Evaluating the impact of interventions addressing health behaviour

Considerations and tools for policy-makers

World Health Organization
European Region

Evaluating the impact of interventions addressing health behaviour

Considerations and tools for policy-makers

World Health Organization
European Region

Abstract

Some of the most persistent public health challenges are dependent on human behaviour. These include, among many others, overuse of antibiotics, use of tobacco and alcohol, suboptimal uptake of vaccination, and cancer screening. These challenges call for evidence-based action that draws on an understanding of these health behaviours and the cultural context in which they take place and that is focused on engaging with those affected. Using evidence, models and methods from behavioural and cultural insights (BCI) allows health-related services, policies and communication to be precisely tailored and refined, thereby improving their outcomes.
A key element of BCI is impact evaluation, the primary objective of which is to evaluate whether an implemented intervention has achieved its expected goal. This guide provides considerations and tools to assist in evaluating the impact of interventions that address health behaviour. It complements the WHO *Guide to tailoring health programmes* by encouraging robust evaluation of interventions and providing starting points for engaging with an expert evaluator. This guide considers the key questions that underlie impact evaluation of interventions addressing health behaviour: Why evaluate? When to evaluate? What to evaluate? How to evaluate? Why did the intervention (not) work? An accompanying toolkit offers frameworks, a decision tool and in-depth information that complement the advice given in the first part of the guide.

Keywords

HEALTH BEHAVIOR
BEHAVIORAL SCIENCES
HEALTH POLICY
EVALUATION STUDY

Document number:
WHO/EURO:2024-10200-49972-75147 (PDF)
WHO/EURO:2024-10200-49972-75181 (print)

Suggested citation. Evaluating the impact of interventions addressing health behaviour: considerations and tools for policy-makers. Copenhagen: WHO Regional Office for Europe; 2024. Licence: CC BY-NC-SA 3.0 IGO.

Cataloguing-in-Publication (CIP) data.
CIP data are available at http://apps.who.int/iris.

Contents

Acknowledgements

This guide was produced by the Behavioural and Cultural Insights (BCI) Unit at the WHO Regional Office for Europe. The lead author of this document is Philipp Schmid (Radboud University). The coauthors are Tiina Likki and Katrine Bach Habersaat (WHO Regional Office for Europe).

The development of the guide was supported by the following members of the Technical Advisory Group on Behavioural and Cultural Insights for Health: Marijn de Bruin (Radboud University Medical Center; Dutch National Institute for Public Health and the Environment), Robert Böhm (University of Vienna) and Marc Suhrcke (Luxembourg Institute of Socio-economic Research).

A warm thanks also to the many people who have contributed by reviewing versions of this document. These include representatives from public health authorities in the WHO European Region: Alice Cline (Public Health Wales), Marina Duishenkulova (Republican Center for Health Promotion and Mass Communication, Kyrgyzstan), Ashley Gould (Public Health Wales) and Robert Murphy (Department of Health, Ireland); and other experts – Matt Barnard (ICF Centre for Behaviour Change) and David Stuckler (Bocconi University).

Abbreviations

BCI	behavioural and cultural insights
BMI	body mass index
COVID-19	coronavirus disease
CVD	cardiovascular disease
EU	European Union
GDPR	General Data Protection Regulation
PICOT	Population, Intervention, Comparison, Outcome, Time
RCT	randomized controlled trial

Introduction

This guide presents various considerations on evaluating the impact of public health interventions informed by behavioural and cultural insights (BCI) (see Box 1); it then describes a number of practical tools to help carry out such evaluations. The document is divided into two parts.

The first part, **Considerations for impact evaluation,** describes why, when, what and how to evaluate. It aims to:

- encourage robust evaluation of interventions; and
- provide some useful templates and starting points for engaging with an expert evaluator.

The second part, **Toolkit,** aims to:

- help identify which evaluation design might be most appropriate;
- help improve the quality of evaluation practices; and
- provide more in-depth information and links to further resources.

This guide accompanies the WHO *Guide to tailoring health programmes (1)*, which describes an approach for developing and implementing evidence-based interventions that address health behaviours. This impact evaluation guide can be used together with the WHO *Guide to tailoring health programmes* which offers insights for example into engaging with stakeholders which also apply to impact evaluation.

Who is this guide for?

This guide is intended for health authorities and other organizations involved in improving the outcomes of health policies, services and communication through interventions that aim to address health behaviours. It is meant for people who do not have expertise in evaluation methods but are involved in the development and evaluation of BCI-informed interventions.

BOX 1

BCI-informed interventions

Public health interventions that aim to address health behaviours are, ideally, informed by BCI and guided by robust methods and evidence. Health behaviours are complex. In the past, it was assumed that people behave in healthy ways if they have the necessary knowledge and motivation. Today, we know that many other factors play a role, including social norms, cultural context, access to the health system, and the way health workers interact with their patients. BCI-informed interventions are public health interventions that acknowledge this complexity.

Whenever there is further information available in the Toolkit, this will be indicated as follows:

TOOLKIT

PART 1

Considerations for impact evaluation

Why evaluate BCI-informed interventions

Impact evaluations are used to assess whether a specific intervention causes[1] the intended effect on health behaviours (for instance, people quit smoking, get vaccinated or engage in more physical activity) and, consequently, on health outcomes in the target population (for non-causal evaluations, see WHO *Guide to evaluating behaviourally and culturally informed health interventions in complex settings (2).*

Impact evaluations can be used to:

- detect (in)effective interventions
- save money
- avoid potential harm
- adapt interventions
- inform policy-making.

With impact evaluations, it is possible to identify the most effective (and cost-effective[2]) interventions and their most effective constituent components, in order to implement effective policies that demonstrably lead to improved health, well-being and equity.

Detecting (in)effective interventions

Even well-intentioned and well-financed interventions can be ineffective. For example, starting in 1998, the United States Congress spent almost US$ 1 billion on the so-called National Youth Anti-Drug Media Campaign, which aimed to educate young people about illegal drugs and reduce their drug use. As part of the impact evaluation, three nationally representative cohorts of 5126 American youths aged 9–18 years were surveyed four times *(3)*. Among other measures, the questionnaires included self-reports of drug consumption, intentions to avoid drugs and perceived anti-drug norms. The results showed no evidence of the effectiveness of the costly intervention on drug consumption among the target population. In fact, the authors found evidence that the campaign even decreased the intention to avoid drugs among participants aged 9–18 years at some measurement points. Without impact evaluations, interventions lacking evidence to support their effectiveness may continue to be implemented and consume resources that could be used elsewhere.

Moreover, simply because an intervention has some impact does not mean that it is the best choice. Starting in 2019, researchers tested four different interventions to increase screening rates for hypertension and diabetes among 6934 Armenians *(4)*. Individuals received one of two versions of a personalized invitation (interventions 1 and 2), a personalized invitation plus a free pharmacy voucher (intervention 3), or a personalized invitation plus a pharmacy voucher that was only given to participants if they attended the screening (intervention 4). Interventions 3 and 4 were more expensive than interventions 1 and 2 because the vouchers added costs. However, the impact evaluation revealed that intervention 3 led to very similar increases in screening rates (+15%) as interventions 1 and 2, while intervention 4 resulted in the highest screening rates (+31%). Thus, the cost-effectiveness of interventions 1, 2 and 4 was very similar, while intervention 3 was estimated to be about twice as expensive for each additional person screened. As a result, the Armenian Ministry of Health decided to scale up an intervention similar to intervention 1. This illustrates how impact evaluations can inform policy decision processes, not only by detecting effective interventions but also by comparing the effectiveness of the most promising candidates while considering their associated costs.

Saving money

Impact evaluations help to make an argument for further investment and scaling of effective approaches, including ones that may be costly. For example, in 1998 the community-based cardiovascular disease (CVD) prevention programme *Hartslag Limburg* [Heartbeat Limburg] started in the Netherlands (Kingdom of the). The programme consisted of several interventions (involving, for instance, more physical activity and less fat intake) aimed at increasing a healthy lifestyle and decreasing

1 For an explanation of causality, see the Glossary.
2 For an explanation of cost-effectiveness, see the Glossary.

the prevalence of CVD in the general population of the Maastricht region *(5, 6)*. As part of the impact evaluation, a cohort of 2414 people from the intervention region and 758 people from a control region were surveyed, first in 1998 and then five years after the implementation of the programme. The results revealed that changes in risk factors for CVD (such as high body mass index (BMI) and high blood pressure) were greater in the intervention region than in the control region *(5)*. Researchers estimated that extending the intervention to large parts of the Dutch population would cost about €45 million; despite this large investment, their cost-effectiveness analysis revealed that the cost of the programme would be around €5100 per life year saved *(7)*. The researchers rated the intervention as cost-effective given (among other things) the costs of other interventions to prevent CVD (such as intensive glycaemic control: €40 881) and the total health-care budget of the country. Impact evaluations paired with cost-effectiveness analysis or economic evaluation enable researchers and policy-makers to consider costs in relation to the effect of an intervention. Thus, the difference between good and bad investments can only be brought to light when impact evaluations that form the basis for cost-effectiveness analysis or economic evaluation are conducted.

Avoiding potential harm

BCI-informed interventions are usually designed and disseminated with good intentions. However, this does not prevent a given intervention from potentially causing unintended effects and, in the worst case, causing more harm than good *(8)*.[3] There are several different types of unintended effects that have the potential to cause harm *(9)*. For example, reminder and warning systems are often used effectively to nudge individuals towards healthier options (for example, reminders to get vaccinated *(10)*). However, excessive exposure to such nudge interventions can lead to so-called desensitization among receivers, with potentially harmful consequences *(9)*. An overload of warnings for clinicians – that is, a high proportion of false alarms – is known to cause alarm fatigue, which can potentially compromise patient safety *(11)*. Several message interventions from various health domains (such as AIDS prevention and breastfeeding) have caused negative impacts, including increased misperceptions or frustration among target groups *(9)*. In addition, there are situations in which interventions can lead to greater health inequality by benefiting only advantaged groups and leaving out those most in need *(12)*. Evidence-based intervention design and involvement of target groups can minimize the risk of unintended effects, but the potential for such effects can ultimately be determined only through impact evaluation. Thus, well-designed impact evaluations that measure unintended effects also serve to avoid harm.

