, that is, an 80% or more chance of detecting an important effect if it is really present. Increasing the power of the study involves increasing the sample size, often very substantially.

Placebo effect – The act of receiving any intervention may itself have an effect on the participants because they are expecting an effect. Control groups in trials of new medications receive a dummy inert medication (*placebo*) which looks exactly like the real medication. Usually around 30% of participants receiving *placebo* treatment report improvement.

Process evaluation – An evaluation on how a programme is implemented. This might involve describing how successful the programme was in engaging with people, whether the participants were from the intended group, how much it cost to deliver in time and resource, etc.

Process measures - Measures that describe how your programme was implemented. These are helpful in explaining the results of the programme and determining if anything needs to be changed. Examples of measures include:

- number of sessions participants attended;
- number of participants attending each session;
- where and when the intervention took place;
- participants' satisfaction with the programme;
- were there any similar programmes that participants attended at the same time as your programme that may affect your results?
- were there any other events in the community that may have affected your participants?

Randomisation by time – A technique used where it is not possible to have a control group which does not receive the intervention. The participants or areas involved are randomly allocated to two groups with one group receiving the intervention in the first phase and the second group receiving the intervention in the second phase. In the first phase the second group acts as the control group for the first group.

Randomisation by topic – Another technique used where it is not possible to have a control group which does not receive the intervention. The participants or areas involved are randomly allocated to two groups with one group receiving one intervention and the other group receiving the other intervention. The groups not receiving an intervention act as controls for those receiving the intervention.

Regression towards the mean – A phenomenon where an individual, group or area chosen because of a high rate tends to have a lower rate in a second period of measurement. This also occurs with groups having low rates at the first measurement. It occurs because many phenomena vary over time and do not have a constant rate. The first measurement may just happen to detect a group or individual with an unusually high or low rate for them at that point in time and subsequent measurements will be nearer their usual or average rate.

Sample size - The number of people or groups which needs to be included in the study to have a good chance of detecting whether the intervention works. A good chance may be referred to as the *power* of the study. The numbers needed will depend on whether the outcome measure is a category (e.g. success, failure) or a continuous measurement (e.g. satisfaction or pain scores), how much natural variation occurs in the measurements and

how much of a difference the intervention is expected to make. Studies using categorical outcomes need much larger sample sizes than those using continuous outcomes. An example of a sample size calculation is included below.

Single blinded randomised controlled trial - This is similar to a *double blinded randomised controlled trial* but where either the investigator or participant knows who is in the intervention or control group.

Systematic reviews – A method of finding all the scientific studies on a particular topic, appraising their quality and summarising the results of the high quality studies. This saves the investigator the trouble of reading a very large number of studies to find out whether an intervention works.

Unblinded (or open) study – A study design where both the investigator and the participants know who is in the intervention and control groups.

Sample Size & Power Calculations:

For many outcome measures, such as a reduction in the number of events or an increase in the proportion of people with a success, the numbers needed in a study to show quite large relative differences (say a 50% change) can be very large. A study which aimed to half the injury rate of 129/1000 to young male children (taken from the table of injury frequency above) would need 264 children in the intervention group and 264 children in the control group to have an 80% *power* of detecting a difference of reducing the rate by half. If a 50% reduction seemed to be too difficult to achieve and a smaller reduction of 25% was thought to be more achievable the sample size is much larger – 1238 in each group.

If a continuous outcome measure was used – such as a knowledge score which could be measured out of 10, then much smaller sample sizes are needed. For example, a study which aimed to improve knowledge scores from an average of 5/10 to 6/10 with a variability of 1/10 in the measurement of knowledge (technically called a standard deviation or sigma), then 16 participants would be needed in each of the intervention and control groups.

These sample sizes were taken from an online sample size calculator from the University of British Columbia - <http://newton.stat.ubc.ca/~rollin/stats/ssize/> . Most injury prevention practitioners would still need to seek advice from a statistician before starting a study.