**Updating Natural Experimental Evaluations Guidance – online consultation**

**Overview of the project**

Welcome to this online consultation to update guidance for the use of natural experimental evaluations.

We are conducting an online consultation of stakeholders experienced in undertaking, funding and publishing research of natural experimental evaluations. Updated guidance is being developed on using natural experiments to evaluate population health interventions, which will provide a single, integrated, up-to-date guide to the use of natural experimental methods to evaluate population health interventions. Since original MRC guidance on natural experiments was published in 2012, new methods have been developed, there has been an increase in the use of natural experimental evaluations, the availability of very large routinely collected data resources, and the application of systems thinking to complex health interventions. This has implications for how natural experimental evaluations are designed and conducted. There is no up-to-date comprehensive overview of all the significant developments in methods, infrastructure or governance, and some key concepts and definitions are still debated. The rise in use of natural experimental approaches means it is important that publishers, funders, and users of intervention research are aware of all the recent developments and of the strengths, weaknesses, applicability and limitations of the full range of methods now available.

This consultation will be open for 30 days. We anticipate that participation will take approximately 1 hour of your time.

You will be asked to provide your expert opinion on content that may be included in guidance for natural experimental evaluations and explain your opinion using open-text boxes.

You may complete the consultation questions in any order, by clicking through the pages you wish to work on later. You can revise your responses at any time while the consultation is open. To move between pages, please click on the navigation boxes located at the top and at the bottom of each page. Once you answer all questions and are on the last page, please click the “Finish” button.

More information about the updating natural experimental evaluations guidance project and participating in this online consultation is available in the participant information sheet [*weblink to pdf Participant Information sheet*].

Thank you

Peter Craig

On behalf of Prof Frank de Vocht, University of Bristol; Prof Martin White and Dr David Ogilvie, University of Cambridge; Prof Judith Green, University of Exeter; Prof Jim Lewsey, Dr Manuela Deidda, Prof Ruth Dundas, Prof Vittal Katikireddi and Ms Mhairi Campbell, University of Glasgow.

**Consultation instructions**

You are invited to provide feedback on the guidance for natural experimental evaluations. We invite you to complete this online consultation by reading the draft guidance and providing feedback (approx. 1 hour)

Please check the box below to acknowledge your consent to taking part in this consultation

I consent to my anonymised data being stored and used in future research 🞎

I consent to take part in this consultation 🞎

**About you**

We would like to gather some broad demographic information about you to help us understand the general background of the participants of this consultation.

Please indicate which of the following describes your role/expertise (choose all that apply):

Member of research funding board 🞏

Representative of research funder 🞏

Journal editor 🞏

Intervention researcher 🞏

Quantitative researcher 🞏

Qualitative researcher 🞏

Policymaker 🞏

Practitioner 🞏

Clinician 🞏

Other (please describe) 🞏

[free text box]

Please indicate which of the following describes your place(s) of employment (more than one may apply):

University 🞏

Public sector organisation (e.g. government department, research council, health service) 🞏

Non-profit organisation (e.g. NGO, charity) 🞏

For-profit organisation (e.g. consultancy) 🞏

Other (please describe) 🞏

[free text box]

I consent to inclusion of my name (and affiliation) in a list of participants acknowledged for contributing expert knowledge in the acknowledgements section of articles and conference presentations reporting results from this project. 🞎 **(optional)**

Contact information (optional)

Name [free text box]

Organisation [free text box]

**Guidance section: Concepts and definitions**

|  |
| --- |
| **Concepts and definitions summary points**   * We define natural experiments as events or processes outside the control of a researcher that divide a population into exposed or unexposed subpopulations. * A natural experimental evaluation uses an event or process associated with the introduction, delivery or withdrawal of an intervention to evaluate the impact of the intervention. * Natural experimental studies are distinguished from the broader range of observational studies by their focus on a specific event or process that determines exposure, rather than by their use of one or more of a prescribed range of methods, or by the extent to which the assignment process approximates randomisation. * Methods used in the evaluation of natural experiments originated in a range of disciplines, leading to the use of different terms for similar concepts, so we provide a glossary defining key terms as they are used in this guidance. |

