NSW Health

www.health.nsw.gov.au

EVIDENCE AND EVALUATION GUIDANCE SERIES

Study Design for Evaluating Population Health and Health Service Interventions

A Guide

NSW Ministry of Health
1 Reserve Road
ST LEONARDS NSW 2065
Tel. (02) 9391 9000
Fax. (02) 9391 9101
TTY. (02) 9391 9900
www.health.nsw.gov.au

Produced by:
Centre for Epidemiology and Evidence

The NSW Ministry for Health acknowledges the traditional custodians of the lands across NSW. We acknowledge that we live and work on Aboriginal lands. We pay our respects to Elders past and present and to all Aboriginal people.

Further copies of this document can be downloaded from the NSW Health website www.health.nsw.gov.au

SHPN (CEE) 230996
ISBN 978-1-76023-717-2

December 2023

Contributors to the development of this guide:
NSW Ministry of Health
Alexandra Schiavuzzi
Amanda Jayakody
Ashleigh Armanasco
Aaron Cashmore
Andrew Milat

Prevention Research Collaboration, University of Sydney
Adrian Bauman

Suggested citation:
Centre for Epidemiology and Evidence. *Study Design for Evaluating Population Health and Health Service Interventions: A Guide*. Evidence and Evaluation Guidance Series, Population and Public Health Division. Sydney: NSW Ministry of Health; 2023.

Contents

1. Executive summary

Appropriate selection of study designs is vital to the production of high quality evaluations

Choosing an appropriate study design and executing it well can produce credible evidence of intervention or program effectiveness. The purpose of this guide is to assist NSW Health staff in the planning and design of evaluations. This guide includes information to assist with the selection of quantitative study designs. It includes consideration of the quality and credibility of different designs, as well as pragmatic considerations for conducting research in healthcare and population settings.

A well-planned study design is a critical first step in evaluation

Planning should include developing a program logic model to assist in determining the mechanism by which the intervention contributes to change, providing a structure upon which to develop an evaluation. Planning should also involve considering good practice principles (e.g. ethical conduct and stakeholder involvement) and generating evaluation questions which unpack the key issues you hope to address in your evaluation. Pragmatic considerations include the stage of implementation of the intervention, the setting, feasibility, acceptability and integrity of the study design, and availability of resources for the evaluation in relation to costs, time and sample size required.

Experimental designs provide the strongest evidence of causality

There are many different study designs which may be used to answer different questions, depending on the stage of implementation of the intervention and evaluation. Experimental designs offer the most rigorous way of determining whether a cause-effect relationship exists between an intervention and an outcome. They include randomised controlled trials (RCTs), cluster randomised controlled trials, and stepped-wedge and multiple baseline designs. These are the preferred designs for health service interventions but may not be a practical approach for evaluating population-based programs.

Quasi-experimental and non-experimental designs offer a pragmatic alternative

Quasi-experimental designs offer a compromise between research integrity and pragmatic considerations, particularly when it is not possible to randomise individuals to intervention or control groups. These designs include controlled before and after and interrupted time series designs. Quasi-experimental designs attempt to demonstrate causality between an intervention and an outcome when random assignment has not occurred. Non-experimental studies may be used when it is not feasible or ethical to conduct a true experiment, for example when an intervention is already underway or where harm would be caused by withholding an intervention. These studies include cohort, repeat cross-sectional and single group pre-post test and post-program designs. Alone, they cannot be used to demonstrate cause and effect.

2. Introduction

The *NSW Treasury Policy and Guidelines: Evaluation (TPG22-22)*[1] sets out mandatory requirements, recommendations and guidance for NSW General Government Sector agencies and other government entities to plan for and conduct the evaluation of policies, projects, regulations and programs.

NSW Health is committed to the generation and use of research and evaluation to improve policy and program effectiveness. NSW Health is committed to the evaluation of population health and health service interventions in order to develop evidence-based policies and programs. This guide will support NSW Health staff in the planning of evaluations of interventions using appropriate study designs.

Study design (also referred to as research design) refers to the different study types used in research and evaluation. In the context of an impact/outcome evaluation, study design is the approach used to systematically investigate the effects of an intervention or a program.

Study designs may be experimental, quasi-experimental or non-experimental. Study design is distinct from the study methods (e.g. structured interviews) and the study instruments (e.g. structured interview questions). The primary purpose of a good study design is to enable us to be as confident as possible that any observed changes were caused by the intervention, rather than by chance or other unknown factors.[2,3] The design should match the scale of the program and the significance of the evaluation, and be as rigorous as possible while meeting pragmatic needs of the real-world context of health interventions.[2] This guide focuses on quantitative study designs used in impact/outcome evaluations and is relevant to simple, complicated and complex population health interventions and health service interventions.

The guide begins with an overview of initial steps needed in choosing an appropriate study design and pragmatic considerations for evaluation of population health and health service interventions. The second section outlines the strengths and weaknesses of quantitative study designs, namely experimental, quasi-experimental and non-experimental designs. It describes when it may be appropriate to use each design in practice, and presents case studies of evaluations of population health and health service interventions.

Evaluation is the systematic and objective process used to make judgements about the merit or worth of a program, usually in relation to its effectiveness, efficiency and appropriateness.[2] Comprehensive program evaluations should integrate process, impact, outcome and economic evaluation, with all components planned at the same time as the development of the intervention. For further information, refer to *Planing and Managing Program Evaluations: A Guide*. Process evaluations assess how a program/intervention is progressing in terms of how it is delivered and to whom, whether the program is being implemented as planned and the level of program quality. Impact or outcome evaluation measures the immediate and long term effects, or unintended effects, of a program as defined in the program logic, and is the primary focus of this guide.[3,4] Economic evaluations are a comparative analysis of the cost-effectiveness or cost-benefit of a program. Please refer to *Engaging an Independent Evaluator for Economic Evaluations: A Guide*.

This guide is primarily intended to be used by NSW Health staff who plan and implement, commission, provide strategic oversight of, or use results of evaluations of population health and health service interventions. It has been developed for a non-specialist audience with key definitions and a resource and further reading list provided.

What makes an intervention simple, complicated or complex?

Evaluations of population health and health service interventions are rarely straightforward. There is increasing recognition that interventions in population health and health services need to address factors operating at the individual, social and system levels which require multifactorial approaches to effectively target the complexity of health and health behaviour.[5] For example, an intervention promoting healthy eating in a defined region may include a mass media campaign, nutrition education in schools, and a program to introduce healthy eating choices in workplaces. A study design to evaluate such an intervention must be able to accommodate its complexity and scope.[3]

Even so, some health interventions can still be defined as simple, in the sense that there is a single intervention being tested in a small group of individuals (e.g. an educational booklet to increase health literacy among teenagers with diabetes). There is a simple linear pathway linking the intervention to its outcome. Complicated interventions may have a number of interrelated components targeting individuals or groups and there is a high degree of certainty that the intervention can be repeated. Complex interventions, on the other hand, are usually defined as interventions that contain several interacting components, targeting multiple problems or designed to influence a range of groups, with a degree of flexibility or tailoring of the intervention permitted.[3,6] This guide broadly applies to all these types of interventions but, given the population and health services context, consideration is mostly given to both complicated and complex interventions. For a more in-depth guide to complex program evaluation please refer to *Evaluation in a Nutshell* by Bauman and Nutbeam.[3]

3. Planning an evaluation

A well-planned and conducted study design is critical to the overall credibility and utility of the evaluation.[2] Before choosing a study design, however, it is important to consider a number of key steps.

3.1 Program logic

An important first stage in planning any evaluation is developing or reviewing a program logic model, even if the program is already being delivered. A program logic model is a schematic representation that describes how a program is intended to work by linking activities with outputs, intermediate impacts and longer-term outcomes. This will assist in determining the mechanism by which the intervention causes change, providing a structure upon which to develop an evaluation, and enabling potential identification and strengthening of causal links. For further information on developing program logic models, please see *Developing and Using Program Logic: A Guide* and/or view this short animation on program logic.

3.2 Good practice principles for evaluation

Good practice principles should be incorporated into the planning and conduct of high quality evaluations, including:

- **Timeliness** – Evaluation planning should ideally be conducted during the program planning phase. Particular consideration should be given to the realistic amount of time needed to conduct an evaluation to ensure findings will be available when needed to support decision making.
- **Appropriateness** – The scope of the evaluation should be realistic and appropriate with respect to the size, stage and characteristics of the program being evaluated, the evaluation budget and availability of data.
- **Stakeholder involvement** – The participation of stakeholders in the planning, conduct and interpretation of findings of program evaluations will increase the likelihood of the evaluation influencing policy and practice.
- **Effective governance** – An advisory group with clear roles and responsibilities should be established to guide and inform the evaluation process.
- **Methodological rigour** – Evaluations should use appropriate study designs and methods, and draw on relevant instruments and data that are valid and reliable, ensuring the design is appropriate to the purpose and scope of the evaluation.
- **Consideration of specific populations** – Consideration of the health context and the needs of different population groups, such as Aboriginal populations, is essential. Engagement with identified specific populations is important throughout the duration of the evaluation.
- **Ethical conduct** – Evaluations should be conducted in an ethical manner that considers legislative obligations, particularly the privacy of participants, and costs and benefits to the individuals or population involved.