[3] For an explanation of causality, see the Glossary.

Adapting interventions

Sometimes even impactful interventions require revision and adaptation to meet the needs of different settings and different groups. Impact evaluations allow occasions when such adaptation is needed to be identified. For example, gamified interventions based on the inoculation theory have been found to be effective measures to mitigate the impact of misinformation (concerning coronavirus disease (COVID-19)), for instance) in developed countries, where they have been shown to raise people's awareness of the threat of misinformation and provide them with strong counterarguments against it *(13,14)*. However, the use of a gamified inoculation intervention in northern India did not protect the target group against misinformation *(15)*. One explanation proposed by those who carried out the study is that the design of the intervention may not have been aligned well enough with cultural and local contexts. Without impact evaluations, policy-makers and researchers may falsely assume that there are one-size-fits-all approaches that can be adopted to solve global threats to health.

Informing policy-making

In addition to offering important data on the effectiveness and cost-effectiveness of interventions, impact evaluations are powerful tools to gain the attention of political decision-makers and to inform policy decision-making. The results of an impact evaluation can help convince policy-makers to implement a BCI-informed intervention at scale. For example, the European Union (EU) Commission uses impact assessments as a standard methodology to analyse the social, economic and environmental consequences of policies *(16,17)*. These assessments include weighing different policy options for a given problem. Impact evaluations of BCI-informed interventions can inform this weighing process and thus lead to policy changes at the EU level.

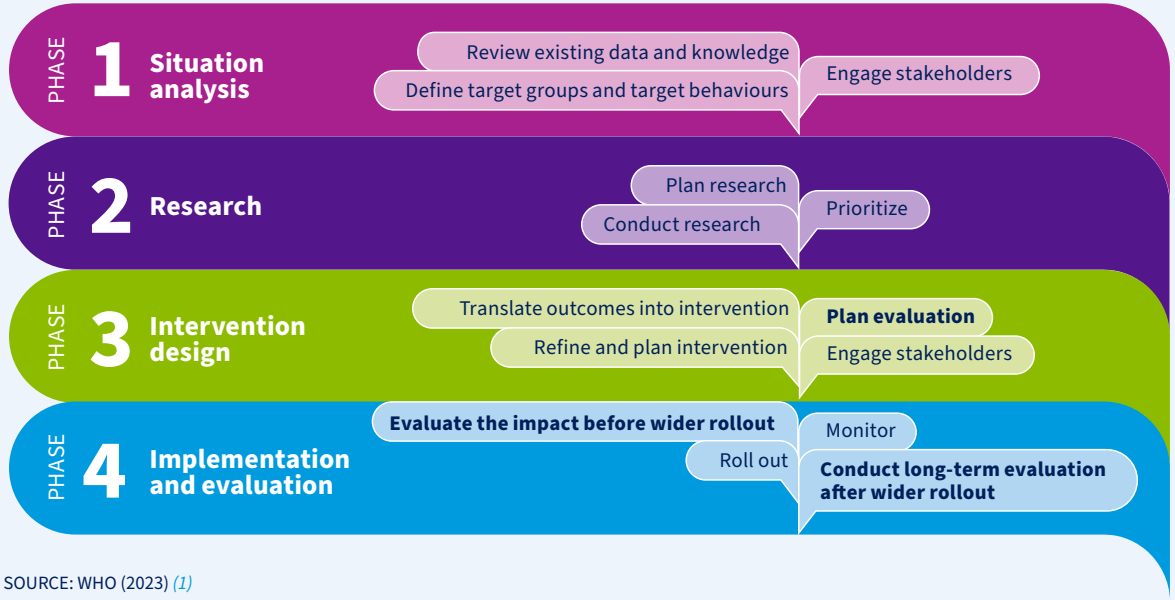
Often the motivations of policy-makers for or against policy changes, such as the realization of or support for a BCI-informed intervention, are multifaceted. Ethical considerations, costs and potential benefits of policy change, and public support for policies *(18)*, among other things, play a key role. Impact evaluations can provide a scientific basis to confirm the ethical, economic and/or empirical advantage of an intervention or to uncover the risks of a potentially harmful intervention. In sum, impact evaluations are a crucial element to inform, justify or change evidence-based policies.

When to evaluate: four phases of planning, implementing and evaluating a BCI-informed intervention

While impact evaluation is usually conducted towards the end of a BCI project (Fig. 1), it is important to plan the evaluation before the intervention is implemented.[4] In the WHO *Guide to tailoring health programmes (1)*, evaluation planning is included in Phase 3 (Intervention design) and includes mapping possible outcome measures[5] and selecting a research design. This is followed by the actual evaluation of a BCI-informed intervention in Phase 4 (Implementation and evaluation).

Once a BCI-informed intervention has been designed, it is recommended to evaluate it for effectiveness before wider rollout. Each evaluation should follow the principle of proportionality, meaning that the resources invested into the impact evaluation should be commensurate with the usefulness of the evaluation findings.

Fig. 1 Four phases of planning, implementing, and evaluating a BCI-informed intervention



SOURCE: WHO (2023) *(1)*

[4] The planning includes, inter alia, the selection of an evaluation design. The evaluator can then implement the intervention in a way that matches the requirements of the chosen design. If the intervention is already implemented, then design choices are limited.

[5] For an explanation of outcome measures, see the Glossary.

BCI-informed interventions can be evaluated before and after wider rollout. The goal of the evaluation before rollout is to evaluate whether the planned BCI-informed intervention is effective in principle – that is, whether the intervention produces the desired outcome in a smaller but potentially more controllable setting (for instance, if tested online); and to identify how the intervention may need to be adjusted or implemented in a given setting (see section “Why it did (not) work” below).

After a successful first evaluation, the intervention can be rolled out and a second evaluation takes place – that is, the evaluation after rollout. The goal of the second evaluation is to evaluate whether the implemented intervention has achieved the expected goal in the target population. This second evaluation is not an “in-principle” assessment but the primary impact evaluation of the intervention in the field.

Numerous research designs are available to evaluate BCI-informed interventions before and after rollout. A selection of key evaluation designs, together with their advantages and disadvantages, is presented in the section “How to evaluate” below, with greater detail provided in the Toolkit.

What to evaluate: forming a research question for impact evaluation

Before conducting any research study, you should have a clear understanding of the primary research question that the study is going to address. This research question guides the selection of a research design. Some designs are simply not suitable to address particular research questions. A clear research question also makes the interaction with an evaluation expert much easier.

Impact evaluations of BCI-informed interventions often address the general research question: "Can the BCI-informed intervention cause the desired effect?" Various tools can then help to further specify the question; one such tool, the Population, Intervention, Comparison, Outcome, Time (PICOT) framework *(19)*, does so using five guiding questions.

Constructing a theory of change

Following the WHO *Guide to tailoring health programmes*, it is recommended that a theory of change is constructed that explains the rationale for the BCI-informed intervention: the barriers and drivers it targets, and how addressing these through the intervention is intended to influence the desired outcomes *(1)*. Investing time in developing a theory of change is key to interpreting the results of any impact evaluation.

TOOLKIT

Find more information on PICOT and theory of change on pages 17-18.

How to evaluate: key impact evaluation designs

Impact evaluations are used to assess whether a BCI-informed intervention causes an intended effect (for instance, more people quit smoking, go for vaccination or follow their treatment plan). Simply exposing a group, organization or region to a BCI-informed intervention and measuring the outcome is not enough because it remains unknown what would have happened without the intervention (in other words, it is necessary to consider the counterfactual[6]). Thus, impact evaluations always require a treatment group and a comparison group.

This guide focuses on two main categories of impact evaluation designs: randomized controlled trials and natural experiments. The evaluation designs included in this guide all provide evidence, with varying degrees of strength, that allow an assessment of whether the BCI-informed intervention had a causal impact on a particular outcome or outcomes. They differ primarily in how the comparison group is determined.

This has important consequences for:

- the quality of findings (for instance, the strength of evidence for causal claims)
- the requirements and costs of conducting the specific design.

In addition to impact evaluation designs (randomized controlled trials and natural experiments), other methods and approaches (surveys and qualitative research) are introduced below. These cannot confirm causal relationships, but they can provide very useful insights for impact evaluations: they can help to identify the strengths and weaknesses of intervention material, to determine how to improve implementation and enhance public attitudes towards policy interventions, and to discern potential reasons for the success or failure of implementing a BCI-informed intervention.

Randomized controlled trials

Randomized controlled trials (RCTs) are considered the gold standard of impact evaluation (for an example of using an RCT, see Box 2). The simplest RCT design consists of a group that is exposed to a BCI-informed intervention (the treatment group) and a group that is not exposed to the intervention (the control group). Whether an individual (or higher-order unit, such as a clinic or region) is part of the treatment group or part of the control group is determined by randomization (that is, a process equivalent to tossing a coin). The random allocation and the comparison of a treatment with a control group give RCTs their name. Another approach, referred to as crossover design, is to expose the entire sample to both an intervention and a control scenario, one after another.

Subtypes of RCTs:

- crossover
- stepped-wedge.

[6] For an explanation of counterfactuals, see the Glossary.

Natural experiments

Natural experiments[7] are typically conducted as a form of impact evaluation when controlled randomization of individuals or higher-order units (clinics, regions, countries, etc.) is not possible. Natural experiments follow the same logic as RCTs. Some individuals or higher-order units are exposed to the intervention (treatment group) and some are not (control group). When multiple measurements are available, more specific types of natural experiments are possible (for an example of using a natural experiment, see Box 3).

Subtypes of natural experiments:

- synthetic control
- difference-in-differences
- repeated measures (pre-post)
- interrupted time series.

BOX 2

Using an RCT to evaluate the impact of information about scientific consensus on vaccine uptake

In 2021, during the COVID-19 pandemic, researchers conducted a field RCT among 2101 Czech adults *(20)*. Participants were randomly allocated to either a treatment group or a control group. The treatment group received information as part of an online survey about the scientific consensus of doctors on the trustworthiness and safety of COVID-19 vaccines. The information (for instance, that 89% of doctors trust the vaccines) was presented to them in the form of charts and written summaries. The control group did not receive any information. The authors measured COVID-19 vaccine uptake in both groups over nine months. The results revealed that the uptake rates in the intervention group steadily increased by 4–5 percentage points compared to the control group. This beneficial outcome due to the intervention remained stable over time. Thus, the impact evaluation of this BCI-informed intervention revealed that simple written information about a scientific consensus could have an impact on actual vaccine uptake rates in the public.

