

Chapter 9 : OUTCOME EVALUATION DESIGNS

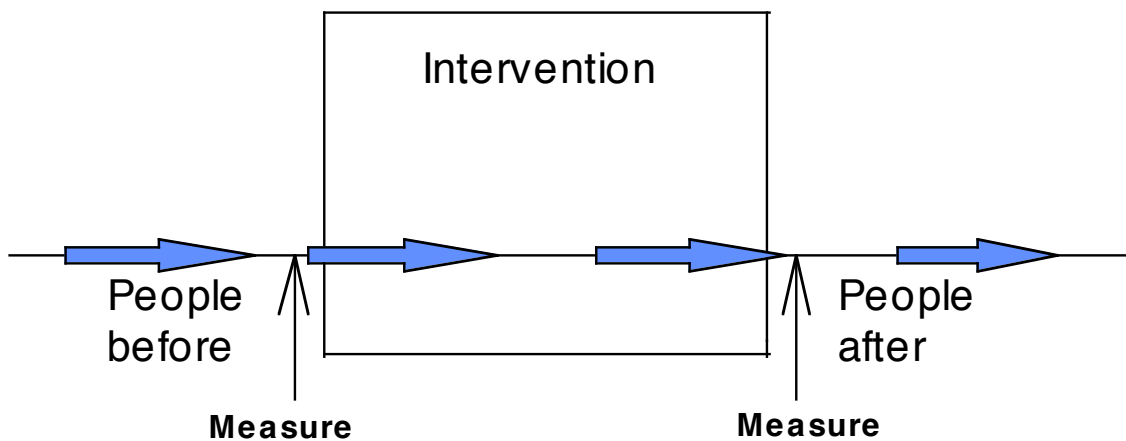
From, Øvretveit, J (2002) Action Evaluation of Health Programmes and Change A handbook for a user focused approach, Radcliffe Medical Press, Oxford.

1: INTRODUCTION

Outcome evaluations designs aim to find out the effects of an intervention. These designs are to answer the question “does it work?” or “does it work better than the alternative?” They do this by comparing the target people or organisations before and after the intervention. For example, by comparing patients states before and after receiving a service, or by comparing the health of a population before and after a health programme or change is introduced. Some may also try to find out if the intervention makes a difference to people other than those who are the targets. For example, what effect does it have on health personnel or patient’s carers? The type 6 designs use the before and after design to look at the effects of an intervention to an organisation, such as a training programme or a health reform.

Some designs do not measure before and after but just ask people after to judge if they thought the intervention had an effect. The difference between the designs is in the degree of certainty which each gives about whether the outcomes were actually produced by the intervention.

2: TYPE 3 – SINGLE BEFORE AND AFTER



If this design was used in the rehabilitation example in chapter 6, the evaluation would only have looked at the active rehabilitation service, and would not also have looked at the outcomes for the traditional service. It would have been less expensive to do so, but we would not know if the outcomes were better or worse than the usual service which people received in primary care.

The box drawing shows that people “pass through” an intervention and are then no longer exposed to it. The drawing thus can be misleading for some types of interventions which do not stop working on patients, or for many health reforms or policies. Sometimes people continue receiving the service – the “box” is not closed at the “right end”. Examples are medication for hypertension, interventions for chronic disorders, or home care services. For these, the diagram shows that a measure is taken before they start and then at periods after first receiving the treatment.

This is the simplest type of design for discovering if an intervention may have made a difference to the people receiving it. It is often designed as an experiment and the evaluator predicts the effects of the intervention on the targets. It can be used retrospectively where the evaluator looks back at “before and after” states. The purpose of the design is to help to judge the value of an intervention by comparing the state of people or organisations before, with their state after the intervention using outcome data or measures. The questions this design aims to answer are: what are the effects? or, what difference does the intervention make to the target?