These principles are described in more detail in *Planning and Managing Program Evaluations: A Guide*.

Table 1. Example outcome evaluation questions

Example question	Example indicator
Have smoking rates decreased in program participants?	Proportion of daily smoking from January 2018 to July 2018 among program participants
To what extent can increased physical activity be attributed to the intervention?	Mean number of sessions of moderate physical activity among intervention participants compared to control group

3.3 Generating evaluation questions

The purpose of the evaluation, and what questions it is intended to answer, will help determine the design of an evaluation.[7] These questions are not survey or interview questions but high level evaluation questions.[2] The evaluation questions need to unpack the key issues that you hope to address in your evaluation. Specifically, the evaluation questions need to include the population in which you wish to observe the change, including a clear control or comparison group if required, and the type of change you expect to see. Indicators which will help answer your evaluation questions must then be selected to ensure they are specific to the evaluation you are conducting, and measurable. This may require an assessment of the data available (e.g. program monitoring data, administrative data) or to generate the data required to answer your evaluation questions (e.g. direct measurement, interview and questionnaires). These aspects of the research must be well established before moving on to the selection of a study design. It is important to keep in mind that more complex interventions may require multiple evaluation questions to be formulated and may need to be changed to suit the practical realities of the situation.[3] For further information on generating evaluation questions, refer to *Planning and Managing Program Evaluations: A Guide* and the Sax Institute's *Translational Research Framework*.

3.4 When to use quantitative methods

In choosing a study design for an evaluation, it is important to understand when you will need quantitative methods. Quantitative methods are used in evaluation for a number of reasons:[3]

- differences or change in an impact or outcome need to be quantified
- validated and reliable measures are needed to answer an evaluation question
- causal evidence of program effects is needed (keeping in mind that the program is rarely the sole cause of change; there may be other activities or environmental factors which provide partial attribution)
- data are needed on a large number of people or populations.

3.5 A note on qualitative methods

Using qualitative methods in your evaluation will depend on your research questions. Other than being used for the formative or developmental stages of an evaluation or for understanding implementation processes (process evaluation), qualitative methods are commonly combined with quantitative methods in a mixed methods evaluation. A mixed methods evaluation allows for the triangulation of both quantitative and qualitative findings, which can strengthen confidence in the findings and provide a deeper understanding of unique contexts.[8,9] Mixed methods are particularly important when programs are more complex and require a multi-pronged evaluation.[3] The more consistent the evidence, the more reasonable it is to assume that the program has produced the observed results.[3] Qualitative approaches use methods such as focus groups, in-depth interviews or observation to analyse and explore complexity, meaning, relationships and patterns.[3] High quality qualitative research involves rigorous methods such as a clear research question, data collection, data analysis and interpretation.[3] For further information on how rigour can be established in qualitative methods please refer to *Qualitative Research Methods* by Liamputtong. Although qualitative methods are important for evaluation, this guide focuses on quantitative study designs. For further information on how to use qualitative and mixed methods in your evaluation refer to the NSW Government *Evaluation Toolkit for Government Agencies*.

3.6 Pragmatic considerations

While scientific rigour, including prudent study design selection, is an important aspect of a high quality evaluation there are also pragmatic considerations to take into account when designing evaluations. Along with a program logic model, the good practice principles and the generation of relevant evaluation questions outlined above, selecting the appropriate design for an evaluation also requires careful consideration of the following factors:[6,10]

- implementation stage of the intervention (see the *Translational Research Framework*)
- setting of the evaluation (e.g. school, hospital, whole population)
- the likelihood of systematic bias (e.g. selection bias)

- availability of data (e.g. program monitoring data)
- budget for and cost of the evaluation
- feasibility of conducting the study design
- acceptability of the study design to stakeholders and participants.

In practice, some compromises may be needed. For example a trade-off between rigour and budget constraints may result in choosing the next best alternative design.[2] Study designs 'lower down' in the hierarchy (see section 4.1) can still produce useful results, but findings need to be interpreted in light of the limitations of the design used.[6]

A good rule of thumb when designing an evaluation is to 'keep it simple'. For examle, it is not necessary to use a complex design when a simple one will suffice. However, more complex interventions may require larger evaluation sample sizes in order to account for additional expected variability and longer study periods, especially if system changes are required. Using a range of outcome measures allows for more efficient use of the data produced by complex interventions. This will enhance the understanding of the practical effectiveness of the program, including the types and causes of variation between individuals, sites and over time, allowing for an assessment of how viable the intervention is in a real-world setting. Finally, more complex interventions are more sensitive to the impacts of local contextual modification. They may present difficulties with adhering to standardised evaluation designs and delivery. Research protocols may therefore need to allow for adaptation to the local setting.[6]

Overall, your chosen study design needs to be fit for purpose, ensuring it is realistic and appropriate with respect to purpose, rigour and context. Focusing on the most relevant evaluation questions will help ensure your evaluation is manageable, cost efficient and useful.

4. Study designs

There are different types of study designs used in evaluation. Different study designs may be appropriate at different stages of implementation of the intervention. This guide focuses on quantitative designs that are used in evaluating the impacts and outcomes of population health and health service interventions.

4.1 Experimental designs

The quality of research designs is often framed around a hierarchy of the 'most scientific' designs to the 'least scientific' designs.[11] Experimental designs are considered to be at the top of the hierarchy as they provide the most compelling evidence that an intervention caused a particular effect.[8] For further reading on levels of evidence see the National Health and Medical Research Council's *How to Use the Evidence*. It is also important to note that methodological quality (how a study is implemented) affects the credibility of study findings. For example, a well-designed and delivered quasi-experimental study may provide more credible evidence than a poorly executed experimental study. For guidance on methodological quality criteria please refer to *Grading of Recommendations Assessment, Development and Evaluation (GRADE)*.

Experiments are characterised by the establishment of two or more groups that are identical except for a single factor of interest, for example, exposure to an intervention. Any observed differences between the groups can hence be attributed to that factor.[10] True experiments are characterised by randomisation of intervention and experimental control groups.[7]

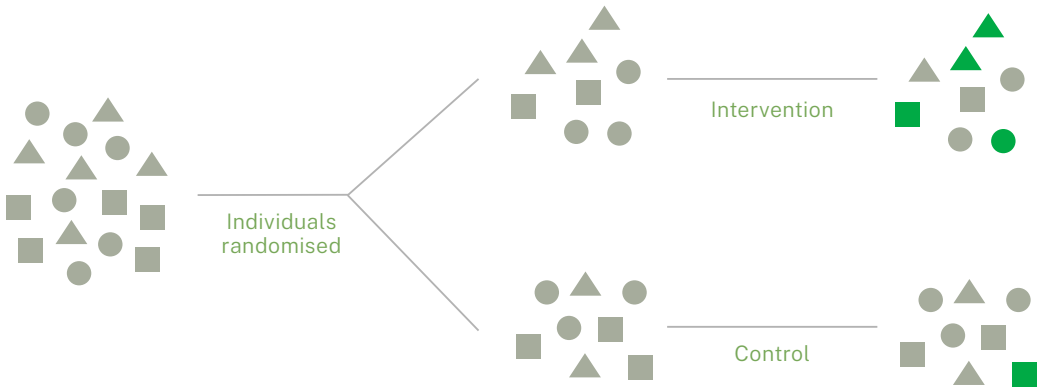
An important consideration in choosing an appropriate study design for an evaluation is that prospective studies (including experimental studies) are not plausible if the intervention has already started, unless an enhancement of the program is being tested.

4.1.1 Randomised controlled trials (RCTs)

Description

In an RCT, individuals are randomly assigned to either a control group or an intervention group at the start of the study (see Figure 1). This ensures that individual factors that may influence the outcomes (including those factors that are not known) will be evenly distributed across the two groups. Theoretically, randomisation allows the groups to be as similar as possible at the outset of the study, though the likelihood of this increases with increasing sample size.[10] Assessment of demographic, other important characteristics of the two groups and the outcome measures should be made at baseline, before the intervention begins and repeated after the intervention has been delivered.[11] Randomisation reduces selection bias so that the differences in the outcomes between the two groups can be attributed to the different treatment of the groups (intervention or control), and not some other confounding factor, effect modifier or due to chance. Again, an adequate sample size is required to ensure that a difference between the two groups is able to be detected. In evaluations of health service interventions, the control group typically receives the usual care or no treatment (rather than receiving a placebo, as in many clinical interventions). It is because of the similarity of the two groups at baseline that RCTs are considered to provide the best evidence of cause and effect of all study designs, as differences between the groups at the study end-point can be attributed to the intervention itself.