**Please continue to read the full text of this section, respond to questions, and provide comments on this section or, if you prefer, you can move to the next section [*move to next section*]**

**Concepts and definitions**

The value of natural experiments is that they widen the range of interventions that can be evaluated beyond those that are amenable to experimental manipulation. They provide opportunities to evaluate outcomes that are by-products rather than the main purpose of the intervention and which would not provide a persuasive or ethically sound basis for experimental manipulation.(1) They can be studied retrospectively and used to evaluate long-term outcomes that might be impractical to include in a prospective study. And they can be used to evaluate very large scale or irreversible interventions, such as national policy changes or large public infrastructure investments, where prior political commitments make experimentation unattractive to decision-makers.

For this updated guidance, as in the first version of the MRC guidance on natural experiments,(2) we adopt a broad definition of natural experiments as events or processes outside the control of a researcher that divide a population into exposed or unexposed subpopulations. A natural experimental evaluation uses an event or process associated with the introduction, delivery or (more rarely) withdrawal of an intervention to evaluate the impact of the intervention on some outcome or range of outcomes. Evidence from such studies can also be accumulated to evaluate questions about more general exposures (such as the effect of income or income inequality on health), but what distinguishes a natural experimental study from the broader range of observational studies is its focus on a specific event or process that determines exposure.

This broad definition contrasts with other approaches that seek to define natural experimental studies in terms of the use of one or more of a prescribed range of methods, such as those that can address unobserved confounding,(3) or that satisfy some other criterion such as ‘as if randomisation.’(4) We question the practical value and applicability of such restrictions. Design labels are an inadequate proxy for study quality, which depends critically on the extent to which assumptions are tested, threats to validity evaluated and robustness checks performed. Lists of ‘approved’ methods can rapidly become dated as new methods are developed and existing ones refined. Whether a method can be applied satisfactorily in a particular case depends as much on the specific details of the process determining exposure to the intervention as it does on the general properties of the method, and a good qualitative understanding of this ‘assignment process’ is vital to the appropriate choice and application of quantitative methods of effect estimation.

The extent to which an assignment process approximates randomisation is a useful starting point for assessing risk of bias due to selective exposure to an intervention, but ’as if randomisation’ is hard to define precisely. It can only be tested by comparing exposed and unexposed groups in terms of observed confounders, whereas the key problem is whether the groups are similar on both observed and unobserved confounders. We suggest that it is helpful to think of ‘as if randomisation’ as defining one end of a spectrum along which natural experimental studies lie, rather than a criterion that distinguishes them from other kinds of observational study. We recommend the use of a target trial approach (see Table 1) for a full assessment of risk of bias in a natural experimental study, (de Vocht et al 5, table 1) rather than reliance either on a single criterion such as ‘as if randomness’, or on the use of one or more of a prescribed range of methods.

Methods used in the evaluation of natural experiments originated in a range of disciplines, including economics, sociology, political science and epidemiology. This has led to the use of different terms for similar concepts. Table 1 defines key terms as they are used in this guidance.