The before-after comparison may be of specific features of the people before and after (eg blood pressure, or stress levels, or how many attend the clinic in a month), or of a number of features of the target which are collected before and after (eg a set of employee perceptions, physical measures, providers assessments of “patient progress”, employee income losses etc). The before after-comparisons are based on theories and predictions about the possible effects of the intervention.

The weaknesses of the design are that it cannot give conclusive evidence of effects. This is because, if a before-after difference is found, the difference may have been caused by things

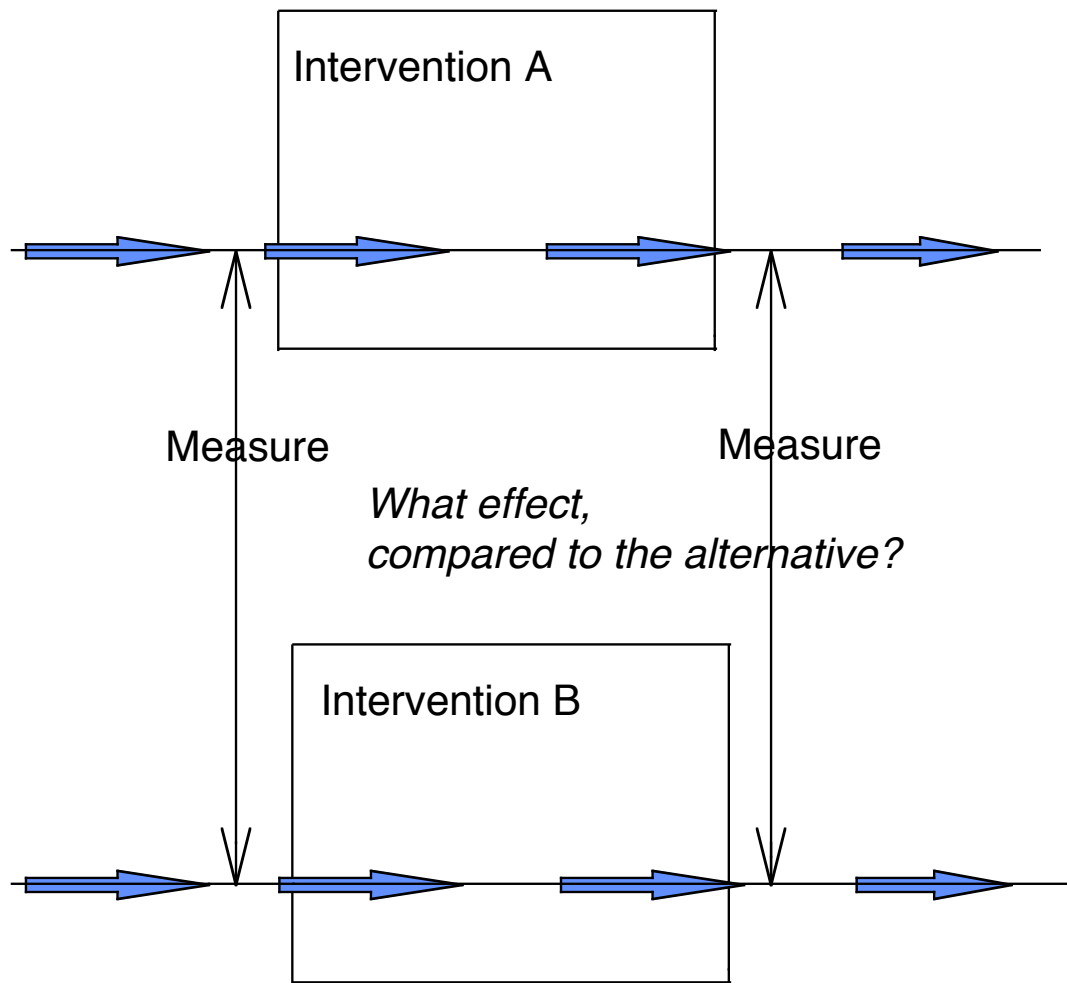
other than the intervention. The difference may be due to selecting subjects who would show these effects over time anyway, for example patients health state would have improved without the intervention. The many possible “confounding variables” are not controlled, controls being one way to exclude variables as explanations for the effects.