Figure 1. Randomised controlled trial



Note: Green shapes represent individuals with a change in the outcome of interest at follow up

Strengths and limitations

RCTs are usually used for simple interventions. They are considered the best way to evaluate the effectiveness of a new treatment or prevention strategy, providing the most compelling evidence of all study designs (particularly if several RCTs are combined in a systematic review).[10] RCTs are the best study design with regard to internal validity, hence they are most relevant when there is a need to generate 'causal evidence'.[7,11] Internal validity refers to the extent to which differences in observed effects between exposed and unexposed groups can be attributed to the intervention and not to some other possible cause. RCTs sometimes lack external validity; generalisations of findings outside the study population may be invalid.[12,13] RCTs are costly to design and implement as they require more control over the program setting than may be achievable.[8] They may be subject to practical problems such as loss to follow-up or an inability to blind participants as to which study group they are in. More complex interventions may require flexible delivery and multiple community-driven strategies; hence, adherence to a strict RCT protocol when delivering and evaluating complex public health interventions is often not pragmatic.[14] This may hamper the implementation and evaluation of the program to the point where no effect can be observed. One of the biggest challenges to maintaining the fidelity of an RCT is the potential for contamination of the control group. This occurs when those in the control group know about the intervention, and this therefore influences their behaviour, making it difficult to detect the effects of an intervention.[11] Randomisation may also involve ethical risks, such as withholding a program intended to improve health outcomes from one group where the intervention is known to be effective in other settings or a different population. This may be overcome through the use of other experimental designs outlined below.

When to use RCTs in practice

You can use RCTs for your evaluation only if it is possible to randomly allocate individuals to intervention and control groups. Given the costs involved in maintaining fidelity to a program protocol required by RCTs, you should only use this design in a well-funded project. You should use RCTs to test a causal hypothesis only after you have used simpler and cheaper designs to determine the feasibility of your intervention.[14] Keep in mind that RCTs are not often used in population health and health services research because they are best used for well-defined (discrete) interventions and controllable settings.

CASE STUDY: RANDOMISED CONTROLLED TRIAL

Effects of a pedometer-based intervention on physical activity levels after cardiac rehabilitation

This case study is an example of an evaluation where it was possible to randomly allocate individuals to receive a pedometer-based intervention or to be in a control group. This type of intervention allowed for a well-targeted and controlled setting. The RCT was conducted to evaluate the efficacy of pedometers for increasing physical activity after a cardiac rehabilitation program (CRP). Patients (n=110) who had attended a CRP were randomised into an intervention or a control group. The six-week intervention included self-monitored physical activity using a pedometer and step calendar and two behavioural counselling and goal-setting sessions. The control group received two generic physical activity information brochures after the baseline questionnaire was administered. Self-reported physical activity and psychosocial status were collected at baseline, six weeks, and six months. At six weeks and six months, improvements in total physical activity sessions (six weeks: change in mean sessions (SD)=2.9 (6.5), p=0.002; six months: change in mean sessions (SD)=0.9 (5.8), p=0.016), walking minutes (six weeks only: change in mean minutes (SD)=80.7 (219.8), p=0.013), and walking sessions (six weeks: change in mean sessions (SD)=2.3 (5.5), p=0.001; six months: change in mean sessions (SD)=0.2 (5.0), p=0.035) in the intervention group were significantly greater than those in the control group after adjusting for baseline differences.

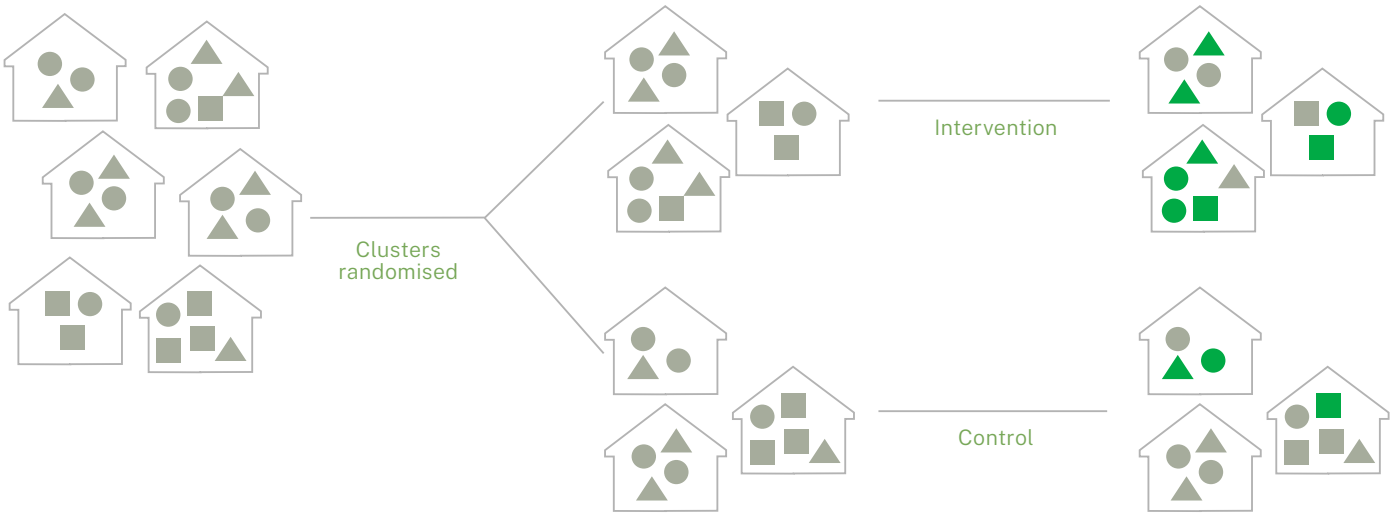
Butler L, Furber S, Phongsavan P, Mark A, Bauman A. Effects of a pedometer-based intervention on physical activity levels after cardiac rehabilitation. J Cardiopulm Rehabil Prev 2009; 29(2): 105-14.

4.1.2 Cluster randomised controlled trials

Description

Cluster randomised controlled trials (cluster RCTs) operate on the same principles as RCTs, however randomisation occurs at the group-level rather than the individual-level (see Figure 2).[5] These groups (clusters) are randomly allocated to intervention or control conditions. This is important where individuals in a group share features in common which may have an impact on the study outcome (e.g. students attending the same school are likely to share behaviours or beliefs that may influence their health). As in a traditional RCT design, randomisation of the intervention and control group ensures that these groups are as similar as possible at baseline so that any effect observed may be attributed to the intervention itself.[5]

Figure 2. Cluster randomised controlled trial



Note: Green shapes represent individuals with a change in the outcome of interest at follow up

Strengths and limitations

Cluster RCTs are often the most appropriate experimental study design when there is a risk of contamination if individuals are randomised, if the intervention is best targeted to a group or population, or if it is more logistically feasible to be delivered at a group-level.[5] Analysis of groups rather than individuals requires more technical statistical expertise given that individuals within groups or clusters are likely to be correlated (and standard statistical methods assume that observations are independent). If this correlation is not accounted for, it may result in a false positive result (Type I error).[5] In cluster RCTs, two sources of variability must be taken into account in analyses: the variation between individuals within groups, and the variation of individuals between groups. It is therefore important to ensure that an adequate number of clusters are used, and that there are an appropriate number of individuals within each cluster group. Careful calculation of sample size, taking into account the intracluster correlation coefficient, is required when designing cluster RCTs. Using stratified randomisation can assist in achieving an appropriate balance between groups when there are a small number of clusters.

It may be difficult to ascertain if observed changes are attributable to the intervention if clusters are selected from the same site. This is because there may be other interventions (e.g. introduced policies, mass media campaigns) or influences that occur within that community at the same time as the intervention.[5] It is therefore important to choose the right cluster unit. For example, children may influence each other within a school, meaning classes are not an appropriate cluster and groups should be clustered at the school-level.

Cluster RCTs are costly as they require more control over the program setting than may be achievable. Cluster RCTs are also ethically complex (e.g. using whole cultural groups as the cluster-unit), and consideration needs to be given to the different ethical needs of different cluster groups.[5] There are also ethical implications of withholding an intervention from the control groups, particularly when evaluating an intervention in vulnerable groups.

When to use cluster RCTs in practice

If your intervention is targeted towards a group rather than individuals, or is based on a modification to the environment that might have an impact on a whole group, you can consider using a cluster RCT for your evaluation.[5,11] However you should bear in mind that cluster RCTs generally require a large budget to design and conduct, and often require more control over the program setting than you may be able to achieve in practice.