**Table 1 Glossary**

|  |  |
| --- | --- |
| Term | Usage |
|  |  |
|  |  |
| Assignment, allocation  Causal estimand  Concurrent interventions (or co-interventions) | The process that determines whether units (which may be individual people, or aggregates of some kind) are exposed to the intervention that is being studied. Also referred to as the data generating process.  See treatment effect.  A common source of confounding in natural experimental evaluations is when one or more interventions other than the intervention being studied are implemented at the same time. Depending on the extent of overlap in exposure, effects can be disentangled, for example by focusing on outcomes specific to the intervention of interest, or on populations exposed to one intervention but not the other(s), see Matthay et al (6). |
| Confounding, confounder (observed, unobserved) | Confounding refers to the mixing up of the effects of exposure to the intervention with the effects of characteristics associated with exposure. Confounding by observed characteristics (i.e., those on which data are available) can be addressed using a statistical model. Some natural experimental study designs go a step further and use a combination of assumptions about the assignment process and statistical modelling to address unobserved confounders. |
| Control, counterfactual | Controls are unexposed units. Outcomes among the controls are used to estimate the counterfactual, in the sense that they represent the outcomes that would occur in the absence of exposure. In an interrupted time series study, for example, the pre-intervention trend can be extrapolated beyond the point at which the intervention occurs in order to estimate the counterfactual. |
| Exposure | A general term for receipt of an intervention; effectively synonymous with treatment in natural experimental studies. Exposure is the preferred term in this guidance to avoid confusion between treatment as a generic term and medical treatments. |
| Exchangeability | Groups are exchangeable if their outcomes would be expected to be the same under identical exposure conditions. If one group is exposed to an intervention and the other unexposed, the difference in outcomes can be interpreted as the effect of exposure. The groups are conditionally exchangeable if outcomes are expected to be the same after conditioning on a set of covariates. |
| Identification, identifying assumption | Identification refers to the method used to obtain an estimate of the effect of the intervention. All methods rely on identifying assumptions, such as the assumption that pre-intervention trends in outcomes will continue in the absence of intervention. Testing how well the assumptions are met is an important element of good natural experimental study design. |
| Instrument, instrumental variable | A variable that is associated with exposure to an intervention. An instrumental variable can be used to identify the effect of exposure on an outcome if: (1) it is associated with the outcome but (2) only through its association with the exposure (the ‘exclusion restriction’) and (3) it is unrelated to any other factors associated with the outcome. |
| Intervention | A general term for any policy, programme, service or treatment that is being evaluated. Interventions may be evaluated as they are implemented, when they are in place (using a process that determines exposure at an individual unit level within a population) or when they are withdrawn. |
| Natural experiment | An event or process outside the control of a researcher that can be used to divide a population into exposed and unexposed sub-populations, or into sub-populations with differing levels of exposure. The division may be spatial, temporal or based on the characteristics of individual units, such as test scores. A natural experimental evaluation uses the differences in exposure generated by that event or process to identify, measure or understand the effect of the intervention. |
| Negative controls, non-equivalent dependent variables | Outcomes that are not expected to change can be used as a robustness check. If changes in such outcomes are observed following exposure to the intervention, that might suggest residual confounding due to selective exposure to the intervention or to the presence of co-occurring interventions. |
| Observational study | A study that does not involve any manipulation of exposure for research purposes. Natural experimental studies are a subset of observational studies that focus on a specific event or process that generates differences in exposure. |
| Outcome, potential outcome | Outcome is used as a general term for the effect of exposure to the intervention being studied. Potential outcomes are the outcomes that would occur in the presence or absence of exposure, only one of which can ever be directly observed. Comparison of actual with potential outcomes is the basis for identifying the effect of exposure in a natural experimental study. The potential outcome for the exposed group is measured in a group that is not exposed to the intervention but is otherwise similar to the exposed group. |
| Placebo tests | A form of robustness check or sensitivity analysis using assignments that do not actually occur, such as dates on which the intervention did not take place in an interrupted time series study, or units that were not exposed to the intervention in a synthetic control study. |
| Positivity | The assumption that any combination of covariate values is possible within any exposure stratum. The combination of covariate values where this assumption holds is referred to as the ‘region of common support’. |
| Quasi-experiment | Natural experimental studies are often referred to as quasi-experiments, but quasi-experiment is also sometimes used to refer to non-randomised experiments. To avoid confusion, the term natural experiment is preferred in this guidance. |
| Selection | A process that leads exposed and unexposed units to differ in ways other than exposure to the intervention that are associated with differences in outcome. Selective exposure to the intervention is a key source of confounding in natural experimental studies. Selection in this sense is different from the idea of selective participation in a study, which may also cause bias if likelihood of participation varies according to characteristics that influence outcome(s). |
| Stable Unit Treatment Value Assumption (SUTVA) | The assumption that the outcome for each unit is independent of the outcomes for all other units, and that each unit has one potential outcome for each level of the exposure. The SUTVA can be violated when outcomes vary according to the prevalence of the exposure (e.g. when stronger effects are observed when a higher proportion of units are exposed) or when the exposure is poorly defined. |
| Target trial  Theory of Change | A hypothetical trial design that would answer the question being addressed in an observational study. Comparison of a natural experimental study design with a target trial can be used to clarify causal questions and identify possible sources of bias.  Theory of how and why the intervention impacts on outcomes of interest. |
| Time varying confounding | In studies where exposure varies over time, confounders whose values vary over time are a common source of bias and one that is not addressed by methods that deal with differences in the fixed characteristics of exposed and unexposed groups, or in observed differences in time varying characteristics. |
| Treatment | Treatment is often used as a general term for exposure to an intervention, rather than to denote a medical treatment. |
| Treatment effect | Also referred to as the causal estimand. The average treatment effect (ATE) is the difference between the average outcome when all units are exposed and the average outcome when none are. The local or complier average treatment effect (LATE or CATE) is the ATE among compliers, i.e., units whose exposure status is determined by their assignment. Different methods estimate different causal effects (see Section x, Table y) |
| Triangulation | In statistical analyses, triangulation refers to the comparison of effects obtained using different methods as a sensitivity or robustness check. If the effect estimates are comparable, despite differences in the assumptions underpinning the methods used, they can be considered more robust. In qualitative and mixed methods research triangulation may be used to refer more broadly to the integration of findings obtained using different methods, without necessarily seeking convergence. |