The strengths are that evaluations using this design can be small-scale and relatively quick. The design can be arranged to use few resources if the evaluator select a small number of subjects, makes one or a few simple “before” and “after” measures, and makes the measurement soon after the intervention. This makes it a useful design for practical professional-self evaluation or in a quality project.

Compared to the process designs 1 and 2 described in the last chapter, this design is, for many, the first recognizable evaluation design because it looks at outcome. However, those who take this view would also be dissatisfied with how this design tries to discover the effect of the intervention: how does this design prove that the difference is due to the intervention and not to something else? Would the difference have happened anyway without an intervention?

One way to improve this design is to gather a series of “before” data and a series of “after” data rather than just gather data once “before” and once “after”. For example, a simple type 3 design might gather data about how many people attend a clinic over one month before and one month after an intervention which publicizes the clinic. An improved type 3 design could gather data about monthly attendance for each of three months before and three months after the intervention. By doing this we can see the variability of attendance over time which will happen for many reasons, and then get a better idea whether the difference before and after the intervention is significant. Another way is a time series design which stops the intervention and starts it again and studies whether the targets revert to their before states after the intervention is stopped and then show changes when the intervention is started again.

3: TYPE 4 – COMPARATIVE



The questions addressed by this design are: what are the effects of the intervention, compared to a similar intervention or to the status quo elsewhere? The rehabilitation example used a design of this type: it compare the before and after measures for two groups, one of which received the active rehabilitation and the other the traditional rehabilitation. The design is also used, for example, for economic evaluations of two different employment policies, or an evaluation of a health reform or policy in one area compared to the status quo elsewhere. It could be used to evaluate a training programme compared to a written information intervention, or to compare two or more different types of services.

The design is like a type 3 outcome evaluation, but compares the outcomes of two groups undergoing different interventions. Normally it is carried out prospectively, but retrospective type 4 designs are possible (eg some comparative evaluations of different policies or working conditions). One variation of the design is a comparison of end-states only, rather than a comparison of the before-after change (outcome).

The weaknesses of the design are that it is expensive, and difficult to prove that the effects were due to the interventions alone, rather than due to other factors. The strengths are that, if the design is made with care, such evaluations can suggest which of the two interventions is

more effective or cost effective. The design is suitable where it is unethical or impractical to treat or intervene on only one group.