BOX 3

Using a natural experiment (difference-in-differences) to evaluate the impact of indoor smoking bans

Researchers analysed the impact of indoor smoking bans on smoking behaviour and lung function in the general population of Denmark *(21)*. Indoor smoking bans were introduced in Denmark in 2007. The researchers used a natural experiment by comparing 62 093 Danish adults (treatment group) with 31 807 Swiss adults (control group) on outcome measures. Switzerland was considered an appropriate control because, inter alia, it introduced indoor smoking bans only in 2010. The authors collected data about outcome measures from 2005 to 2010. Having repeated measures and a control group, they decided to perform a difference-in-differences analysis. Their results revealed: “Nationwide indoor smoking ban is associated with less smoking and improved lung function in the general population” *(21)*.

TOOLKIT

Find more information on RCTs on pages 19-20.
Find more information on natural experiments on pages 20-21.

[7] It is common that the term ‘natural experiments’ also includes ‘quasi-experiments’. In line with this, we use the term ‘natural experiment’ throughout this document. For a discussion and differences in definitions of natural and quasi-experiments, see de Vocht et al. 2021 *(48)*

Why did it (not) work: complementing impact evaluation

While impact evaluation is effective in determining the causal effects of an intervention, it often leaves little room to explore the perceptions and experiences of the target audience or of those implementing the intervention, to gain a better understanding of why an intervention did or did not work, and how it could be improved.

A range of methods can be used to complement impact evaluation, many of them stemming from the field of implementation research *(22–24)* (for an example of using implementation research, see Box 4). By using such methods, it is possible:

- to identify potential reasons for the success or failure of a BCI-informed intervention;
- to understand the mechanisms that explain the effectiveness of an intervention;
- to identify the strengths and weaknesses of the intervention process and implementation (for example, acceptability among those involved, such as health workers, or possible limitations in an organization's effectiveness and ability to implement the intervention);
- to understand the attitudes of those affected towards intervention elements (for example, opinions about mandatory vaccination or perceptions of health communication materials);
- to understand intended and unintended effects (such as impacts on well-being, sense of social cohesion, or trust among those involved and targeted); and
- to assess the broader positive and negative implications of an intervention.

Taken together, this complementary knowledge can inform decisions on why an intervention did or did not work, and how it could be improved. It can be gained, among others, through survey studies and qualitative studies.

Survey studies

Surveys can be used to collect descriptive data about how an intervention is perceived in the target group or about the self-reported impact of an intervention (for instance, the emotional response or the knowledge or behaviour change following a training). Surveys can be conducted:

- before rollout (to incorporate feedback as soon as possible);
- during rollout (to understand how the intervention is being implemented); and/or
- after rollout (to gain a better understanding of the intervention's impact).

Surveys aim to provide an accurate indication of average views and experiences across a population, and to do so, they rely on the sample being representative of the target group.[7] Public perceptions of a BCI-informed intervention can be useful when assessing why an intervention was effective or ineffective.

BOX 4

Using implementation research to find applicable solutions

The Food Dude programme is a multicomponent intervention that increases children's in-school consumption of fruits and vegetables in the United States *(25)*. One of the components of the programme is a reward system for consuming a specific amount of fruits and vegetables. Researchers evaluated the impact of different reward systems. For example, one study revealed that the programme with monetary rewards ($12.50 per prize) led to a 92% increase in fruit and vegetable consumption among children compared to a control group *(26)*. Moreover, the effect was still significant in a six-month follow-up. In contrast, when teacher praise was used as a reward instead of monetary rewards, the benefit decreased by about 50% and was not detectable in the six-month follow-up. Considering effect sizes only, the Food Dude programme should be implemented with a monetary reward system. However, teachers and policy-makers raised several issues when asked about the monetary reward system. For example, the financial costs of the intervention and the administrative work that was needed to manage the system at schools were raised as barriers to implementation. More manageable reward systems were needed, especially when financial and labour resources were limited. Thus, researchers began to evaluate different implementation strategies for systems to reward healthy eating. For example, researchers invented a game-based reward system that used fictional rewards for a fictional hero within a children's story *(27)*. These game-based approaches were effective at increasing fruit and vegetable consumption and were deemed promising as a low-cost intervention that required little additional administrative work *(28)*. However, the longevity of most reward systems remained a challenge. The studies reveal how important it is not only to focus on impact but also to consider different types of implementation. Effective BCI-informed interventions are only useful if implementable.

[7] On representativeness, see Ramsey & Hewitt (2005) *(29)*.

Qualitative studies

Qualitative studies can help to explore what worked well in an intervention and what could be improved (for an example of using qualitative studies, see Box 5). They can shed light on intended and unintended effects, as well as on the perceptions, emotional responses and self-reported experiences of those involved and targeted. Such studies may also help to analyse in more depth the contribution of an intervention to the health outcome: how, why, when and with whom it had an effect *(2,30)*. A variety of qualitative methods are available, including:

- observational studies
- focus groups
- in-depth interviews.

When conducting observational studies, researchers usually observe members of a target group visually and take field notes about the relevant behaviour. Other approaches include in-depth interviews in which lists of topics are used as the starting point to explore participants' experiences and perceptions in order to gain specific insights from the target group *(31)*. Interviews can also be conducted with groups rather than individuals (for example, focus group discussions *(30)*).

BOX 5

Using qualitative studies to gain additional insights into an anti-smoking campaign

In 2002 an impact evaluation of an anti-smoking campaign, the so-called "truth" campaign, revealed that implementation of the campaign was associated with more negative attitudes towards smoking among young people in the United States *(32).* The campaign included slogans such as "Your pee contains urea. Thanks to tobacco companies, so do cigarettes. Enjoy." The goal of slogans of this kind is usually to increase disgust towards the ingredients of the unhealthy product and thus to reduce willingness to consume the product among the target group. In addition to the quantitative evaluation, researchers conducted over 100 qualitative interviews with college students who were either smokers or non-smokers *(33)*. The qualitative interviews revealed that the slogan seemed to have produced the intended effect among non-smoking college students. For example, one of the students stated: "If I smoked, it would really make me think twice. I would be so disgusted by this fact that I would try to stop smoking right away." However, the opposite observation was often made by smoking college students. For example, one of the students stated: "All the 'truth' campaign does is convince me that I should go outside and light up another cigarette." The researchers concluded that anti-smoking campaigns, even if found to be successful in an impact evaluation, could be a waste of money when applied to specific relevant subgroups – in this case, current smokers. Such insights can be very important in adapting BCI-informed interventions to specific target groups and increasing their likelihood of success across populations.

Quality checklist: improving the evaluation with the Toolkit

- [ ] I know how to formulate a research question.
  If NO or in doubt, go to TOOL 1: PICOT on page 17.
- [ ] I know what is meant by a theory of change.
  If NO or in doubt, go to TOOL 1: Theory of change on page 17.
- [ ] I have a good understanding of different RCTs and natural experiments.
  If NO or in doubt, go to TOOL 2: Research designs on page 19.
- [ ] I know which research designs are appropriate for my evaluation.
  If NO or in doubt, go to TOOL 3: Decision Matrix on page 22.
- [ ] I have a good justification for the sample size of the evaluation study.
  If NO or in doubt, go to TOOL 4: Selecting a sample size on page 26.
- [ ] I have considered an a priori power analysis.
  If NO or in doubt, go to TOOL 4: Selecting a sample size on page 26.
- [ ] I have considered preregistering the evaluation study
  If NO or in doubt, go to TOOL 4: Preregistering a study on page 27.
- [ ] I have checked whether the study meets ethical standards.
  If NO or in doubt, go to TOOL 4: Following ethical standards on page 27.
- [ ] I know what to do with the evaluation results if they are not as expected.
  If NO, go to TOOL 4: Sharing stories of failure or inconclusive results on page 28.

Concluding remarks

Impact evaluations can be used to detect effective or ineffective BCI-informed interventions, to save money, to avoid potential harm, to adapt interventions, and to inform policy decision-making. As shown through examples in the preceding sections, the benefits of impact evaluations usually outweigh their costs. In spite of this, discussing impact evaluation with an expert and choosing an appropriate evaluation design are often met with hesitation because of the assumed complexity and the additional costs associated with impact evaluations.

The preceding text offers a few guiding questions (and some suggested answers) that can make the process easier: Why evaluate? When to evaluate? What to evaluate? How to evaluate? Why did it (not) work? The text that follows describes some tools that will help to conduct impact evaluations in a user-friendly manner. The goal remains to encourage public health authorities to consider impact evaluation when planning BCI-informed interventions.

The toolkit in Part 2 offers in-depth information, frameworks and a decision tool that complement the advice given in Part 1 of this guide. Fill out the checklist on the previous page to assess which of the following tools should be considered before talking to an evaluation expert.

PART 2

TOOLKIT

Tool 1. What to evaluate

PICOT

Various tools are available that can help to further specify your research question. Research shows that using structured frameworks such as PICOT to determine your research question is associated with higher-quality research or better research reports *(34)*. The PICOT framework *(19)* increases the specificity of your research question by posing five guiding questions:

P Who is the target **population** for the impact evaluation?

I What is the BCI-informed **intervention** being considered for the impact evaluation?

C What is an appropriate **comparison** or **control** group for the impact evaluation?

O What is the desired **outcome** measure for the impact evaluation?

T At what **time** will the impact be detected?

Using this framing, an example of a research question for the impact of a sugar tax on the health of school-aged children could then be:

What is the effect of the introduction of a sugar tax [intervention] on the prevalence of obesity [outcome] among school-aged children living in country X [population] three years after the introduction of the tax [time] compared to a period before the tax was introduced in country X [comparison]?

Theory of change

A theory of change is a tool used to design more precise research hypotheses. Investing time in developing a theory of change is key to interpreting the results of any impact evaluation. This logic model should provide answers to the following questions:

- What was the challenge?
- What was the end goal?
- What were the assumptions about each of the steps towards the goal?