**Q1.** Do you agree with the content of the proposed **Concepts and definitions** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

**Q2**. If there are any key terms missing from the glossary, please describe in the comments section below

[free text box]

References

1. Petticrew M, McKee M, Lock K, Green J, Phillips G. In search of social equipoise. BMJ. 2013;347(f4016).

2. Craig P, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, et al. Using natural experiments to evaluate population health interventions: new Medical Research Council guidance. Journal of Epidemiology and Community Health. 2012;66(12):1182-6.

3. Tugwell P, Knottnerus JA, McGowan J, Tricco A. Big-5 quasi-experimental designs. Journal of Clinical Epidemiology. 2017;89:1-3.

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

5. de Vocht F, Katikireddi SV, McQuire C, Tilling K, Hickman M, Craig P. Conceptualising natural and quasi experiments in public health. BMC Medical Research Methodology. 2021;21(1):1-8.

6. Matthay EC, Glymour MM. Causal Inference Challenges and New Directions for Epidemiologic Research on the Health Effects of Social Policies. Current Epidemiology Reports. 2022:1-16.

**Guidance section: Design and planning**

|  |
| --- |
| **Design and planning summary points**   * The MRC/NIHR framework for the development and evaluation of complex interventions highlights the value of a ‘complex systems perspective’ for evaluation. Considering a natural experiment event as a disruption within a system can help identify the breadth of potential intended and unintended impacts, as well as the role of context. * For planning natural experiment studies, three important phases include identifying and theorising natural experiments, assessing their evaluability and conducting feasibility studies for a future evaluation. * Mixed-methods approaches are needed to provide a comprehensive understanding of a natural experiment, with planning required for how different study designs, types of data and analyses can be brought together. * Publication or registration of a study protocol is good practice because it enables other researchers and users of the evidence to understand what aspects of study design were decided in advance of obtaining data, and how plans were modified in light of the data. |

**Design and planning**

The MRC/NIHR framework for the development and evaluation of complex interventions(1) sets out a model for planning and conducting evaluations that can readily be adapted for natural experimental evaluations. By definition, natural experiments are identified or ‘discovered’ rather than developed or refined by researchers and are often evaluated retrospectively, rather than prior to full scale implementation. This implies a process of evidence generation that works in the opposite direction to the conventional translational research pipeline (figure x).