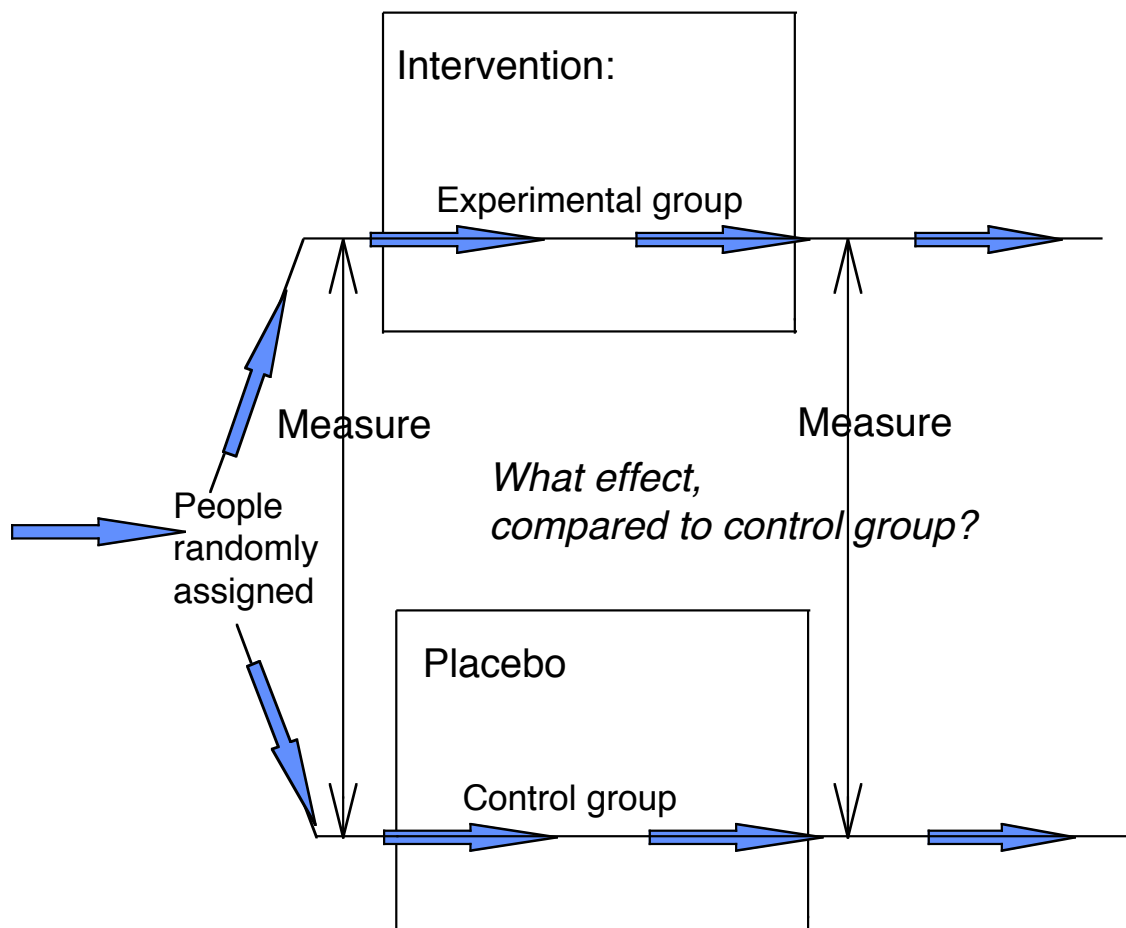
Unlike type 3 designs this design compares two interventions. It is a common design for evaluations of services or health programmes, for example where a new service is started at one site which people want to compare it with a traditional service at another site. An example is a comparison of the health state of patients before and after a new service for stroke patients compared to conventional rehabilitation. Evaluations with these designs are often carried out according to experimental principles (“controlled trials”), with hypotheses to test and with different methods for controlling for influences other than the interventions, such as patient characteristics (eg age, sex, severity of illness, duration of previous illness etc). A common control technique is “matching” the characteristics of the people experiencing each intervention, which is what was done in the example evaluation. The aim of matching is to try to exclude influences which might affect the outcome other than the intervention, but matching is less good for this purpose than random allocation (see design 5 below).

The design, like the type 3 design, can also be used when conditions allow a “natural experiment”. An example is to make a retrospective evaluation where there are records or measures already available. Comparative outcome evaluations of this type can give some objective evidence of the effect of the intervention compared to another one, but they take longer and use more resources than type 3 designs. The cost and time required increases with more subjects (which calls for more time and care matching subjects), with more complex measures (eg quality of life measures), with more than one “after” measure, and with measurement some time after the intervention.

The difference between type 4 and the type 3 design is that two interventions are compared, and the differences to type 5 which we consider next are that people are not randomly allocated. Also in type 4, one of the interventions is not a placebo: the controls are therefore fewer than for the type 5 full experimental design.

4: TYPE 5 - RANDOMISED CONTROLLED TRIAL

This design is not suitable for evaluations of most types of health programme, policy or reforms. It is described here to show the design which is most well known and respected in medical research and is the standard against which many other designs are judged in healthcare, often inappropriately.