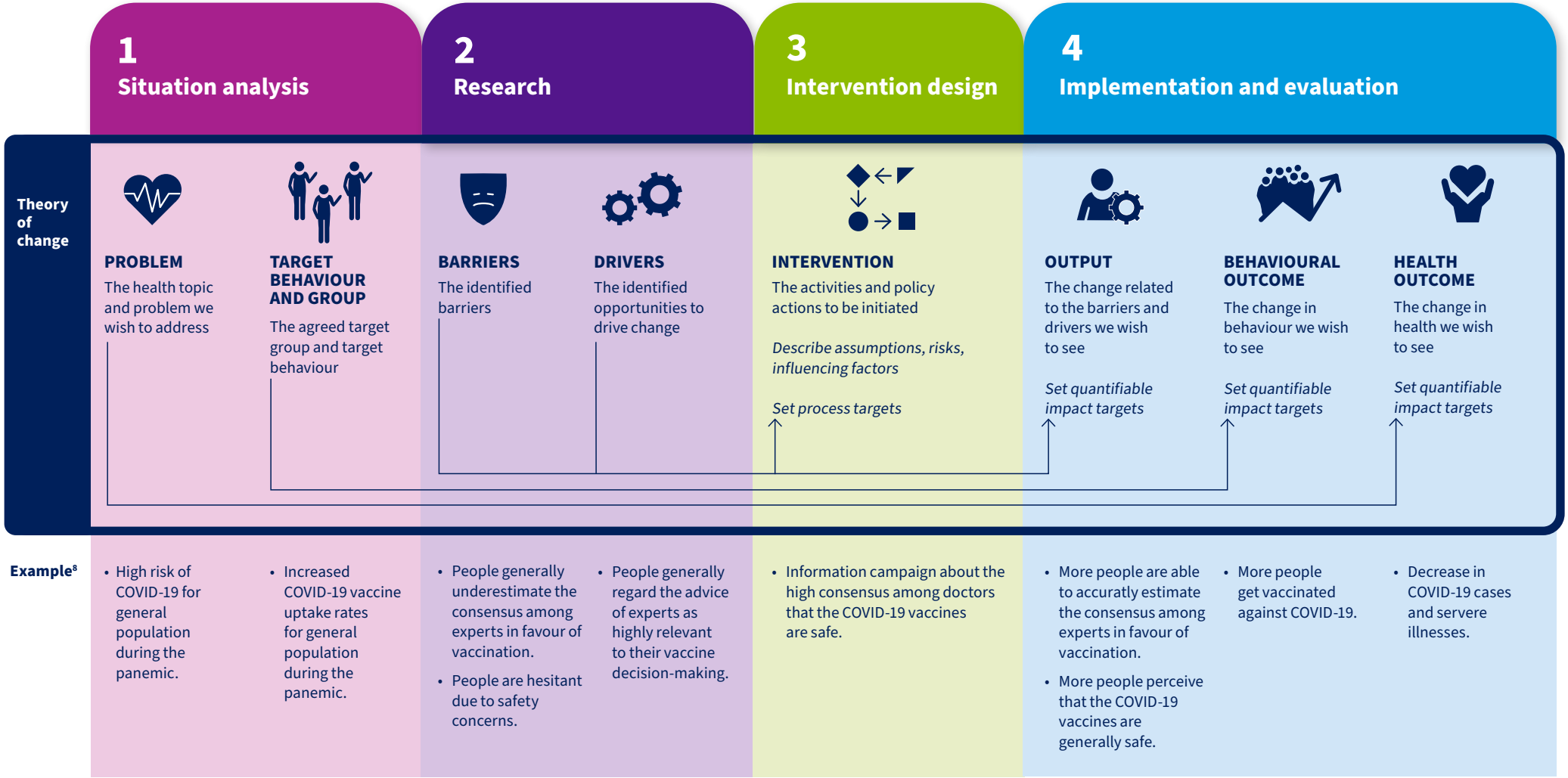
These questions go beyond formulating an initial research question because they help the evaluator to think about the exact process of how a BCI-informed intervention is meant to cause impact. Once this process is documented, it becomes a theory of change that can be revisited and refined during the evaluation process as new knowledge emerges.

For example, reverting to Box 2 above (page 10), the researchers who used scientific consensus messaging to increase vaccine uptake among Czech adults during the COVID-19 pandemic hypothesized the following: by exposing individuals to factual scientific consensus information ("89% of doctors trust the vaccines"), individuals' misperceptions about how many doctors trust the COVID-19 vaccines decrease *(20)*. This, in turn, should lead to more vaccine uptake because medical doctors are usually considered a highly trustworthy source for health decision-making. The described process can be reduced to:

exposure to consensus messaging → decrease in misperceptions → increase in vaccine uptake.

This assumption about how a BCI-informed intervention works can be described as a theory of change, and it can be incorporated into the four phases of planning, implementing and evaluating a BCI-informed intervention (Fig. 2). In fact, the Czech research found the hypothesized impacts: vaccine uptake was indeed increased, and this increase could be explained by a decrease in misperceptions about the scientific consensus *(20).* Theories of change can be much more complex than this example, depending on the specific research question and the associated process that leads to change.

Fig. 2 Theory of change



[8] The example is based on the study presented in Box 2, from a study in Czechia *(20)*.

Tool 2. How to evaluate: research designs

RCT

RCT are considered the gold standard for analysing causal relationships *(35)*. The simplest RCT design consists of a group that is exposed to a BCI-informed intervention (the treatment group) and a group that is not exposed to the intervention (the control group). Whether an individual (or higher-order unit, such as a clinic or region) is part of the treatment group or part of the control group is determined by randomization – that is, every individual is randomly allocated to one of the groups. This procedure is used to eliminate selection bias, to facilitate blinding (see Box 6), and to reduce the influence of unknown confounders *(36)*. Confounders are variables (such as gender, age, education, attitudes towards the outcome, contextual changes, or the presence of other interventions or campaigns) that can influence the study results and thereby violate the assumption that the results are only due to differences between the treatment group and the control group.[8] Despite their strength, RCTs are not immune to biases *(37)*.

Cluster randomization

Sometimes, randomization cannot be conducted at the individual level. For example, a BCI-informed intervention may be implemented at a clinic level, where everyone in a clinic is (intentionally) exposed to the intervention, but spillover from intervention wards to other wards cannot be controlled (it is unintentional). In this case, one needs to find additional clinics and randomly select which clinics serve as treatment clinics and which as control clinics. This sort of randomization on higher-order units (such as clinics, villages or regions) is often referred to as cluster randomization. Cluster trials can also be used when researchers want to know whether BCI-informed interventions that are effective at the individual level are also effective at scale *(38)*.

Crossover design

Another way of selecting a control group is to use the same individuals or higher-order units (such as clinics, villages or regions) but to expose them to the intervention or the control at different points in time. Randomization is used in these so-called crossover designs to eliminate order effects by randomly determining which exposure (treatment or control) comes first. Theoretically, these designs are more efficient than classic RCTs because they produce more precise estimates given the same number of participants *(39)*. Moreover, they are often considered a good choice from an ethical perspective because all participants receive the treatment sooner or later – that is, everyone is treated equally. This kind of design is typically used in clinical research, in which a period with medication (for instance) is compared

BOX 6

Blinding to treatment assignment

Blinding is an essential part of RCTs to ensure that the results of impact evaluations are not influenced by participants who want either to please or to displease the researcher. Blinding refers to the fact that whether an individual (or higher-order unit) is allocated to the treatment or control group remains unknown to participants. This is often accomplished by random allocation of participants to one or other group. Thus, participants cannot choose the group they are assigned to, and they are also often not informed about specific research hypotheses until the end of a BCI impact evaluation (but see also "Following ethical standards", page 27 below). Moreover, the source of the data (control group or treatment group) can also be hidden from the analysing researchers to ensure that their analysis is not influenced by what they would like or expect to find.

[8] For further explanation of confounders, see the Glossary.

with a period without medication. While there are BCI-informed interventions that can be evaluated using a full crossover design – for instance, when interventions involve changes in the environment, such as adding opportunities for physical activity or healthier food choices – it is often difficult to realize this kind of design in the case of BCI-informed interventions. The problem is that, after being exposed to a particular BCI-informed treatment, such as a weight-loss programme, a workplace intervention for physical activity or a group-based smoking cessation, participants in the crossover trial cannot undo or unthink the effect of the treatment. Thus, participants who are allocated to the control group after the treatment will be biased by the treatment effects. In these cases, trials can only be conducted with one-order control: first no treatment and then treatment (see "Stepped-wedge randomized trials" below).

Stepped-wedge randomized trials

In stepped-wedge trials, all participants receive the treatment, but the starting point is randomized *(40,41)*. Such trials offer some of the strengths of crossover trials. For example, everyone is treated equally – that is, by the end of the trial every participant will have been exposed to the treatment. In contrast to crossover trials, stepped-wedge trials do not require that participants undo or unthink the effects of the treatment, because the trial is unidirectional (usually from no treatment to treatment). Some trials are unidirectional but do not randomly allocate participants to different starting points. These solutions do not count as RCTs and are prone to bias. However, they can still provide useful results and are discussed in the next section under "Repeated measures trials".

Field, laboratory or online

RCTs can be conducted in the field (that is, in a real-world setting, such as a health clinic), either before or after rollout, or in a laboratory setting or online, before rollout. In a laboratory, researchers are in the best position to control the setting and confounding factors, and thus to ensure that the results are due only to exposure to a BCI-informed intervention as opposed to no exposure. To conduct an RCT in a laboratory, it is necessary to have appropriate facilities and access to target groups that can be invited to act as participants in the laboratory. An alternative to such resource-intensive research is online RCTs. In many countries, there are panel providers (that is, corporations that offer research samples) that can provide access to and responses from target groups within days. Alternatively, participants can be recruited through social media and other networks, although the sample produced is likely to be less representative of the population than one provided by a panel provider. Online RCTs are more suitable for certain types of intervention (ones involving delivery of a message or some visual or other form of communication), but less so for interventions with a physical, social or environmental dimension. Compared to laboratory RCTs, they also often lack experimental control;[9] nevertheless, they offer an efficient way to evaluate an intervention before rollout in a time- and cost-efficient manner.