Diagram

Description automatically generated

Figure x Two complementary modes of evidence generation (Source: Ogilvie et al(2))

Otherwise, the phases and the core elements of the MRC/NIHR framework provide a useful structure for planning and conducting a natural experimental evaluation, by drawing attention to practices that can help maximise the usefulness of natural experimental evidence for decision-making. Evaluation of natural experiments is covered in detail in Sections x and y. In this section we focus on the early stages of identifying opportunities for a natural experimental study and working out a feasible and appropriate design. (Figure xx)

Diagram

Description automatically generated

Figure xx A framework for planning natural experimental evaluations.

1. **Discovering and theorising natural experiments**

A wide variety of circumstances may give rise to opportunities for a natural experimental evaluation. We identify five such kinds of opportunity (Table x). One is where there is a clear division in presence, level, or type of exposure between two or more otherwise similar sub-populations by time and/or place of implementation. Examples include policies implemented in some units within a federal jurisdiction but not others, such as state level gun control laws in the US or the minimum unit price at which alcohol can be sold in some UK countries but not others. A second uses individual level allocation mechanisms, such as eligibility criteria embedded within a policy. Examples include upper or lower age limits for leaving school, acquiring a driving licence, purchasing alcohol, or being exempt from deportation, and means tests that define entitlement to social security benefits. A third is where a policy is implemented in a phased way across a population in which outcome data is continuously accumulating, so that observations can be categorised as exposed or unexposed. A counterpart is where the timing of data gathering, for example in a general population survey, is unrelated to the timing of policy implementation.(3) A fourth is where randomisation is built into the policy, as in the case of the Vietnam draft lottery(4) and other lotteries used to allocate housing or school places. Flaws or shortcomings in policy delivery are a potentially useful fifth source of variation, especially if the process is either unplanned or more abrupt than is usually the case with implementation of an intervention.(5)

**Table x Opportunities for natural experimental evaluations**

|  |  |
| --- | --- |
| **Type of opportunity** | **Examples** |
| Difference over time or between places in presence or level of exposure between otherwise similar subpopulations | State level gun control laws(6); English Teenage Pregnancy Strategy(7) |
| Eligibility criteria within a policy that identify some units within a population but not others as exposed | Minimum legal age for driving or purchasing alcohol(8); eligibility rules within social security systems(9) |
| Phased implementation of a policy across a population in which outcome data is continuously accumulating | Rollout of Universal Credit(10, 11) |
| Randomisation used as an assignment mechanism within a policy | Vietnam Draft Lottery(4); housing vouchers(12) |
| Flaws or shortcomings in policy implementation | Database errors(13, 14) and false negative test results in the UK’s Test and Trace programme |

Most, if not all, population health interventions can be characterised as complex, either in the sense of having multiple components or because their impacts are moderated by interactions with elements of the wider system in which are implemented (Box x).

**Box x Taking account of complexity in evaluations of natural experiment**

Natural experiments typically occur within complex systems that influence health. Systems comprise several interacting elements which can be considered as part of a broader whole, defined by a system boundary. Such systems may be resistant to change after an event occurs (showing adaptation) while the interplay between different elements of a system may lead to unexpected or unpredictable effects. For example, if we want to evaluate the introduction of a tobacco tax, we may consider smokers, retailers, producers, smugglers, the mass media, think tanks, tobacco control advocates and the government taxation department as elements of the cigarette taxation system, whose behaviour in response to the introduction of the tax may dampen or amplify its effects. A complex systems perspective can help researchers understand why an intervention may fail to have the desired impact, or what the contingencies might be for it to succeed. Developing a holistic understanding of the important elements contained within a system and how they contribute to the system’s response to disruption, will typically require a plurality of methodological approaches and therefore has major implications for planning research.

In subsection 2, we suggest ways in which the design of a natural experimental study can take account of these interactions.