The idea behind this design is to create two groups which are exactly the same in all respects apart from the fact that one (target) group received the intervention. The design can rule out many alternative explanations for a change which is detected in the targets, and can give evidence of causal mechanisms. The purpose of the design is to compare the effects of an intervention on one group compared to another group which does not get the intervention, but who are in all other possible respects similar. The questions the design answers are, what are the effects of the intervention, compared to a control group? It is used for an experimental and economic evaluation to gain evidence of probable the effect of a treatment or service on one or a few measures of the health state of a group of patients.

This design is the “classic” evaluation design which many think of for a “proper” evaluation of a treatment or a service. The design is like the type 4 design, but the people selected for the intervention are randomly assigned to a control (placebo) and an intervention group. If people are randomly assigned to the two groups then statistical methods can be used to compare the outcomes and assess if the difference between the groups is greater than would be expected by chance alone. We can predict using statistics that by chance alone a certain number of people may show that the intervention has a sizable effect, for example anything up to 5% may show this effect. We want to be sure that the difference in effect is significantly greater than what we would expect by chance. This requires not only randomisation, but a large enough number of subjects to be entered to the experiment to be sure that any differences are statistically significant.

The design is able to reduce the number of possible explanations for any differences between the outcomes for the two groups being due to things other than the intervention. Ideally the second control group do not “get nothing”, but receive an intervention called a “placebo” which is as similar as possible but without the “active ingredient” (One estimate is that 30-40% of patients will show an improvement without any intervention). Also they do not know which group they are in (the subjects are blind) and the service providers do not know who is in which group (provider blinding). Such double blinding is not always possible.

The weaknesses of the design are that it is expensive, takes time, needs many carefully selected patients, and need evaluators with experience, skill and statistical expertise to produce credible results. Often these designs do not examine patient’s subjective experience, and may lose important effects for a few individuals in the group average. The design has been criticized as not allowing generalization of findings to normal settings because the subjects are carefully selected (to maximize controls) which means that the subjects are not typical of those in normal settings - the design maximizes “internal” validity at the expense of “external” validity. Nine other criticisms are listed in Øvretveit (1998a).

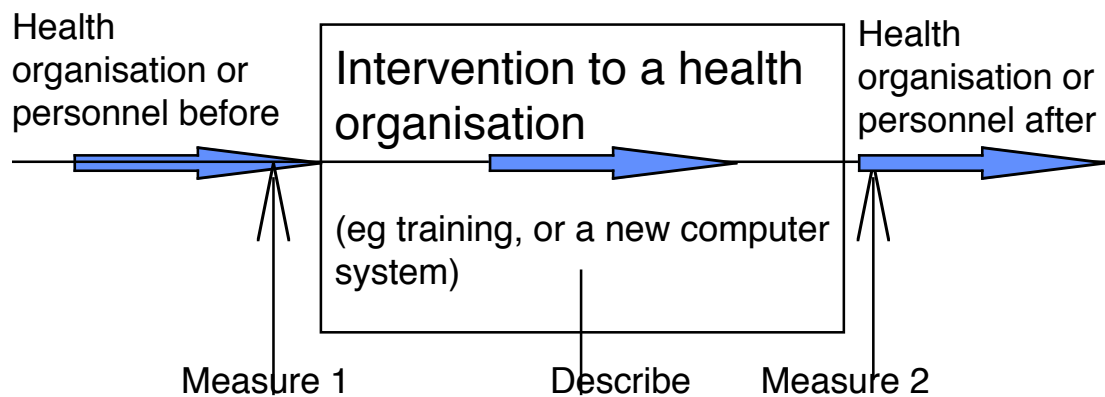
The strengths are that the design gives more reliable and valid information about the effect of the intervention than type 3 simple outcome evaluation, and a type 4 non-random, non-placebo comparison. The results from a well conducted evaluation using this design has high credibility with most clinicians. In principle the design can be retrospective, but, when looking back over time, it is usually difficult to get true randomisation or a good control group or to check if the intervention was stable. Type 5 designs are prospective and called “experimental” because the evaluator intervenes to create a change and then studies it using similar principles and controls to those used by a natural scientist in a laboratory.