Further reading
- on the probabilistic theory of causality and RCTs: Cartright (2010) *(42)*
- on using RCTs to evaluate complex public health interventions: Bonell et al. (2012) *(43)*.

Natural experiments

In natural experiments, much like in RCTs, some individuals or higher-order units (such as clinics, villages or regions) are exposed to the intervention (treatment group) and some are not (control group) *(44)*. The primary difference from an RCT is that the allocation into treatment and control groups is not controlled by the researchers but determined by the given circumstances – that is, determined by "nature". For example, policy-makers could decide to introduce plant-based diets in all primary school canteens. In this case, the intervention group is already fixed and cannot be chosen by a control mechanism such as randomization. Natural experiments are usually used as an evaluation method after wider rollout, but not before. (For an example of the use of natural experiments, see Box 3 on page 10.)

[9] In laboratory settings, it is possible to control the tools, information and interaction with other participants because researchers can prepare the laboratory in certain ways and are usually present when the study is conducted. This reduces the influence of potential confounders. In online studies, it is much more difficult to control whether, for example, participants are using secondary sources or are distracted by other tasks while taking part in the study.

The difficulty with a natural experiment is to find a comparison group – that is, a control group of individuals or higher-order units (such as clinics, villages or regions) that is as similar to the treatment group as possible. To create a comparison group, it is recommended to identify relevant potential confounders and statistically control for the influence of these variables and/or to find a control group that matches the treatment group with respect to these confounding variables. As an example of the latter, when analysing the impact of mandatory vaccination in a country, choosing a different country with similar uptake rates at the start of the study as a control group can provide a meaningful comparison.

Synthetic control

An alternative to finding a single real appropriate comparison group is to create a synthetic control, where several control groups are averaged to best represent groups that are not exposed to the intervention *(44)*. This average is then used as the so-called synthetic control and compared to the treatment group.[10]

Difference-in-differences

One approach that can be used even when treatment and control groups in a natural experiment differ in baseline values (for instance, if regions have different vaccine uptake rates at the start of a study) is the so-called difference-in-differences method. In difference-in-differences, researchers measure the primary outcome in the treatment group and the control group before and after implementation of the BCI-informed intervention. The impact evaluation focuses on change scores – that is, it compares changes in the outcome measure from before the intervention to after the intervention between the two groups: difference (treatment versus control group) in difference (baseline versus after intervention). For example, difference between a neighbourhood that received financial support for physical exercise programmes and a neighbourhood that did not receive such support (treatment versus control group) in difference of BMI of people living in those neighbourhoods measured before and after the introduction of the intervention (baseline versus after intervention). Difference-in-differences is also often referred to as "comparative interrupted time series design" or "nonequivalent control group pretest design" *(46)*. (For an example of the use of difference-in-differences analysis, see Box 3 on page 10.)

Repeated measures trials (pre-post)

If no control group is available, outcome measures in the treatment group from before the intervention and after the intervention can be analysed on their own. In this case, individuals, clinics, regions or countries serve as their own control group. These so-called repeated measures trials, or pre-post analyses, lack a proper control group – that is, even if a desired change is noted after the intervention is implemented, it may also have occurred without the intervention, and there is no way to be certain that the change is due to the intervention. One way to increase the quality of this kind of design is to monitor change in the outcome measure for a longer period before the intervention (often referred to as "waiting-phase control"). This monitoring allows positive or negative trends in outcome measures that are present even without the intervention to be detected.

Interrupted time series

Interrupted time series design uses the idea of simple repeated measures trials. Instead of measuring the outcome twice (before and after the BCI-informed intervention), researchers measure the outcome several times before and after treatment *(47)*. In the absence of a control group, they use the change between measures before the intervention and estimate how the treatment group would have performed without the BCI-informed intervention. This artificial trend, the so-called counterfactual data, is then compared with the observed treatment data. This design is more complex than simple pre-post designs because it requires long-term trend data.

Further reading

- on conceptualizing natural and quasi-experiments in public health: de Vocht et al. (2021) *(48)*
- on natural experiments in the social sciences: Dunning (2012) *(49)*.

[10] For a more detailed overview of adjustments for natural experiments, see Craig et al. *(45)*.

Tool 3. How to evaluate: selecting a research design

Decision Matrix to assist in choosing the appropriate evaluation design

The Decision Matrix below serves as a framework for the selection of a suitable evaluation method. The Matrix consists of six steps, each with a statement that can be answered either YES or NO. Depending on the answer to each step, certain evaluation designs may become excluded, meaning they are not suitable to evaluate the impact of the intervention that is being considered. Once a design is excluded in a step, it remains excluded for the rest of the steps. If, after five steps, multiple designs are available, the decision can be based on step 6 on the strength of the design.

To learn how to use the Decision Matrix, two illustrative examples of using it to guide evaluation design choice are provided on following pages.

Your own scenario may be more complex, and some statements from the Matrix may not be applicable to your case. The Decision Matrix is a decision tool, not a dogma. Deviations from the tool or the advice given in this guidance document can be discussed with an evaluation expert.

Decision Matrix for policy-makers

Evaluating the impact of interventions addressing health behaviour

Excluded ✕

Step	Exclusion criteria	Answer: YES or NO		Impact evaluation: Randomized controlled trial – Crossover	Stepped-wedge	Randomized controlled trial	Natural experiment – Natural experiment	Difference-in-differences	Pre-post	Interrupted time series
1	**The intervention has already been implemented**	If NO, then go to the next step	If YES, then the following designs are excluded →	✕	✕	✕				
2	**Group comparison is possible** (you may answer NO if, for example, data is only available for entire population)	If YES, then go to the next step	If NO, then the following designs are excluded →	✕	✕	✕	✕	✕		
3	**Randomization is possible** (you may answer NO if, for example, a policy intervention is already planned in a fixed region)	If YES, then go to the next step	If NO, then the following designs are excluded →	✕	✕	✕				
4	**Outcome can be measured before implementation** (you may answer NO if, for example, no data is available from before implementation)	If YES, then go to the next step	If NO, then the following designs are excluded →	✕	✕			✕	✕	✕
5	**Intervention can be undone** (you may answer NO if, for example, the intervention will not or cannot be interrupted)	If YES, then go to the next step	If NO, then the following designs are excluded →	✕						✕

If, after the previous five steps, multiple designs are available, the decision can be based on the following step 6 on the strength of the design.

		Crossover	Stepped-wedge	Randomized controlled trial	Natural experiment	Difference-in-differences	Pre-post	Interrupted time series
6	Weak / Medium / Strong	Strong	Strong	Strong	Medium	Strong	Weak	Medium

Illustrative example 1. Intervention to increase health through physical activity

Scenario: Neighbourhood A receives financial funding from the State to provide more opportunities for the 1000 residents to engage in physical activity (for instance, building green areas that motivate residents to become physically active). The decision to fund this neighbourhood is made, and there is no possibility to intervene – rollout is happening as planned by the State. The neighbourhood was chosen because it has an especially low health index compared to other neighbourhoods, including a high average BMI score. The State wants to know whether the intervention has worked or not, to justify future investments in other neighbourhoods. There are several options to conduct this impact evaluation, so the responsible team uses the Decision Matrix for a first assessment of an appropriate evaluation design.

1. They answer **NO** to **The intervention has already been implemented** because the State has not started building new facilities and providing opportunities for physical activity. ***Thus, all design types are still possible.***

2. They answer **YES** to **Group comparison is possible** because the neighbourhood can be considered a group and there are other similar neighbourhoods available that allow a comparison. ***Thus, all design types are still possible.***

3. They answer **NO** to **Randomization is possible** because the State determined a specific intervention neighbourhood and changing that is (a) not possible and (b) might cause ethical problems, given the reasons the State chose this neighbourhood in the first place. ***Thus, RCTs are no longer an option.***

4. They answer **YES** to **Outcome can be measured before implementation** because the State has knowledge about at least one important outcome measure before the intervention rollout – that is, BMI. ***Thus all natural experiments are still possible.***

5. They answer **NO** to **Intervention can be undone** because undoing the intervention (a) is nearly impossible as it requires removal of buildings and (b) can be judged as unethical.
   Thus, interrupted time series studies are no longer an option.

In sum, all RCTs and interrupted time series studies are excluded after using the Decision Matrix. The remaining options are now: natural experiment, difference-in-differences and pre-post trials.
The next step in the Matrix is to judge the quality of the data that can be expected from the remaining options.

6. The evaluation team sees that Difference-in-differences studies have the best quality level among the remaining options. This is because difference-in-differences combines natural experiments with repeated measures.

7. They consult their evaluation expert and suggest using a difference-in-differences. Now the expert can provide further advice.

A real-world intervention similar to the scenario sketched out above is described in Box 7.

BOX 7

Does funding for neighbourhood improvement increase physical activity?

A similar intervention to the one described in Illustrative example 1 was designed and evaluated in the United States *(50)*. One neighbourhood received funding for a better, more stimulating environment that should increase physical activity among participants; another neighbourhood was chosen as a control group. The BMI of participants was measured before and after the intervention. In addition, the researchers measured daily minutes of physical activity with wearable devices that participants were asked to wear on their nondominant wrist for seven consecutive (24-hour) days. The researchers used a difference-in-differences approach to evaluate the impact of the intervention. They found a significant decrease in BMI in the treatment neighbourhood. However, the same trend was observed in the control neighbourhood, and the difference-in-differences revealed no evidence for a stronger effect in the treatment neighbourhood than in the control. Moreover, the analyses revealed no advantage in terms of daily minutes of physical activity due to the intervention. This impact evaluation highlights the importance of a proper control group. If the researchers had chosen a pre-post trial rather than the difference-in-differences approach, they may have come to the conclusion that the intervention was promising in reducing BMI among participants. Seeing that the same trend was observable in the control neighbourhood casts doubt on the idea that the intervention causes benefits.