CASE STUDY: CLUSTER RANDOMISED CONTROLLED TRIAL

A cluster randomised controlled trial of a telephone-based intervention targeting the home food environment of preschoolers (The Healthy Habits Trial): the effect on parent fruit and vegetable consumption

This is an example of an evaluation where groups (preschools) were randomly allocated to intervention or control conditions. The unit of randomisation was the preschools. This ensured the fidelity of the intervention was not compromised because parents within the preschools may have shared their beliefs which may have influenced their health behaviours. The intervention was delivered between April and December 2010, targeting the home food environment of preschool children (i.e. the fruit and vegetable consumption of parents). In 2010, 394 parents of children aged 3–5 years from 30 preschools in the Hunter region of New South Wales were recruited and randomly assigned to an intervention or control group. Intervention group parents received four weekly 30-minute telephone calls and written resources. The scripted calls focused on: fruit and vegetable availability and accessibility, parental role-modelling, and supportive home food routines. Two items from the Australian National Nutrition Survey were used to assess the average number of serves of fruit and vegetables consumed each day by parents at baseline, and 2-, 6-, 12-and 18-months later, using generalised estimating equations (adjusted for baseline values and clustering by preschool) and an intention-to-treat approach. At each follow-up, vegetable consumption among intervention parents significantly exceeded that of controls. At 2-months the difference was 0.71 serves (95% Confidence Interval (CI): 0.58-0.85, $p < 0.0001$), and at 18-months the difference was 0.36 serves (95% CI: 0.10-0.61, p=0.0067). Fruit consumption among intervention parents was found to significantly exceed consumption of control parents at the 2-,12-and 18-month follow-up, with the difference at 2-months being 0.26 serves (95% CI: 0.12-0.40, p=0.0003), and 0.26 serves maintained at 18-months, (95% CI: 0.10-0.43, p=0.0015). A four-contact telephone-based intervention that focuses on changing characteristics of preschoolers' home food environment can increase parents' fruit and vegetable consumption.

Wyse R, Campbell KJ, Brennan L, Wolfenden L. A cluster randomised controlled trial of a telephone-based intervention targeting the home food environment of preschoolers (The Healthy Habits Trial): the effect on parent fruit and vegetable consumption. Int J Behav Nutr Phys Act 2014; 11: 144.

4.1.3 Stepped-wedge and multiple baseline designs

Description

A stepped-wedge design is a modified cluster RCT where the intervention is implemented sequentially in clusters (or sites).[15] All clusters are in the control phase at the start of the study and by the end of the study all clusters are in the intervention phase (see Figure 3). The order in which the different clusters receive the intervention is randomised, and the intervention may be implemented in multiple clusters at the same time. Outcomes are measured prior to implementation of the intervention in any of the sites and at the end of each implementation period before implementation in the next site.[5]

As with stepped-wedge design, multiple baseline design involves implementing the intervention in multiple sites using a phased approach, where the timing of the intervention implementation is randomised.[5] Some consider there to be no real difference between a multiple baseline and stepped-wedge design as they both involve observations pre-and post-intervention in all sites or groups.[5] However, unlike a stepped-wedge design where multiple individual measurements can be taken at each time point, a multiple baseline design usually involves a single measurement at each time point and at more frequent intervals both before and after intervention implementation (see Figure 4).[5] Another difference is that the intervention is implemented at different times in multiple sites, while stepped-wedge designs can be implemented in multiple sites at the same time.[5] With a multiple baseline, multiple data points should be collected prior to the intervention being implemented in any site. The purpose of this is to ensure that a stable baseline is achieved. When there is an effect (determined by assessing both change in slope, from pre-to post-implementation, and change in intercept) in the site where the intervention had been implemented, but not in other sites at the same point in time, the effect can be attributed to the intervention itself and not some other cause. The repeated observation of an effect before and after the implementation of the intervention in each site, and the absence of substantial fluctuations in measurements before and after the intervention, strengthens the conclusion that the effect is attributable to the intervention itself.[16]

Figure 3. Stepped-wedge design

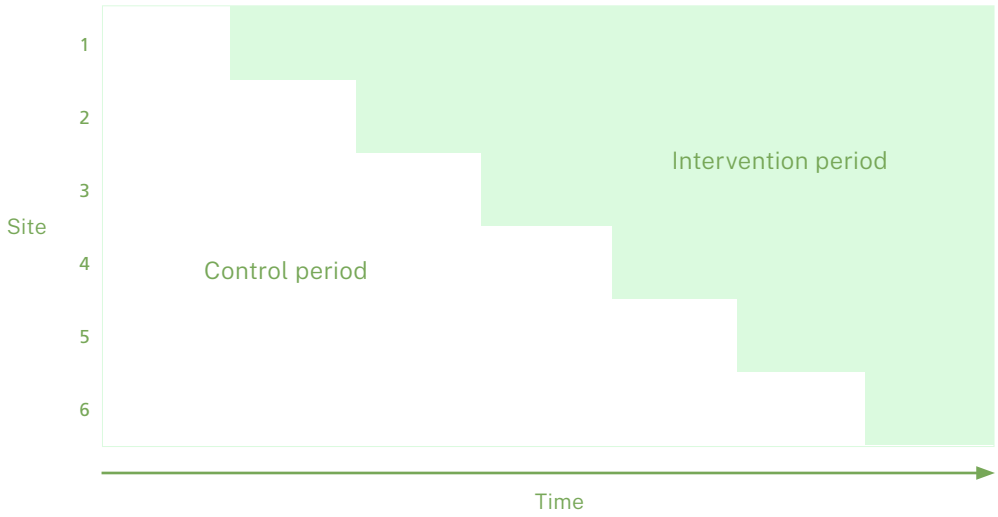
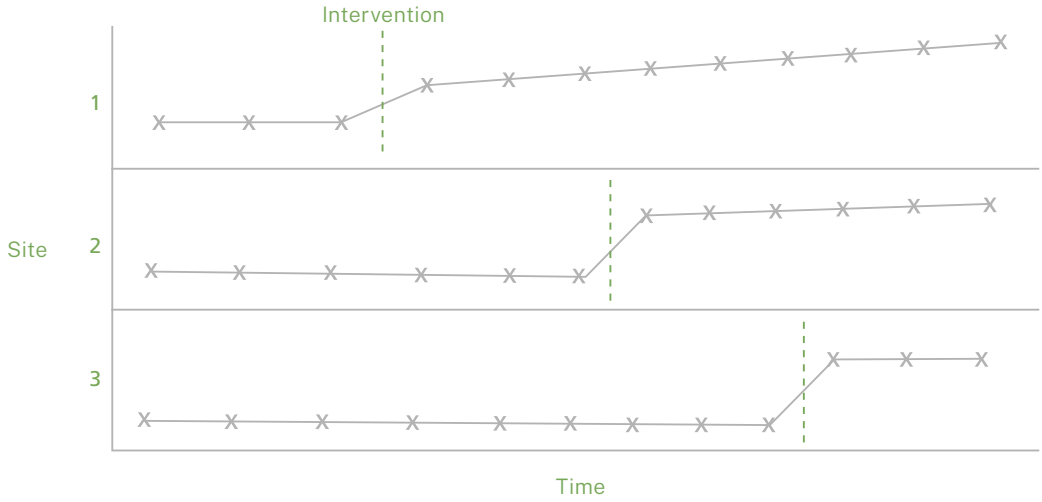


Figure 4. Multiple baseline design



Strengths and limitations

In both stepped-wedge and multiple baseline designs, the order in which the intervention is delivered to different clusters is randomised, reducing selection bias. These designs also minimise the ethical risks of RCTs and cluster RCTs as all groups eventually receive the intervention. Staggering of the intervention and measurements also offers greater efficiency than standard RCTs and cluster RCTs, by spreading the use of resources over time, which improves the feasibility of these designs. They also allow for analysis both between groups and within groups over time.

However, both stepped-wedge and multiple baseline designs require more measurements over a longer period of time than traditional cluster RCTs, and for multiple baseline, even when there is no active intervention. In particular, for multiple baseline designs where outcomes are self-reported, the burden on the participant can be great and could result in greater dropout rates. They are also relatively complex to design and analyse, and it may be challenging to determine the optimal number of clusters, individuals per cluster, measurements required, and the most appropriate methods of analysis. Variation in the cluster size may reduce the efficiency of this design. In the context of health services research, multiple baseline and stepped-wedge designs can be challenging to implement as intended, as health services are not always ready to start implementation at the allocated date due to a host of factors including competing priorities.

When to use stepped-wedge and multiple baseline designs in practice

You can use a stepped-wedge design if it is important all your groups receive the intervention (rather than half the groups, as in a cluster RCT). For example, a health intervention can be rolled out in a staggered way across a whole health system. You could use this design if you have a high demand for a policy or a program of unknown effectiveness but insufficient resources to roll the program out universally at one time. This allows local demand for the program to be met without compromising the rigour of your evaluation.[17] A stepped-wedge design is used more in health service interventions rather than population health interventions.