1. **Assessing the ‘evaluability’ of natural experiments**

A formal evaluability assessment is one way of ensuring that natural experimental evaluations are well-designed and address questions of relevance to decision-makers. Evaluability assessment is a systematic, collaborative approach to evaluation planning that is increasingly widely used in public health research.(15) It involves structured engagement with stakeholders to develop an agreed conceptual model of how the intervention is expected to achieve impact, identify relevant data sources and appraise evaluation options in terms of their likely cost and the usefulness of the evidence they will yield. Evaluability assessment can be used to address questions such as: what are the key uncertainties given what is already known about interventions of the kind identified as a candidate for evaluation; how will the evidence acquired from an evaluation influence future policy decisions and is it practical to obtain results in time to influence such decisions; what kinds of effects is it plausible to expect given what is already known; and how might the results of an evaluation contribute to the wider evidence base.(2, 16)

An evaluability assessment is particularly useful when an evaluation is being commissioned by policy-makers or other stakeholders because those questions may not have been explicitly addressed in the process of developing the policy and because the direct involvement of stakeholders in the evaluation planning process helps to ensure that there is a shared understanding of what an evaluation can and cannot deliver given the resources available and other constraints on evaluation design. If the study has been initiated by researchers rather than by policy or other decision-makers, an evaluability assessment may still be useful. It can help to make stakeholders aware of the research, and allow the study team to draw on their practical knowledge of intervention delivery, what kinds of monitoring information is collected and how such data can be accessed.

A diverse range of methods can be considered when planning the evaluation of a natural experiments. In many cases, it will be appropriate to use mixed methods approaches, drawing upon both quantitative and qualitative data. Natural experiment studies will also often benefit from consideration of multiple methods within each tradition. Consideration should be given at the start as to how to bring together different research components .(17) (Box xx)

**Box xx Combining study designs: alternative approaches**

Embedded design: one study design is dominant, with another playing a supportive role. For example, a process evaluation conducted alongside an outcome evaluation.(18)

Exploratory design: one study design is first used to explore and understand the natural experiment and its context, followed by using another method to conduct a more definitive evaluation. For example, to inform the evaluation of minimum unit pricing of alcohol in Scotland, qualitative interviews with policy stakeholders and document analyses were conducted, identifying both anticipated positive but also potentially unintended adverse impacts which were then evaluated in further quantitative and qualitative studies.(19)

Explanatory design: one method is used to provide explanations for the findings observed by a different method. For example, in an evaluation of free bus travel for young people in London, qualitative data was used to explore the reasons for a decline in cycling that was observed in epidemiological analyses, highlighting the importance of the greater sociability of bus travel.(20)

Triangulation: an approach which is studying the same phenomenon using different data sources, methods or theoretical perspectives. Triangulation can be used to bring together qualitative and quantitative data, or to combine evidence from a range of qualitative or a range of quantitative sources. In an evaluation of the effects of the North American Free Trade Agreement (NAFTA) on the supply of high-fructose corn syrup, both synthetic control and fixed-effects study designs were used to assess the effects of this natural experiment.(21)

Integration: drawing on multiple analyses to answer different but highly inter-related research questions. For example, Alvarado and colleagues assessed whether the announcement of a sugar-sweetened beverage tax itself led to the health risks of such beverages being more readily appreciated by the public. The authors used ‘process tracing’ whereby qualitative and quantitative data were used to test the different steps in the causal chain that would be expected to occur if such an effect existed.(22)

Natural experiment studies drawing purely on qualitative data, and using methods such as qualitative comparative analysis(23), process tracing(24) and case study designs(25), can also be valuable, although they are not the focus of this guidance.

1. **Feasibility studies for natural experimental evaluations**

Once the basic design options have been identified, a detailed assessment of their feasibility should be carried out. Whereas feasibility studies for a randomised controlled trial need to consider both the practicalities of delivering the trial intervention and the feasibility of implementing the trial procedures, the focus of feasibility work for a natural experimental evaluation will be on the practicalities of implementing the evaluation design. Questions that such a feasibility study might address include: are there routinely collected sources of data that capture (change in) exposure and outcomes, and are such data sources accessible for researchers; is it feasible to model the assignment process, given available data; are co-occurring policies likely to confound the effect of the intervention under study, and if so how can the effects of the different interventions be disentangled; if there is no comprehensive source of routinely collected data, are there feasible ways of collecting primary data?