Some of the principles can be applied in programme or reform evaluations, such as randomly assigning providers or units rather than patients to an “intervention” and a control group. But this may not be possible, neither may “blinding” of subjects. The full design is not suitable for evaluating many health interventions because of the difficulties of controlling for extraneous variables, and or of controlling the intervention itself. For some interventions - such as many diagnostic or treatment interventions - there is no doubt that it can be the best design for answering certain questions, if the resources and time are available.

“People criticize the randomized controlled trial. But if they had cancer and did not know which treatment to take, or whether to take any, it only evidence from such trials that they would use to make their decisions. They would not stake their lives on evidence from less powerful designs “

“If I had cancer and had to choose a treatment I would want to know what the treatment experience is like and what other patients thought about the treatment process. There is no point in living for five years more if the treatment is so awful you cannot live properly.”

5: TYPE 6A – INTERVENTION TO A HEALTH ORGANISATION – IMPACT ON PERSONNEL

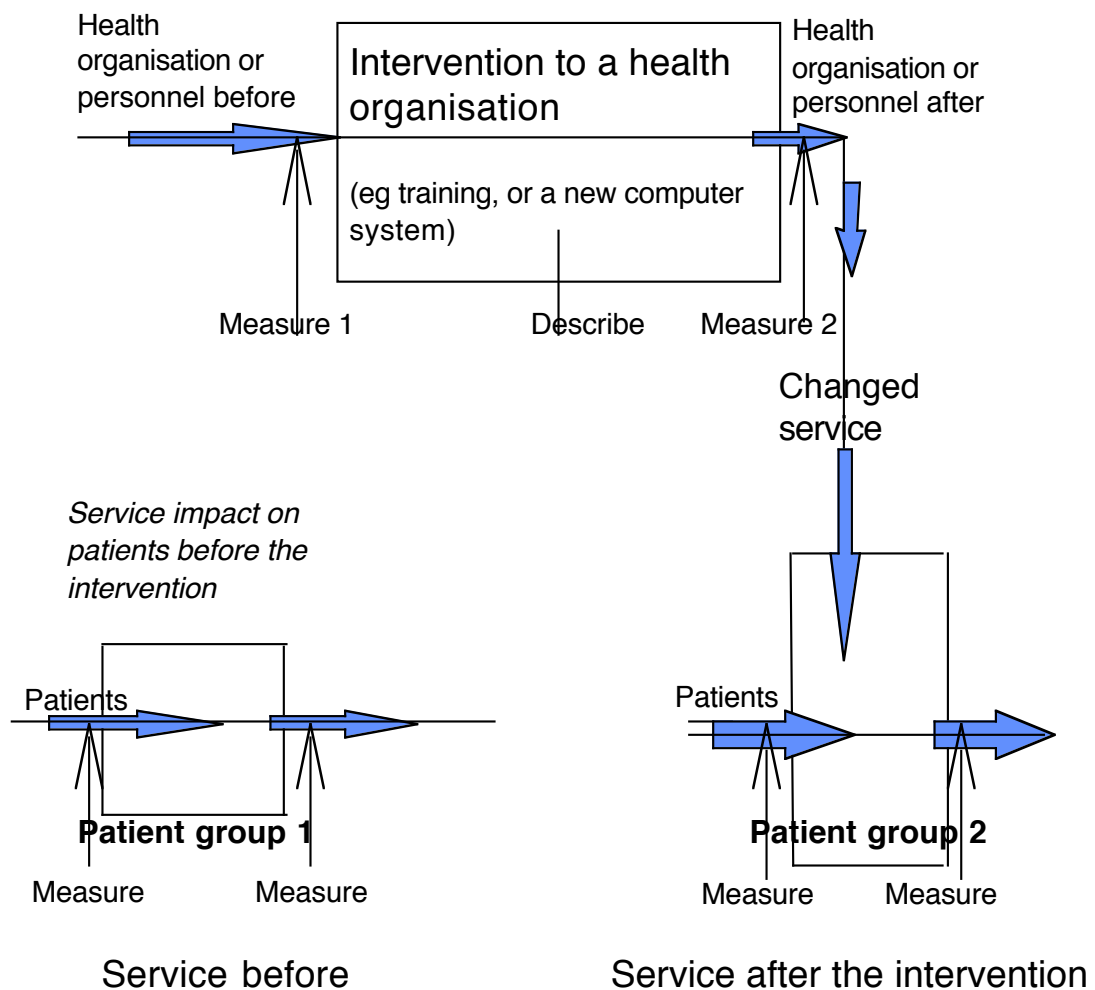


The two type 6 designs consider the effect of an intervention on a health organisation. Examples are evaluations of a training programme for physicians, of an intranet system, of a new nursing protocol, or of a new financing system. The Type 6a design answers the question, “what is the effect of an intervention to a service on health personnel or organisation?” It is similar to the before-after type 3 design. The purpose of this design is to judge the value of an intervention to a service by comparing the state of health personnel or organizational functioning before, with the state after the intervention.

The data to be gathered before and after depend on the user’s questions and future decisions, the objectives of the intervention, and any theories about how the intervention may affect personnel or organisation. Common methods are surveys, observation, and measures such as personnel work stress. Before and after data could also be about resources, as in an economic evaluation of the effect of an intervention, such as new technology, on the service costs.

The weaknesses are, as with type 3, that it is difficult to be sure that any effects on health personnel or organizational functioning which are detected are only due to the intervention and not to other influences. One variation of this design to use a comparison organisation to control for some factors. Another weakness for some purposes is that it does not consider the effect on patients - if a change to health personnel or organisation is detected, how does this change affect patients? The strengths are that it is cheap, quick and usually better than nothing, so long as the limitations are spelled out. It is more useful when another evaluation has already discovered how the intervention or its outcome for providers affects patient care.