Illustrative example 2. All designs are possible

Scenario: Let us consider the same scenario as in the previous case example – a fictitious neighbourhood that receives financial funding from the State to provide more opportunities for the 1000 residents to engage in physical activity. This time the State agrees to a lottery system. All neighbourhoods participate and the funding is randomly allocated to one neighbourhood. Moreover, the funding is intended for specific facilities (such as sports equipment in the park) and will be removed after one year and allocated to another neighbourhood. These changes to the scenario mean that the impact evaluation team can now also conduct RCTs (random allocation is possible), crossover trials (one neighbourhood receives the intervention, and after one year the control and the treatment groups change over), or an interrupted time series study (the intervention is removed). In this case, no options are excluded using steps 1-5 of the Decision Matrix. Step 6 of the Matrix, focused on the strength of design, provides some decision aid in this case.

1. The evaluation team sees that RCTs and difference-in-differences provide strong data quality. So these designs should be preferred.

2. The evaluation team read that crossover trials are usually more efficient than RCTs and difference-in-differences but that they are only appropriate if the intervention can be undone (this often rules them out in the case of educational interventions). Opportunities for physical activity can be removed, so they think that crossover trials may be the best option. However, they are unsure about this.

3. To discuss their rationale, they contact the evaluation expert and suggest a difference-in-differences, an RCT or perhaps a crossover trial.

Tool 4. Improving quality of evaluation

Selecting a sample size

There are several approaches to sample size selection *(51)*. If measuring the entire population is not possible and resources are not especially limited, then so-called "a priori statistical power analyses" are a recommended approach to determine sample sizes. When conducting power analyses, the number of participants that is recommended for a quantitative impact evaluation depends on the desired statistical power of the evaluation – that is, the probability of detecting an impact if the impact exists. An "underpowered" study is one where the sample size is not big enough to detect an impact of the intervention even if it exists. It is common in many BCI fields to aim for a statistical power of at least 80–90%. The actual sample size required for a specific power (say, 80%) can vary greatly across impact evaluation designs and even within designs. One of the main reasons for this variation is the different effect sizes of BCI-informed interventions: large effect sizes require fewer participants in order to statistically detect an effect. Basically, the larger you expect the effect of an intervention to be, the smaller the required sample size. (For an example showing how important it is to get the right sample size, see Box 8.)

Conducting an a priori power analysis can save resources and reduce the probability of inconclusive results from your impact evaluation if the study is underpowered (the sample is too small to detect an effect). An evaluation expert can conduct power calculations and determine an appropriate sample size, given a particular (fixed) effect size.

In the absence of a more specific smallest effect size of interest, meta-analyses[11] and systematic reviews can help to gain a sense of a reasonable expected effect size for a specific BCI-informed intervention. Moreover, there are several freely available tools that can be used for quick power and sample size calculations for simple impact evaluation designs *(54,55)*.[12]

BOX 8

Correctly matching sample size and effect size

Suppose that you have chosen an RCT as the preferred research design for your impact evaluation. You aim to compare a group that is exposed to your BCI-informed intervention (say, an anti-smoking intervention) with a group that is not exposed to the intervention. You aim to reach a statistical power of 80% for your analysis. Now, a meta-analysis states that the impact of health interventions on oral health behaviours is $r = .13$, while for smoking $r = .05$ *(52)*.[a] The required sample size can vary between $n = 460$ (for oral health behaviours) and $n = 962$ (for smoking) because of the difference in expected effect sizes. Now, if you chose the sample size for oral health behaviours (only 460 participants instead of 962), the statistical power to detect an effect for your smoking intervention is only around 50%. This means there is a high probability that a nonsignificant result of your impact evaluation may simply reflect that the power was too low to detect an effect – the intervention worked, but you could not see it because of the small sample size. Thus, investing in a decent sample size and conducting a power analysis can greatly increase the meaningfulness of your impact evaluation.

[a] For an explanation of meta-analysis, see the Glossary; for further information on correlation coefficient *(r)*, see Asuero, Sayago & González *(53)*.

[11] For an explanation of meta-analysis, see the Glossary.

[12] Power calculation tools available online free of charge include G*Power and Superpower.

Preregistering a study

It is recommended that you preregister your impact evaluation. Preregistration involves "defin[ing] the research questions and analysis plan before observing the research outcomes" *(56)*. This procedure has several advantages for impact evaluations *(56–58)*. First, preregistration encourages researchers to report all results of the analysis plan and not to ignore results they consider irrelevant after seeing them (see also below, "Sharing stories of failure and inconclusive results"). Second, by using preregistration, researchers can avoid potential accusations that a result of an impact evaluation is based on questionable research practices (such as adjusting your hypothesis after the results are known *(59)*). Thus, preregistration increases the perceived reliability of reported results and can help to convince other researchers and policy-makers that an impact evaluation is of high quality. Questionable research practices are often the result not of intentionally inappropriate behaviour but of unintended errors. No one is immune to such errors, so preregistration is considered a valuable investment for any impact evaluation.

You can preregister your impact evaluation on one of several preregistration platforms, such as OSF *(60)* and AsPredicted *(61)*. You can also publish your analysis plan as a study protocol in journals such as Trials *(62)* and Pilot and Feasibility Studies *(63)*.

Following ethical standards

It is important that all BCI-informed impact evaluation studies follow nationally and internationally acknowledged ethical principles such as the Declaration of Helsinki *(64)*. Thus, it is recommended that you consult your institutional review board or an external review board before starting an impact evaluation. The WHO Guide to tailoring health programmes offers guidance on ethical approval and what to include in a research protocol for a BCI-related study *(1)*.

An ethical review board can provide feedback on whether the design of an impact evaluation raises any ethical concerns. Some issues that are often raised include:

- **Debriefing.** Every participant in a study should have the opportunity to learn about the study goals and to raise questions after completion of the study.
- **Data protection.** Personal data should be treated confidentially and participants' fundamental rights strengthened by complying with data security standards such as the General Data Protection Regulation (GDPR) *(65)*.
- **Informed consent.** Every participant should be informed about the risks and benefits of participating in an impact evaluation study and consent should be obtained. WHO provides templates for informed consent forms *(66)*.
- **Safety.** All participants should be protected from physical or psychological harm, and they should have the opportunity to refuse participation or quit from the study at any time.
- **Equity.** All members of the target group should have equal opportunity to participate in the study. This ensures that all those affected by the research can have their say.

Sharing stories of failure or inconclusive results

Inconclusive or statistically nonsignificant results are often not reported as they may be considered uninteresting or irrelevant. This can lead to what is known as publication bias *(67,68)*, which is "any tendency on the parts of investigators or editors to fail to publish study results on the basis of the direction or strength of the study findings" *(69)*. Such selective publication of results can have severe consequences. For example, if a BCI-informed intervention was successful in two countries but failed in 10 others, and only the successful results are reported, public health implementers might incorrectly conclude that the intervention is highly effective and recommend it despite the mixed pattern of results. It is just as important to share stories of failure as it is to share stories of success. Other teams that aim to design and evaluate a BCI-informed intervention can learn from inconclusive or nonsignificant results, and discussing different results across countries may even lead to new insights into the conditions under which a BCI-informed intervention works and under which it fails. In that sense, impact evaluation and discussion of evaluation results should be understood as a collective endeavour *(70)*.

Inconclusive or nonsignificant results can be published in public reports, but there is also a growing number of scientific peer-reviewed journals (such as the *Journal of Trial and Error (71)*) that encourage publication of unsuccessful attempts and experiences, or even focus specifically on them. Publishing results in these indexed journals can be particularly useful for other researchers when conducting meta-analysis or literature reviews.

Glossary

Causality. In the context of BCI-informed interventions, the term "causes" denotes that a given intervention leads to a desired outcome (for instance, more people quit smoking, go for vaccination or follow their treatment plan). In other words, the intervention is responsible for the occurrence of the outcome. For an intervention to be the cause of a desired outcome, three conditions generally need to be met:

1. The intervention needs to precede the desired outcome.
2. If the intervention is present, the desired outcome happens.
3. If the intervention is absent, the desired outcome does not happen.

These conditions are usually evaluated in RCTs. The key feature of RCTs is that all other influences and potential causes of the desired outcome, except for the intervention itself, are held constant. This design ensures that the only thing that can explain any observed change in the outcome is the intervention. All three conditions are necessary to conclude that an intervention caused a behaviour, and if one of them is not confirmed, we cannot conclude that the BCI-informed intervention caused the desired outcome. For example, if researchers find that vaccine uptake is higher in region A, where a BCI-informed intervention was administered, than in region B, where the intervention was absent, this seems to fulfil conditions 2 and 3. However, without knowledge of vaccine uptake rates in both regions before the intervention, we cannot confirm condition 1. For instance, the desired outcome may have been present before the intervention (region A always had higher uptake rates than region B), and thus the intervention did not cause this result. If the presence of a BCI-informed intervention and a desired outcome seem to be related in some way but causality cannot (yet) be concluded, researchers usually use the term "correlation". As shown by the example, correlation does not imply causation.

Further reading on correlation and causation:
Rohrer (2018) *(72).*

Confounder. A variable (such as gender, age, education, attitudes towards the outcome, contextual changes, or the presence of other interventions or campaigns) that can influence the outcome measure but is not part of the BCI-informed intervention. Confounders can lead to false conclusions about the impact of a BCI-informed intervention. Carefully designed RCTs are a way to limit the impact of confounders on the evaluation results and to determine whether an effect is really caused by a BCI-informed intervention.

Further reading on confounders:
VanderWeele & Shpitser (2013) *(73).*

Counterfactual. This term refers to hypothetical outcomes under "business as usual" – that is, if participants had not been exposed to the BCI-informed intervention. In RCTs, this is mimicked by the introduction of a control group. However, even where there is no control group, the counterfactual can be mimicked based on prior data of the outcome measure. This is done, for example, in an evaluation design called interrupted time series.

Further reading on counterfactuals:
Höfler (2005) *(74).*

Cost-effectiveness analysis. This kind of analysis compares the costs of a BCI-informed intervention with its health-related effects. The result is often presented as a ratio of net costs per one unit of outcome. The outcome units of cost-effectiveness analyses in BCI contexts vary. Some studies use the ratio of net costs per life year saved; others use net costs per case of disease prevented; and still others use net costs per death averted. The general idea is always to assess the economic impact of a BCI-informed intervention. This information can complement the results from impact evaluations that do not usually focus on financial costs as an outcome measure.