The same applies to multiple baseline studies. You should consider a multiple baseline design in particular if you have a single outcome of interest, it is feasible to collect data at multiple regular intervals, and you are able to ensure the intervention occurs at different times in each site. You can also use a multiple baseline design if you are only able to attain a small sample size (i.e. only a small number of sites will participate in the evaluation).[16]

CASE STUDY: STEPPED-WEDGE DESIGN

Increasing the provision of preventive care by community healthcare services: a stepped wedge implementation trial

This is an example of where the investigators were able to evaluate a complex community healthcare intervention in all facilities in a local health district (LHD) in New South Wales without the need for a site to miss out on the potentially effective intervention. The study aimed to determine the effectiveness of an intervention in increasing community-based clinician implementation of multiple elements of recommended preventive care for four risk behaviours. A three-group stepped-wedge trial was undertaken with 56 community-based primary healthcare facilities in one LHD. A 12-month implementation intervention was delivered sequentially in each of three geographically and administratively defined groups of facilities. The intervention consisted of six key strategies: leadership and consensus processes; enabling systems; educational meetings and training; audit and feedback; practice change support; and practice change information and resources. Client-reported receipt of three elements of preventive care: assessment; brief advice; and referral for four behavioural risks (smoking, inadequate fruit and/or vegetable consumption, alcohol overconsumption, and physical inactivity), individually and for all such risks combined was collected for 56 months in total. Significant increases were found for receipt of four of five assessment outcomes (smoking Odds Ratio (OR)=1.53; fruit and/or vegetable intake OR=2.18; alcohol consumption OR=1.69; all risks combined OR=1.78) and two of five brief advice outcomes (fruit and/or vegetable intake OR=2.05 and alcohol consumption OR=2.64). No significant increases in care delivery were observed for referral for any risk behaviour, or for physical inactivity.

Wiggers J, McElwaine K, Freund M, Campbell L, Bowman J, Wye P, et al. Increasing the provision of preventive care by community healthcare services: a stepped wedge implementation trial. Implement Sci 2017; 12(1): 105.

CASE STUDY: MULTIPLE BASELINE DESIGN

Chronic Care Service Enhancements Program Evaluation

This study is an example of where an alternative to a cluster RCT design was needed due to the importance of all clusters receiving the intervention, but it was not logistically possible for the investigators to deliver the intervention simultaneously. This study aimed to assess the impact of a variety of interventions on improving rates of chronic disease screening and management of diabetes in NSW Aboriginal Community Controlled Health Services (ACCHSs). The intervention included: staff training; standardised care through the use of health assessment templates and diabetes assessment templates; recall/reminder systems; clinical audit and feedback; and problem solving via ACCHSs consulting with one another on strategies to achieve set targets and increase adherence. The study involved the sequential implementation of the intervention across ACCHS sites (five ACCHSs for screening and six ACCHSs for diabetes management). The order of implementation was planned to be randomised across the sites however due to practical circumstances the order of commencing the intervention was chosen by some ACCHSs themselves. Due to time constraints, services were paired, with the implementation of interventions for each pair staggered over three months. It is distinctive from a stepped-wedge design as it used count data, more data collection points and interrupted time series analysis. Outcomes of interest were health screens and specific tests related to diabetes. At monthly intervals throughout the study, the proportion of active participants receiving a specific test or having a target result was recorded. Segmented logistic regression was used to test the hypothesis that the intervention would increase the proportion of patients receiving important health screens or having biochemical test results within a goal range.

Overall, the study found significant variation across ACCHSs for both preventive health screening rates and management of diabetes. Combining all five screening sites, interventions were associated with a statistically significant increase in the proportion of patients being screened for diabetes using random blood glucose measurements, but no change for fasting blood glucose. Combining all six diabetes management sites, no changes in patient health outcomes in terms of achieving recommended goals for optimum diabetes management were observed.

Health Behaviour Research Group. Chronic Care Service Enhancements Program evaluation: summary report. University of Newcastle; 2016.

4.2 Quasi-experimental designs

Many population health and health service interventions are incompatible with study designs that randomly assign individuals or locations to intervention or control groups (e.g. evaluations at the policy or system-level).[5,6]

It is often not possible to randomise due to:

- difficulty of randomising participants
- difficulty of randomising by location
- small available sample size
- policy timing
- limited resources/funding.

For these interventions, a quasi-experimental design is usually the most appropriate choice whereby it is possible to construct and apply quasi-experimental study designs that measure the effects of the intervention where random assignment has not occurred but rather has been pre-determined.[11] Similar to true experiments, they aim to demonstrate causality between an intervention and an outcome, compared to a control or comparison group* who did not receive an intervention.[11,18] These designs may address one or more of the threats to internal validity (but not all of them).[8]

4.2.1 Controlled before and after design

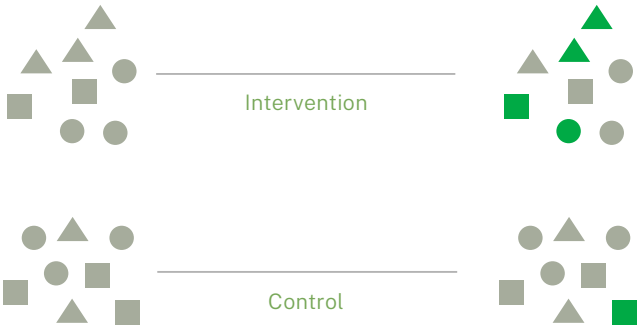
Description

Controlled before and after studies (also known as controlled pre-post) measure observations before and after the implementation of an intervention and compare them to a control group who did not receive the intervention (see Figure 5). The main difference between this design and an RCT is that there is no random assignment to treatment groups. A matched control group is used for comparison purposes and helps to ascribe the effect to the intervention itself. This provides information on whether an intervention is responsible for the change observed.

Strengths and limitations

Controlled before and after studies are relatively simple to undertake and offer a practical design for the evaluation of population health and health service interventions. They may be used for as few as two groups, which may make them relatively less expensive than some other designs.[16] As controlled before and after designs use single before and after points of data collection, it can be difficult to determine, or control for, the impact of other influences or events on the outcome, which may not be evenly distributed between the groups.[16] The confidence with which one can say that the results are due to the intervention, rather than temporal trends or other external factors, is therefore limited.

Figure 5. Controlled before and after design



Note: Green shapes represent individuals with a change in the outcome of interest at follow up

When to use controlled before and after designs in practice

Controlled before and after designs are common. They can be used to explore the effects of an intervention over time as measures can be compared over time within a group and between groups. This design may be useful if you need to provide an initial estimate of change, for example in pilot studies, or if other more robust designs are not possible.

* A control group (or treatment/experimental group) generally refers to a group created through random assignment who do not receive an intervention, or receive the usual program compared to a newly developed intervention. For quasi-experimental and non-experimental designs, groups are not randomly assigned. These groups can be referred to as control groups, but they are more generally referred to as comparison groups. You may also hear them being referred to as non-exposed groups. For further information on these groups, please see the BetterEvaluation website or the *NSW Government Evaluation Toolkit for Government Agencies*.

4.2.2 Interrupted time series design

Description

An interrupted time series design measures the effect of an intervention in one group, site or cluster over a period of time.[5] A single group can be compared over time using many repeated measures, and with the baseline as the control comparison.[16] A time series is fit pre-and post-intervention and the effect of the intervention is measured by changes in the slope or level (intercept) of the outcome of interest from the pre-to the post-intervention period (see Figure 6). A change in the level suggests a change in the outcome following the intervention, whereas a change in slope indicates a different trend following the intervention.[5] The interrupted time series design is considered quasi-experimental because there is no random allocation of the intervention.[5] The strength of this approach can be enhanced by using a comparison group, for example, one or more comparison regions which also have trend data.

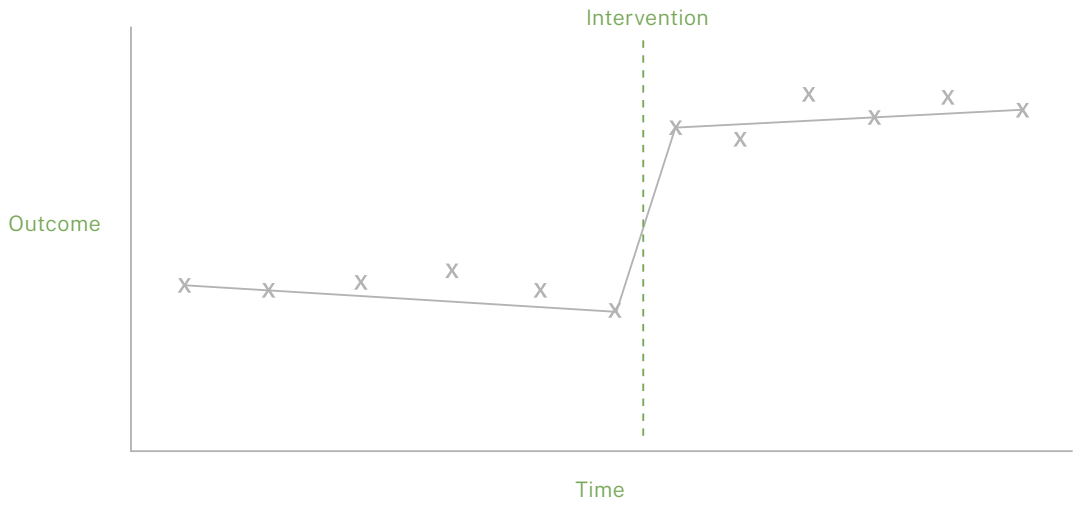
Figure 6 demonstrates how an interrupted time series design works. A comparison can be made with a multiple baseline design which uses multiple sites (see Figure 4).