6: TYPE 6B- INTERVENTION TO A SERVICE - IMPACT ON PATIENTS



This design looks at the effect of an intervention to organisation on patients, as well as the effect on health personnel or organizational functioning. By comparing patient outcomes before and after the change, the design can help to judge the value of a change to organisation or of other interventions to a service. It is used for evaluations of training programmes, of quality assurance and of other interventions intended to improve patient care.

The two patient groups before and after need to be carefully matched to increase certainty that any difference in outcome between the two groups is due to the intervention and not due to the characteristics of the patients in the two groups. Variations of this design involve comparing two or more services receiving the intervention and the effects on the patients of these services.

The weaknesses of this design are that it is difficult to match or control for patient characteristics between the before and after patient groups. Also the measures chosen to study the effect of the service on patients might not detect important benefits or disbenefits produced by the intervention. The strengths are that it does give a method for assessing how changes to health organisations affect patient care, whether or not this is the primary aim of the change or

intervention. This design and variations of it allow an evaluation of the impact on patients and on health personnel.

7: SUMMARY

- Outcome evaluation designs are used to find out if an intervention or action has any effects. Usually the design measures the state of target people or organisations before and after the intervention to see if it makes a difference.
- The simple “before-after” type 3 design is low cost and easy to arrange, but it cannot give conclusive evidence of effects. If a before-after difference is found, we cannot be sure if the difference was caused by things other than the intervention.
- The “comparative” type 4 design compares the before-after difference in two groups of people or organisations, one of which receive the test intervention and the other receives a traditional or comparable intervention. If both groups are similar and both interventions take place in a similar environment, then it is likely that any difference in the before after states of the groups is due to the interventions alone.
- The “randomized controlled trial” type 5 design allocates patients or other subjects randomly to the test intervention and to a placebo which is the same as the test but without the active ingredient. By using statistics, any difference between the two groups which is greater than chance is significant if confounders are controlled for.
- The type 6a design is used to evaluate the effect of an intervention such as a training programme or health reform on health providers or organizational functioning. The type 6b design is used when the effects of such interventions on patients or populations is also of interest.