Further reading on cost-effectiveness analysis:
Murray et al. (2000) *(75).*

Meta-analysis. In this kind of analysis, results from several individual studies are summarized to assess the overall impact of a BCI-informed intervention. A common result is an aggregated effect size of the BCI-informed intervention across all impact evaluations. It is also common to report the heterogeneity of study results. Heterogeneity measures reveal whether results across studies vary a lot or are consistent. If the heterogeneity is high, this could indicate the presence of a relevant moderator variable for the impact of BCI-informed interventions (for instance, the age of participants). Studies for a meta-analysis are usually identified via systematic reviews.

Further reading on meta-analysis:
Borenstein et al. (2021) *(76).*

Outcome measure. Any measure that is used to assess the effect of a BCI-informed intervention for an impact evaluation. These measures are also referred to as “endpoint measures”. Outcome measures can range from outputs (such as self-reported beliefs, experiences and emotions), to behavioural outcomes (such as healthy food intake) and health outcomes (such as absence of CVD). It is often useful not to rely entirely on one type of measure to assess the impact of an intervention. Self-reports can deviate from actual behaviour, while relying only on observational measures may miss individual evaluations of an intervention that can be crucial for its long-term success.

Further reading on self-reported measures:
Newell et al. (1999) *(77).*

References*

1. A guide to tailoring health programmes: using behavioural and cultural insights to tailor health policies, services and communications to the needs and circumstances of people and communities. Copenhagen: WHO Regional Office for Europe; 2023 (https://iris.who.int/handle/10665/367041).

2. Guide to evaluating behaviourally and culturally informed health interventions in complex settings. Copenhagen: WHO Regional Office for Europe; 2022 (https://iris.who.int/handle/10665/362317).

3. Hornik R, Jacobsohn L, Orwin R, Piesse A, Kalton G. Effects of the National Youth Anti-Drug Media Campaign on youths. Am J Public Health. 2008;98(12):2229–36 (https://doi.org/10.2105/AJPH.2007.125849).

4. de Walque D, Chukwuma A, Ayivi-Guedehoussou N, Koshkakaryan M. Invitations, incentives, and conditions: a randomized evaluation of demand-side interventions for health screenings. Soc Sci Med. 2022;296:114763 (https://doi.org/10.1016/j.socscimed.2022.114763).

5. Schuit AJ, Wendel-Vos GCW, Verschuren WM, Ronckers ET, Ament A, Van Assema P et al. Effect of 5-year community intervention Hartslag Limburg on cardiovascular risk factors. Am J Prev Med. 2006;30:237–42 (https://doi.org/10.1016/j.amepre.2005.10.020).

6. Ronckers ET, Groot W, Steenbakkers M, Ruland E, Ament A. Costs of the "Hartslag Limburg" community heart health intervention. BMC Public Health. 2006;6:51 (https://doi.org/10.1186/1471-2458-6-51).

7. Bemelmans W, van Baal P, Wendel-Vos GCW, Schuit AJ, Feskens E, Ament A et al. The costs, effects and cost-effectiveness of counteracting overweight on a population level: a scientific base for policy targets for the Dutch national plan for action. Prev Med. 2008;46:127–32 (https://doi.org/10.1016/j.ypmed.2007.07.029).

8. Ringold DJ. Boomerang effects in response to public health interventions: some unintended consequences in the alcoholic beverage market. J Consum Policy. 2002;25:27–63 (https://doi.org/10.1023/A:1014588126336).

9. Cho H, Salmon C. Unintended effects of health communication campaigns. J Commun. 2007;57:293–317 (https://doi.org/10.1111/j.1460-2466.2007.00344.x).

10. Siddiqui FA, Padhani ZA, Salam RA, Aliani R, Lassi ZS, Das JK et al. Interventions to improve immunization coverage among children and adolescents: a meta-analysis. Pediatrics. 2022;149:e2021053852D (https://doi.org/10.1542/peds.2021-053852D).

11. Sendelbach S, Funk M. Alarm fatigue: a patient safety concern. AACN Adv Crit Care. 2013;24(4):378–86 (https://doi.org/10.4037/NCI.0b013e3182a903f9).

12. Lorenc T, Petticrew M, Welch V, Tugwell P. What types of interventions generate inequalities? Evidence from systematic reviews. J Epidemiol Community Health 2013;67(2):190–3 (https://doi.org/10.1136/jech-2012-201257).

* All references were accessed on 28 December 2023

13. Basol M-S, Roozenbeek J, Berriche M, Uenal F, McClanahan WP, Linden SVD. Towards psychological herd immunity: cross-cultural evidence for two prebunking interventions against COVID-19 misinformation. Big Data Soc. 2021; 8:205395172110138 (https://doi.org/10.1177/20539517211013868).

14. Roozenbeek J, van der Linden S. Fake news game confers psychological resistance against online misinformation. Palgrave Commun. 2019;5:65 (https://doi.org/10.1057/s41599-019-0279-9).

15. Harjani T, Basol M-S, Roozenbeek J, van der Linden S. Gamified inoculation against misinformation in India: a randomized control trial. Journal of Trial and Error (JOTE). 2023 (https://doi.org/10.36850/e12).

16. Kozyreva A, Smillie L, Lewandowsky S. Incorporating psychological science into policy making: the case of misinformation. Eur Psychol. 2023;28:206–24 (https://doi.org/10.1027/1016-9040/a000493)

17. Adelle C, Weiland S. Policy assessment: the state of the art. Impact Assess Proj Apprais. 2012;30:25–33 (https://doi.org/10.1080/14615517.2012.663256).

18. Public support for health policies: why it matters and how it can be maximized. Copenhagen: WHO Regional Office for Europe; 2024 (https://iris.who.int/handle/10665/376735).

19. Richardson WS, Wilson MC, Nishikawa J, Hayward RS. The well-built clinical question: a key to evidence-based decisions. ACP J Club 1995;123(3):A12–13. PMID: 7582737.

20. Bartoš V, Bauer M, Cahlíková J, Chytilová J. Communicating doctors' consensus persistently increases COVID-19 vaccinations. Nature. 2022;606:542–9 (https://doi.org/10.1038/s41586-022-04805-y).

21. Strassmann A, Çolak Y, Serra-Burriel M, Nordestgaard BG, Turk A, Afzal S et al. Nationwide indoor smoking ban and impact on smoking behaviour and lung function: a two-population natural experiment. Thorax. 2023;78(2):144–50 (https://doi.org/10.1136/thoraxjnl-2021-218436).

22. Peters DH, Adam T, Alonge O, Agyepong IA, Tran N. Republished research: implementation research: what it is and how to do it. Br J Sports Med. 2014;48(8):731–6 (https://doi.org/10.1136/bmj.f6753).

23. Hamilton AB, Finley EP. Qualitative methods in implementation research: an introduction. Psychiatry Res. 2019;280:112516 (https://doi.org/10.1016/j.psychres.2019.112516).

24. Brown CH, Curran G, Palinkas LA, Aarons GA, Wells KB, Jones L et al. An overview of research and evaluation designs for dissemination and implementation. Annu Rev Public Health. 2017;38:1–22 (https://doi.org/10.1146/annurev-publhealth-031816-044215).

25. Horne PJ, Tapper K, Lowe CF, Hardman CA, Jackson MC, Woolner J. Increasing children's fruit and vegetable consumption: a peer-modelling and rewards-based intervention. Eur J Clin Nutr. 2004;58:1649–60 (https://doi.org/10.1038/sj.ejcn.1602024).

26. Morrill BA, Madden GJ, Wengreen HJ, Fargo JD, Aguilar SS. A randomized controlled trial of the Food Dudes program: tangible rewards are more effective than social rewards for increasing short- and long-term fruit and vegetable consumption. J Acad Nutr Diet. 2016;116(4):618–29 (https://doi.org/10.1016/j.jand.2015.07.001).

27. Jones BA, Madden GJ, Wengreen HJ. The FIT Game: preliminary evaluation of a gamification approach to increasing fruit and vegetable consumption in school. Prev Med. 2014;68:76–9 (https://doi.org/10.1016/j.ypmed.2014.04.015).

28. Joyner D, Wengreen HJ, Aguilar SS, Spruance LA, Morrill BA, Madden GJ. The FIT Game III: reducing the operating expenses of a game-based approach to increasing healthy eating in elementary schools. Games Health J. 2017;6(2):111–18 (https://doi.org/10.1089/g4h.2016.0096).

29. Ramsey CA, Hewitt AD. A methodology for assessing sample representativeness. Environ Forensics. 2005;6:71–5 (https://doi.org/10.1080/15275920590913877).

30. Kitzinger J. The methodology of focus groups: the importance of interaction between research participants. Sociol Health Illness. 1994;16:103–21 (https://doi.org/10.1111/1467-9566.ep11347023).

31. DiCicco-Bloom B, Crabtree BF. The qualitative research interview. Med Educ. 2006;40(4):314–21. https://doi.org/10.1111/j.1365-2929.2006.02418.x.

32. Farrelly MC, Healton CG, Davis KC, Messeri P, Hersey JC, Haviland ML. Getting to the truth: evaluating national tobacco countermarketing campaigns. Am J Public Health. 2002;92:901–7 (https://doi.org/10.2105/AJPH.92.6.901).

33. Wolburg JM. College students' responses to antismoking messages: denial, defiance, and other boomerang effects. J Consum Affairs. 2006;40:294–323 (https://doi.org/10.1111/j.1745-6606.2006.00059.x).

34. Rios LP, Ye C, Thabane L. Association between framing of the research question using the PICOT format and reporting quality of randomized controlled trials. BMC Med Res Methodol. 2010;10:11 (https://doi.org/10.1186/1471-2288-10-11).

35. Hariton E, Locascio JJ. Randomised controlled trials: the gold standard for effectiveness research. BJOG. 2018;125(13):1716 (https://doi.org/10.1111/1471-0528.15199).

36. Altman DG, Schulz KF, Moher D, Egger M, Davidoff F, Elbourne D et al. The revised CONSORT statement for reporting randomized trials: explanation and elaboration. Ann Intern Med. 2001;134(8):663–94 (https://doi.org/10.7326/0003-4819-134-8-200104170-00012).

37. Lewis SC, Warlow CP. How to spot bias and other potential problems in randomised controlled trials. J Neurol Neurosurg Psychiatry. 2004;75(2):181–7 (https://doi.org/10.1136/jnnp.2003.025833).