Strengths and limitations

Since this study design only requires one group or cluster, it is relatively simple and cost-effective to implement.[5] There is also no need to allocate clusters to a control condition, which further simplifies the study design. The major limitation of this study design is that it is not possible to determine conclusively whether the changes observed are due to the intervention itself or other factors.[5]

Methodological issues include determining the number of observations (time periods) required and the duration of the time period measurement.[5]

Figure 6. Interrupted time series design



When to use interrupted time series designs in practice

Time series design is more commonly used in health service interventions, for example in evaluations of screening or immunisation programs. You should use an interrupted time series design if your intervention is being delivered to a single group, site or cluster, and/or when you have sources of data that are routinely collected by health authorities, such as hospital admission records.[11] You may also find this design useful if you are evaluating a policy innovation. This is because it allows for structured observations of change in a population where an intervention has already been introduced, with little consideration of the evaluation.[11] This design is also useful for evaluating social marketing campaigns where you may find it very difficult to have a comparison group.

CASE STUDY: INTERRUPTED TIME SERIES DESIGN

Alcopops, taxation and harm: a segmented time series analysis of emergency department presentations

This case study is an example of using an interrupted time series design (also known as segmented) to evaluate the impact of two government taxes on alcohol-related emergency presentations using emergency department (ED) records. The evaluation estimated the change in incidence of emergency department presentations for acute alcohol problems associated with the 2000 Goods and Services Tax (GST) and the 2008 ready-to-drink ('alcopops') tax. Segmented regression analyses were performed on age-and sex-specific time series of monthly presentation rates for acute alcohol problems to 39 hospital EDs across New South Wales over 15 years (1997–2011). Indicator variables represented the introduction of each tax. Retail liquor turnover controlled for large-scale economic factors such as the global financial crisis that may have influenced demand. The GST was associated with a statistically significant increase in ED presentations for acute alcohol problems among females aged 18–24 years (0.14/100 000/month, 95% CI 0.05-0.22). The subsequent alcopops tax was associated with a statistically significant decrease in males aged 15–50 years, and females aged 15–65 years, particularly in females aged 18–24 years (-0.37/100 000/month, 95% CI -0.45 to -0.29). An increase in retail turnover of liquor was positively and statistically significantly associated with ED presentations for acute alcohol problems across all age and sex strata.

Gale M, Muscatello DJ, Dinh M, Byrnes J, Shakeshaft A, Hayen A, et al. Alcopops, taxation and harm: a segmented time series analysis of emergency department presentations. BMC Public Health 2015; 15(1): 468.

4.3 Non-experimental designs

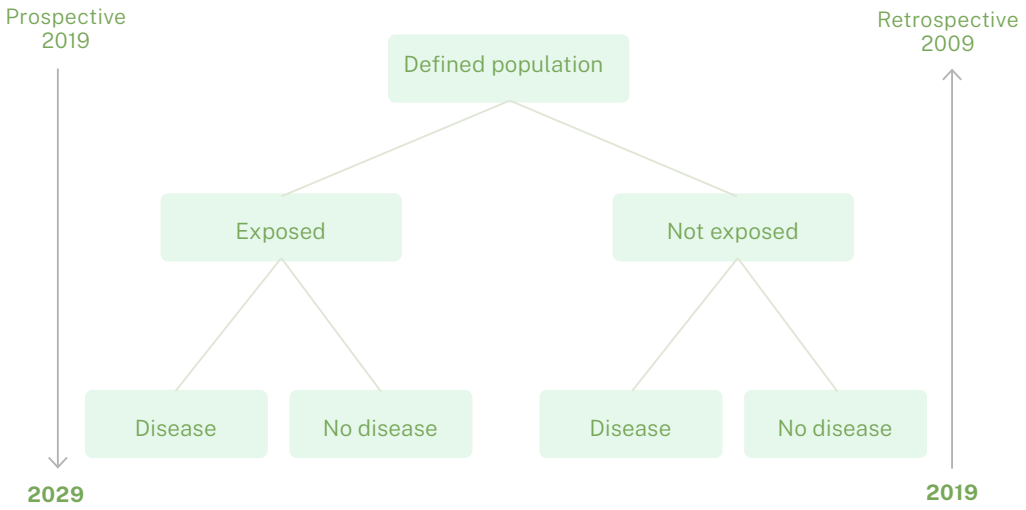
A number of study designs that can be used in evaluations are described as non-experimental, observational or pre-experimental. Unlike experimental designs, they do not include a control group whereby participants or groups are randomly assigned to receive the intervention or usual care, nor is there randomisation in terms of sampling.[7] They are most useful in evaluating population health and health service interventions when a whole community or whole state is exposed to the intervention (and therefore there is no unexposed group), or when it is ethically unreasonable, or not feasible for practical purposes (e.g. a lack of time), to conduct a true experiment.[7,14] These study designs are applicable for describing a health issue and the context in which it is occurring, and may help to illustrate an association between certain factors and a health issue. It is difficult to demonstrate cause-and-effect from these studies, hence they provide the weakest evidence of all the study designs.[11] They should be used only after other design possibilities have been considered.

4.3.1 Retrospective and prospective cohort studies

Description

A cohort study is a non-experimental study which starts with the selection of a study population (cohort). A cohort is any group of individuals sharing a common characteristic such as age, ethnicity, or exposure to a factor of influence, presence of the disease of interest or receipt of an intervention. A cohort study may be prospective or retrospective and sometimes more than two groups are compared. A prospective (concurrent) cohort study recruits participants at the beginning of the study and follows them into the future.[7] Prospective cohort studies follow a group of similar subjects over time who differ with respect to certain factors under study. For example, one could group subjects based on their body mass index and compare their risk of developing heart disease or cancer in the future. Retrospective (historical) cohort studies are carried out at the present time and examine existing data consisting of observations about the cohort (collected in the past over a period of time) to determine the incidence of disease or other health outcome in the cohort and factor/s of influence. This type of study might consist of chart data or public records. Figure 7 depicts the timeframes for a hypothetical prospective cohort study and a hypothetical retrospective cohort study.

Figure 7. Retrospective and prospective cohort studies



Strengths and limitations

Cohort studies represent the most comprehensive of the non-experimental study designs as they can enable the long-term effects of population-level interventions and policies in a real-world setting to be evaluated. As retrospective cohort studies use existing data, they offer efficiency in terms of time and cost as most of the evaluation resources are directed towards data analysis, rather than data collection.

Cohort studies typically require a large sample size and long follow-up period in order to determine the associations between exposures and outcomes.[7] Retrospective cohort studies rely on previously collected datasets to provide this. Because the data have been collected in the past, there is little control over the data collection methods so data may be incomplete, inaccurate or inconsistently measured between subjects. It is important when using administrative data to formulate hypotheses *a priori* (when planning the study) so that the data are not being 'dredged' looking for associations. Since the allocation to groups is not controlled by the researcher, it is likely that the groups will differ in ways other than the exposure or outcome of interest.[7] This needs to be accounted for in the data analysis of these study types, and data may or may not be available to do this adequately.

Prospective cohort studies are both expensive and time-consuming. A comparison of the outcome between groups can only take place after enough time has elapsed so that some subjects have developed the outcomes of interest.

When to use retrospective and prospective cohort studies in practice

You should use a retrospective cohort design for your evaluation if you have previously collected relevant (administrative) data and your intervention has already been rolled out. You can also use a retrospective cohort design if you need a cost-effective and efficient way to generate hypotheses for more rigorous testing using other study designs.

You can use an existing prospective cohort for your evaluation such as the 45 and Up Study. For example, a randomised controlled trial of a web-based intervention to improve mood, cognitive function and adherence in people with cardiovascular disease (CVD) recruited their eligible participants from the 45 and Up Study prospective cohort.[19] Self-reported history of CVD or CVD risk factors and psychological distress were analysed from the 45 and Up Study baseline dataset.

CASE STUDY: RETROSPECTIVE COHORT STUDY

The impact of telephone follow up on adverse events for Aboriginal people with chronic disease in New South Wales, Australia: a retrospective cohort study

This case study is an example of where a program had already been rolled out across services over a number of years and there was program monitoring data in place. The evaluation used a retrospective cohort design using program monitoring and administrative data to assess the effectiveness of the intervention. The intervention was a NSW telephone follow-up service, 48 Hour Follow Up, for Aboriginal people recently discharged from hospital. The health outcomes of interest were unplanned hospital readmissions, ED presentations, mortality within 28 days of discharge from hospital and at least one adverse event (composite outcome). A data analysis was conducted of the 48 Hour Follow Up Program Register, a public health register comprising linked data from the following five sources: 48 Hour Follow Up Program Dataset (all records of patients identified by LHDs as eligible for 48 Hour Follow Up); NSW Admitted Patient Data Collection; NSW Registry of Births, Deaths and Marriages Death Registrations; Chronic Disease Management Program Minimum Dataset; and NSW Emergency Department Data Collection. The data sources were linked by the Centre for Health Record Linkage. Results found no significant difference in 28-day unplanned hospital readmission and 28-day deaths for those receiving 48 Hour Follow Up, compared to eligible patients who did not receive follow up. However, there was evidence that patients who received 48 Hour Follow Up were significantly less likely to experience an unplanned ED presentation (OR=0.92; 95% CI: 0.85-0.99; p=0.0312) and at least one adverse event (OR=0.91; 95% CI: 0.85-0.98; p=0.0136) within 28 days of discharge compared to eligible patients who did not receive follow up.

Jayakody A, Passmore E, Oldmeadow C, Bryant J, Carey M, Simons E, et al. The impact of telephone follow up on adverse events for Aboriginal people with chronic disease in New South Wales, Australia: a retrospective cohort study. Int J Equity Health 2018; 17: 60.

4.3.2 Repeat cross-sectional studies

Description

In cross-sectional studies, information about the exposure/s and outcome/s of interest are collected simultaneously from a representative sample of individuals within a defined population. This provides a snapshot of the frequency and characteristics of a health problem in the population at a particular point in time.[20] A repeat cross-sectional study (also referred to as a serial cross-sectional) involves the same or similar information being asked of a different (independent) sample of individuals (from the target population) at different time points. Estimates of changes can be made over time at the aggregate or population level.

Strengths and limitations

Repeat cross-sectional studies are simple to conduct as they do not require assignment of participants to an intervention or control group. It is important to ensure that the sample of subjects selected is representative of the whole population to whom the results are being extrapolated. Random sampling techniques (as distinct from random allocation to intervention and control groups) may help to minimise selection bias. While repeat cross-sectional studies are a feasible design for many health system evaluations, they are not as methodologically strong as cohort studies (which follow the same individuals over time) for explaining how and why observed changes occurred.[11]

When to use repeat cross-sectional studies in practice

You can use a repeat cross-sectional design if you need to estimate prevalence, distribution and determinants of common conditions of relatively long duration, such as the health effects of being a smoker. You can also use this design to estimate the prevalence of acute or chronic conditions in the population at repeat points in time. Repeat cross-sectional studies are often used in population health, for example, in evaluating the effectiveness of a smoking campaign at the population level.

CASE STUDY: REPEAT CROSS-SECTIONAL STUDY

Impact of Australia's introduction of tobacco plain packs on adult smokers' pack-related perceptions and responses: results from a continuous tracking survey

This case study describes the use of a repeat cross-sectional design used in evaluating the introduction of Australia's tobacco plain packaging legislation in October 2012. This study aimed to evaluate the impact of Australia's plain tobacco packaging policy on two stated purposes of the legislation — increasing the impact of health warnings and decreasing the promotional appeal of packaging — among adult smokers. Weekly telephone surveys were conducted from April 2006 to May 2013 with 15,745 adult smokers in NSW. Random selection of participants involved recruiting households using random digit dialling and selecting a smoker for interview. The main outcomes of the study were salience of tobacco pack health warnings, cognitive and emotional responses to warnings, avoidance of warnings and perceptions regarding one's cigarette pack. Analyses found, 2–3 months after the introduction of the new packs, a significant increase in the absolute proportion of smokers having strong cognitive (9.8% increase, $p=0.005$), emotional (8.6% increase, $p=0.01$) and avoidant (9.8% increase, $p<0.001$) responses to onpack health warnings. There was a significant increase in the proportion of smokers strongly disagreeing that the look of their cigarette pack is attractive (57.5% increase, $p<0.0001$), says something good about them (54.5% increase, $p<0.0001$), influences the brand they buy (40.6% increase, $p<0.0001$), makes their pack stand out (55.6% increase, $p<0.0001$), is fashionable (44.7% increase, $p<0.0001$) and matches their style (48.1% increase, $p<0.0001$). Changes in these outcomes were maintained 6 months post-intervention.

Dunlop SM, Dobbins T, Young JM, Perez D, Currow DC. Impact of Australia's introduction of tobacco plain packs on adult smokers' pack-related perceptions and responses: results from a continuous tracking survey. BMJ Open 2014; 4(12): e005836.

4.3.3 Single group pre-post test and post-program only

Description

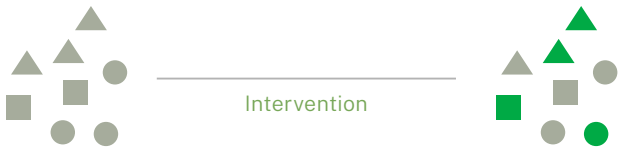
A single group pre-post test design (also known as a prepost one group, before and after study or case-series design) is commonly used in evaluation. It uses a single group and measures the change in health outcome or behaviour in this group before and after an intervention (see Figure 8). This study design is useful for determining whether a change has occurred after the implementation of an intervention. In this design, each individual is his or her own control. It is important to note that data can be collected at multiple points (see the *Evaluation of the NSW Knockout Health Challenge case study*). A post-program only evaluation is where participants are only assessed after the program has finished.[11]

Strengths and limitations

The major flaw of a single group pre-post test design is the absence of a control group, which makes it difficult to ascribe the outcome to the intervention itself. There is potential for temporal (over time) trends and extraneous factors to influence the outcomes observed and this design offers limited ability to assess the impact of these factors.[16] Therefore any observed effects cannot be ascribed to the intervention itself, and the intervention can only be said to be associated with the outcomes observed. This design can be strengthened by including multiple data collection points which increases the methodological rigour.

A post-program only evaluation may be used for process evaluations, including participants' assessment of their experience of a particular intervention component, however the risk of social desirability bias is high.[11] A post-program only evaluation should never be used for assessing the effectiveness of a program as it is not possible to measure any changes due to the intervention and therefore causal inference is not possible.

Figure 8. Single group pre-post test design



Note: Green shapes represent individuals with a change in the outcome of interest at follow up

When to use pre-post test and post-program studies in practice

You may find that a single group design is the only design possible for your type of intervention. For example, you may be evaluating a mass media campaign where the program has broad reach so it is difficult to obtain a control group for comparison purposes. A single group design offers greater simplicity in design and analysis, and is a relatively cost-effective method, however attribution of the outcome to the intervention is difficult.

Importantly, a post-program only design is not useful for assessing the effectiveness of a program and cannot be used for any causal inferences.

CASE STUDY: SINGLE GROUP PRE-POST TEST DESIGN

The impact of a community-led program promoting weight loss and healthy living in Aboriginal communities: the New South Wales Knockout Health Challenge

A single group pre-post design was used to evaluate an Aboriginal community-based weight-loss challenge (the NSW Knockout Health Challenge). In 2013, the Challenge involved Aboriginal community-based teams from across NSW competing in a 16-week weight loss Challenge, followed by a 12-week maintenance period. A total of 586 people participated in 22 teams across NSW. Data on weight, diet and physical activity levels were collected at four time points: at the start of the Challenge (via paper form, n=576); end of the Challenge (via paper form, n=377); five months after the Challenge (via telephone interviews (n=271) and paper form (n=76)); and nine months after the Challenge (via telephone interview, n=195). Among participants who provided data at all time points (n=122), there was a significant mean weight loss from the start to the end of the Challenge (2.3kg, 95% CI -3.0 to -1.9, p<0.001) and from the start to nine months after the Challenge (2.3kg, 95% CI -3.3 to -1.3, p<0.001). By the end of the Challenge, participants reported they were more physically active and had increased their fruit and vegetable consumption compared with the start of the Challenge.

Passmore E, Shepherd B, Milat A, Maher L, Hennessey K, Havrlant R, et al. The impact of a community-led program promoting weight loss and healthy living in Aboriginal communities: the New South Wales Knockout Health Challenge. BMC Public Health 2017; 17(1): 951.

5. Key resources and further reading

For further information on evaluation and study design, please see the resources listed below (the reference list might also be of interest). When you are deciding on the most appropriate study design for your evaluation it is recommended you seek further advice from data, research or evaluation specialists.

Evidence and Evaluation Guidance Series, Population and Public Health Division

Developing and Using Program Logic: A Guide and *Program Logic animation*

Planning and Managing Program Evaluations: A Guide

Engaging an Independent Evaluator for Economic Evaluations: A Guide

Increasing the Scale of Population Health Interventions: A Guide

Further reading about evaluation design and methods

NSW Government Evaluation Toolkit for Government Agencies

Translational Research Framework, Sax Institute

Translational Research Framework: Source Book, Sax Institute

BetterEvaluation

Bauman A, Nutbeam D. *Evaluation in a nutshell: a practical guide to the evaluation of health promotion programs*. 2nd ed. Sydney: McGraw-Hill; 2014.

Liamputtong P. *Qualitative Research Methods*. 4th ed. Melbourne: Oxford University Press; 2013.