38. Dron L, Taljaard M, Cheung YB, Grais R, Ford N, Thorlund K et al. The role and challenges of cluster randomised trials for global health. Lancet Glob Health. 2021;9:e701–10 (https://doi.org/10.1016/S2214-109X(20)30541-6).

39. Sibbald B, Roberts C. Understanding controlled trials: crossover trials. BMJ. 1998;316:1719–20 (https://doi.org/10.1136/bmj.316.7146.1719).

40. Hemming K, Haines TP, Chilton PJ, Girling AJ, Lilford RJ. The stepped wedge cluster randomised trial: rationale, design, analysis, and reporting. BMJ. 2015;350:h391 (https://doi.org/10.1136/bmj.h391).

41. Hargreaves JR, Copas AJ, Beard E, Osrin D, Lewis JJ, Davey C et al. Five questions to consider before conducting a stepped wedge trial. Trials. 2015;16:350 (https://doi.org/10.1186/s13063-015-0841-8).

42. Cartwright N. What are randomised controlled trials good for? Philos Stud. 2010;147:59–70 (https://doi.org/10.1007/s11098-009-9450-2).

43. Bonell C, Fletcher A, Morton M, Lorenc T, Moore L. Realist randomised controlled trials: a new approach to evaluating complex public health interventions. Soc Sci Med. 2012; 75:2299–306 (https://doi.org/10.1016/j.socscimed.2012.08.032).

44. Abadie A, Diamond A, Hainmueller J. Synthetic control methods for comparative case studies: estimating the effect of California's tobacco control program. J Am Stat Assoc. 2010; 105:493–505 (https://doi.org/10.1198/jasa.2009.ap08746).

45. Craig P, Katikireddi SV, Leyland A, Popham F. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. Annu Rev Public Health. 2017;38:39–56 (https://doi.org/10.1146/annurev-publhealth-031816-044327).

46. Wing C, Simon K, Bello-Gomez RA. Designing difference in difference studies: best practices for public health policy research. Annu Rev Public Health. 2018;39:453–69 (https://doi.org/10.1146/annurev-publhealth-040617-013507).

47. Turner SL, Karahalios A, Forbes AB, Taljaard M, Grimshaw JM, McKenzie JE. Comparison of six statistical methods for interrupted time series studies: empirical evaluation of 190 published series. BMC Med Res Methodol. 2021;21:134 (https://doi.org/10.1186/s12874-021-01306-w).

48. de Vocht F, Katikireddi SV, McQuire C, Tilling K, Hickman M, Craig P. Conceptualising natural and quasi experiments in public health. BMC Med Res Methodol. 2021;21:32 (https://doi.org/10.1186/s12874-021-01224-x).

49. Dunning T. Natural experiments in the social sciences: a design-based approach. Cambridge: Cambridge University Press; 2012.

50. Dubowitz T, Ghosh Dastidar M, Richardson AS, Colabianchi N, Beckman R, Hunter GP et al. Results from a natural experiment: initial neighbourhood investments do not change objectively-assessed physical activity, psychological distress or perceptions of the neighbourhood. Int J Behav Nutr Phys Act. 2019;16(1):29 (https://doi.org/10.1186/s12966-019-0793-6).

51. Lakens D. Sample size justification. Collabra: Psychology. 2022;8(1):33267 (https://doi.org/10.1525/collabra.33267).

52. Snyder LB, Hamilton MA, Mitchell EW, Kiwanuka-Tondo J, Fleming-Milici F, Proctor D. A meta-analysis of the effect of mediated health communication campaigns on behavior change in the United States. J Health Commun. 2004;9:71–96 (https://doi.org/10.1080/10810730490271548).

53. Asuero AG, Sayago A, González AG. The correlation coefficient: an overview. Crit Rev Anal Chem. 2006;36:41–59 (https://doi.org/10.1080/10408340500526766).

54. Lakens D, Caldwell AR. Simulation-based power analysis for factorial analysis of variance designs. Adv Methods Pract Psychol Sci. 2021;4:251524592095150 (https://doi.org/10.1177/2515245920951503).

55. Faul F, Erdfelder E, Lang A-G, Buchner A. G*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. Behav Res Methods. 2007;39(2):175–91 (https://doi.org/10.3758/BF03193146).

56. Nosek BA, Ebersole CR, DeHaven AC, Mellor DT. The preregistration revolution. Proc Natl Acad Sci USA. 2018;115:2600–6 (https://doi.org/10.1073/pnas.1708274114).

57. Nosek BA, Beck ED, Campbell L, Flake JK, Hardwicke TE, Mellor DT et al. Preregistration is hard, and worthwhile. Trends Cogn Sci. 2019;23:815–18 (https://doi.org/10.1016/j.tics.2019.07.009).

58. Wagenmakers E-J, Wetzels R, Borsboom D, Van Der Maas HLJ, Kievit RA. An agenda for purely confirmatory research. Perspect Psychol Sci. 2012;7:632–8 (https://doi.org/10.1177/1745691612463078).

59. Kerr NL. HARKing: hypothesizing after the results are known. Pers Soc Psychol Rev. 1998;2:196–217 (https://doi.org/10.1207/s15327957pspr0203_4).

60. Center for Open Science. OSF Support [online platform] (https://help.osf.io/).

61. Wharton School of the University of Pennsylvania, Wharton Credibiltiy Lab. AsPredicted [online platform] (https://aspredicted.org/).

62. Trials. BioMed Central. Study Protocol; 2023 (https://trialsjournal.biomedcentral.com/submission-guidelines/preparing-your-manuscript/study-protocol).

63. Pilot and Feasibility Studies. BioMed Central ; 2023 (https://pilotfeasibilitystudies.biomedcentral.com/).

64. Williams J. The Declaration of Helsinki and public health. Bull World Health Organ. 2008;86:650–1 (https://doi.org/10.2471/BLT.08.050955).

65. Data protection in the EU. Brussels: European Commission; 2023 (https://commission.europa.eu/law/law-topic/data-protection/data-protection-eu_en).

66. Templates for informed consent forms. Research Ethics Review Committee. Geneva: World Health Organization; n.d. (https://www.who.int/groups/research-ethics-review-committee/guidelines-on-submitting-research-proposals-for-ethics-review/templates-for-informed-consent-forms).

67. DeVito NJ, Goldacre B. Catalogue of bias: publication bias. BMJ Evid Based Med. 2019;24(2):53–4 (https://doi.org/10.1136/bmjebm-2018-111107).

68. Francis G. Publication bias and the failure of replication in experimental psychology. Psychon Bull Rev. 2012;19(6):975–91 (https://doi.org/10.3758/s13423-012-0322-y).

69. Dickersin K, Min Y-I. Publication bias: the problem that won't go away. Ann NY Acad Sci. 1993;703:135–48 (https://doi.org/10.1111/j.1749-6632.1993.tb26343.x).

70. Holford D, Fasce A, Tapper K, Demko M, Lewandowsky S, Hahn U et al. Science communication as a collective intelligence endeavor: a manifesto and examples for implementation. Sci Commun. 2023;45(4):539–54 (https://doi.org/10.1177/10755470231162634).

71. Devine S, Bautista-Perpinya M, Delrue V, Gaillard S, Jorna T, van der Meer M et al. Science fails: let's publish. Journal of Trial and Error (JOTE). 2020; 1:1–5 (https://doi.org/10.36850/ed1).

72. Rohrer JM. Thinking clearly about correlations and causation: graphical causal models for observational data. Adv Methods Pract Psychol Sci. 2018; 1:27–42 (https://doi.org/10.1177/2515245917745629).

73. VanderWeele TJ, Shpitser I. On the definition of a confounder. Ann Stat. 2013;41(1):196–220 (https://doi.org/10.1214/12-AOS1058).

74. Höfler M. Causal inference based on counterfactuals. BMC Med Res Methodol. 2005; 5:28 (https://doi.org/10.1186/1471-2288-5-28).

75. Murray CJ, Evans DB, Acharya A, Baltussen RM. Development of WHO guidelines on generalized cost-effectiveness analysis. Health Econ. 2000;9(3):235–51 (https://doi.org/10.1002/(SICI)1099-1050(200004)9:3<235::AID-HEC502>3.0.CO;2-O).

76. Borenstein M, Hedges LV, Higgins JPT, Rothstein HR. Introduction to meta-analysis. 2nd edition. Hoboken (NJ): Wiley; 2021.

77. Newell SA, Girgis A, Sanson-Fisher RW, Savolainen NJ. The accuracy of self-reported health behaviors and risk factors relating to cancer and cardiovascular disease in the general population. Am J Prev Med. 1999; 17:211–29 (https://doi.org/10.1016/S0749-3797(99)00069-0).

The WHO Regional Office for Europe

The World Health Organization (WHO) is a specialized agency of the United Nations created in 1948 with the primary responsibility for international health matters and public health. The WHO Regional Office for Europe is one of six regional offices throughout the world, each with its own programme geared to the particular health conditions of the countries it serves.

Document number: WHO/EURO:2024-10200-49972-75147 (PDF)
WHO/EURO:2024-10200-49972-75181 (print)

Member States

Albania
Andorra
Armenia
Austria
Azerbaijan
Belarus
Belgium
Bosnia and Herzegovina
Bulgaria
Croatia
Cyprus
Czechia
Denmark
Estonia
Finland
France
Georgia
Germany
Greece
Hungary
Iceland
Ireland
Israel
Italy
Kazakhstan
Kyrgyzstan
Latvia
Lithuania
Luxembourg
Malta
Monaco
Montenegro
Netherlands (Kingdom of the)
North Macedonia
Norway
Poland
Portugal
Republic of Moldova
Romania
Russian Federation
San Marino
Serbia
Slovakia
Slovenia
Spain
Sweden
Switzerland
Tajikistan
Türkiye
Turkmenistan
Ukraine
United Kingdom
Uzbekistan

World Health Organization
Regional Office for Europe
UN City, Marmorvej 51,
DK-2100 Copenhagen Ø, Denmark
Tel.: +45 45 33 70 00
Fax: +45 45 33 70 01
Email: euinsights@who.int
Website: www.who.int/europe