Further reading about reviewing and appraising evidence

National Health and Medical Research Council, Identifying the Evidence

Grading of Recommendations Assessment, Development and Evaluation (GRADE)

Guidelines for reporting evaluations

CONSORT statement for the reporting of Randomised Controlled Trials

Standards for Reporting Implementation Studies (StaRI) Statement

TREND statement for reporting evaluations using non-randomised designs

STROBE statement for reporting observational studies

SPIRIT statement for the reporting of study protocols

6. Key definitions

Complex interventions are usually defined as interventions that contain several interacting components, targeting multiple problems/behaviours, and/or designed to influence a range of groups.[6]

Confidence intervals are a statistical term referring to the range of values within which the true value in the population is likely to lie.

Confounding refers to where a variable may influence the association between two other variables and thereby confound the results in a study.[3]

Contamination refers to the extent to which participants in control groups might be exposed to the intervention.

Evaluation is the systematic and objective process used to make judgements about the merit or worth of a program, usually in relation to its effectiveness, efficiency and appropriateness.[2]

External validity refers to the extent to which findings from an intervention are generalisable to all potential recipients.

False positive or Type I error are terms used in statistical hypothesis testing. A type I error is the incorrect rejection of a true null hypothesis (concluding a program works when it doesn't), while a type II error is incorrectly retaining a false null hypothesis (concluding a program doesn't work when it does). A type III error is correctly rejecting the null hypothesis but for the wrong reason.

Health service intervention is a multidisciplinary activity with the objective of improving the health services people receive.

Internal validity refers to the extent to which differences in observed effects between exposed and unexposed groups can be attributed to the intervention and not to some other possible cause. Examples of threats to internal validity could be another current event coincided with the intervention which affects the outcome, or those in the control group find out about, or interact with the intervention group.

Intracluster correlation coefficient is used particularly in the design of cluster RCTs. It is a statistical measure of the relatedness of clustered data. It accounts for the relatedness of clustered data by comparing the variance within clusters with the variance between clusters.[21]

Levels of evidence refers to a ranking system to describe a research study, whereby certain study designs produce more valid and reliable results than others. Level I evidence is attributed to a systematic review or meta analyses of RCTs, Level II evidence is attributed to an individual RCT, followed by quasi-experimental and so on. For further information regarding the levels and quality of evidence refer to National Health and Medical Research Council, *Identifying the Evidence*.

Matched control group is where non-participants (individuals, organisations or communities) are selected so that they are similar to the participants in certain characteristics that are thought to be relevant, such as age, sex, socioeconomic status and occupation.[10]

Population health intervention refers to any action that directly or indirectly addresses a health issue in the population as a whole, or in a particular subgroup within the population, with the aim of producing change or identifiable outcomes.[14]

Protocol refers to a document that describes your study design and its methods. It normally includes: the background, rationale, objectives, design, methodology and planned analyses.

Qualitative research involves using methods such as focus groups, in-depth interviews or participant observation to analyse and explore complexity, meaning, relationships and patterns. This method is commonly used in process evaluation (in understanding how a program has been implemented) and in interpreting findings from quantitative studies.[3]

Quantitative research involves using statistical approaches and is based on quantifiable measurements of phenomena such as physical, behavioural, psychological, social and environmental factors. This is particularly useful for determining the impacts and outcome (effectiveness) of a program.[3] Examples of common data sources for evaluation are self-report survey, structured interview, BMI measurements, and administrative data.

Randomisation refers, firstly, to the sampling method where each individual from a population has an equal probability of being chosen to be included in a study. Secondly, it refers to where individuals or groups are randomly allocated to receive the intervention or not.

Sample refers to a subset of the population that may be used to assess the impact of an intervention, ideally a sample that is representative of the population of interest so that you can extrapolate to the broader population with confidence. Individuals or groups can be chosen via a number of different sampling methods (e.g. random sample, purposive sampling or convenience sampling).

Sample size is calculated using statistical formulas to work out how many people are needed for a study to be confident that the results are likely to be true. It is necessary to pre-specify what quantitative change is expected for the intervention.[3] Complex interventions may require larger sample sizes in order to account for additional expected variability across multiple outcomes.

Selection bias refers to the way in which people are selected into a study and whether the selected people are representative of the target population.[3]

Stratified randomisation is a variation to simple randomisation, whereby the population is divided into smaller groups called strata. Strata are based on members' shared characteristics. Random samples are then selected from each stratum.

Triangulation refers to comparing information obtained from different methods of data collection (e.g. comparing results from a survey with what participants discussed in a focus group).[7]

7. References

1. NSW Treasury. *NSW Treasury Policy and Guidelines: Evaluation*. Sydney: NSW Treasury; 2023. Available from: www.treasury.nsw.gov.au/finance-resource/evaluation-policy-and-guidelines
2. NSW Department of Premier and Cabinet. *Evaluation Toolkit for Government Agencies*. Sydney: NSW Department of Premier and Cabinet. Available from: www.nsw.gov.au/departments-and-agencies/premiers-department/evaluation-toolkit
3. Bauman A, Nutbeam D. *Evaluation in a nutshell: a practical guide to the evaluation of health promotion programs*. 2nd ed. Sydney, Australia: McGraw-Hill; 2014.
4. Centre for Epidemiology and Evidence. *Planning and Managing Program Evaluations: A Guide*. Evidence and Evaluation Guidance Series, Population and Public Health Division. Sydney: NSW Ministry of Health; 2023. Available from: www.health.nsw.gov.au/research/Publications/planning-evaluations.pdf
5. Sanson-Fisher RW, D'Este CA, Carey ML, Noble N, Paul CL. Evaluation of systems-oriented public health interventions: alternative research designs. *Annu Rev Public Health* 2014; 35: 9-27.
6. Craig P, Dieppe P, Macintyre S, Michie S, Nazareth I, Petticrew M. Developing and evaluating complex interventions: the new Medical Research Council guidance. *Int J Nurs Stud* 2013; 50(5): 587-92.
7. Wludyka P. *Study Designs and their Outcomes*. In: Macha K and McDonough JP, editors. Epidemiology for Advanced Nursing Practice. Sudbury, MA: Jones and Bartlett Learning; 2012. p 81-114.
8. McDavid JC, Huse I, Hawthorn LRL. *Program Evaluation and Performance Measurement: An Introduction to Practice*. 2nd ed. California: Sage Publications; 2012.
9. Bamberger M. *Introduction to mixed methods in impact evaluation*. Impact Evaluation Notes (No. 3.). Interaction, 2012. Available from: www.interaction.org/wp-content/uploads/2019/03/Mixed-Methods-in-Impact-Evaluation-English.pdf
10. Webb P, Bain C. *Essential Epidemiology: An introduction for Students and Health Professionals*. 2nd ed. Cambridge: Cambridge University Press; 2011.
11. Sax Institute. *Translational Research Framework: Source Book*. Sydney: Sax Institute; 2016. Available from: www.saxinstitute.org.au/wp-content/uploads/Translation-Research-Framework_Source-Book.pdf
12. Rothwell PM. External validity of randomised controlled trials: "to whom do the results of this trial apply?" *Lancet* 2005; 365(9453): 82-93.
13. Bothwell LE, Greene JA, Podolsky SH, Jones DS. Assessing the gold standard — lessons from the history of RCTs. *N Engl J Med* 2016; 374: 2175-81.
14. Rychetnik L, Frommer M, Hawe P, Shiell A. Criteria for evaluating evidence on public health interventions. *J Epidemiol Community Health* 2002; 56: 119-27.
15. Brown CA, Lilford RJ. The stepped wedge trial design: a systematic review. *BMC Med Res Methodol* 2006; 6: 54.
16. Hawkins NG, Sanson-Fisher RW, Shakeshaft A, D'Este C, Green LW. The multiple baseline design for evaluating population-based research. *Am J Prev Med* 2007; 33(2): 162-8.
17. Hawe P, Potvin L. What is population health intervention research? *Can J Public Health* 2009; 100(1): Suppl I8-14.
18. Harris AD, McGregor JC, Perencevich EN, Furuno JP, Zhi J, Peterson DE, et al. The use and interpretation of quasi-experimental studies in medical informatics. *J Am Med Inform Assoc* 2006; 13(1): 16-23.
19. Cockayne NL, Glozier N, Naismith SL, Christensen H, Neal B, Hickie IB. Internet-based treatment for older adults with depression and co-morbid cardiovascular disease: protocol for a randomised, double-blind, placebo controlled trial. *BMC Psychiatry* 2011; 11: 10.
20. dos Santos Silva I. *Cancer Epidemiology: Principles and Methods*. Lyon, France: International Agency for Research on Cancer. Available from: https://publications.iarc.fr/Non-Series-Publications/Other-Non-Series-Publications/Cancer-Epidemiology-Principles-And-Methods-1999
21. Killip S, Mahfoud Z, Pearce K. What is an intracluster correlation coefficient? Crucial concepts for primary care researchers. *Ann Fam Med* 2004; 2(3): 204-8.