1. **Study registration**

A natural experimental evaluation will often use multiple datasets and analysis methods. This makes it challenging to know whether all analyses and results are reported, rather than a subset which show effects in line with some prior expectation about the pattern of benefits or harms. Because NEEs are often conducted retrospectively, researchers may be using a dataset with which they are already familiar and it is important for other researchers and evidence-users to know how far the research questions and analytical choices were informed by prior knowledge of the data. For the sake of transparency, it is good practice for a study protocol to be available in the public domain before analysis commences. Registration of study protocols is now a condition of publication in some leading journals(26) and many journals now publish protocol papers. Some research funders require publication of a protocol on their webpages as a condition of funding. Protocols can also be made available through platforms such as the Open Science Framework (<https://osf.io/>) or on study authors’ institutional websites.

Currently there is no established guidance for reporting a NEE protocol, although recommendations have been made(27) and a guideline is in development for reporting protocols for observational studies.(28) In the meantime, the Standard Protocol Items: Recommendations for Interventional Trials (SPIRIT 2013) include in a protocol.(29) We recommend that, as a minimum, a NEE protocol should include: the rationale for the study, based on an up-to date review of existing evidence; a description of the intervention that is being evaluated; a description of the assignment process and definitions of the exposed and unexposed groups; the proposed design, methods and analysis plan, including sensitivity analyses, robustness checks, and the methods that will be used to integrate findings from the different study components; information about ethical approval and or information governance.

For natural experimental studies, the requirement for transparency must be balanced with the need for flexibility. There are good reasons to include a range of approaches to data analysis in the study protocol. Analysis plans will often require alteration after an initial inspection of the data, particularly when the study will use existing survey or administrative data, with a final analysis plan decided once the distributions of key variables, extent of missingness, and access to datasets and variables, are known. Protocols may also require amendments as new information emerges about how the intervention works or, in a prospective NEE, because the intervention itself is modified during the study. The quantitative analysis plan may need to be amended based on emerging findings from qualitative analysis. Registration should allow such amendments, but the process must be transparent. Baldwin et al.(30) recommend a ‘decision-tree’ approach, noting and time-stamping any changes to the analyses as the study progresses.

Preregistration of protocols will not solve all problems of selective reporting and retrofitting of hypotheses. An important additional safeguard, mandated by some research funders, is independent oversight, for example by a study steering committee to check and approve deviations from the protocol and to ensure that analyses are reported comprehensively.

**Q1.** Do you agree with the content of the proposed **Design and planning** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

[free text box]

References

1. Skivington K, Matthews L, Simpson SA, Craig P, Baird J, Blazeby JM, et al. A new framework for developing and evaluating complex interventions: update of Medical Research Council guidance. BMJ. 2021;374:n2061.

2. Ogilvie D, Adams J, Bauman A, Gregg EW, Panter J, Siegel KR, et al. Using natural experimental studies to guide public health action: turning the evidence-based medicine paradigm on its head. Journal of Epidemiology and Community Health. 2020;74(2):203-8.

3. Goodman A, van Sluijs EMF, Ogilvie D. Impact of offering cycle training in schools upon cycling behaviour: a natural experimental study. International Journal of Behavioral Nutrition and Physical Activity. 2016;13(1):34.

4. Angrist JD. Lifetime Earnings and the Vietnam Era Draft Lottery: Evidence from Social Security Administrative Records. The American Economic Review. 1990;80(3):313-36.

5. Craig P, Gibson M, Campbell M, Popham F, Katikireddi SV. Making the most of natural experiments: What can studies of the withdrawal of public health interventions offer? Preventive Medicine. 2018;108:17-22.

6. Humphreys DK, Gasparrini A, Wiebe DJ. Evaluating the impact of Florida’s “stand your ground” self-defense law on homicide and suicide by firearm: an interrupted time series study. JAMA Internal Medicine. 2017;177(1):44-50.

7. Baxter AJ, Dundas R, Popham F, Craig P. How effective was England's teenage pregnancy strategy? A comparative analysis of high-income countries. Social Science & Medicine. 2021;270:113685.

8. Callaghan RC, Gatley JM, Sanches M, Asbridge M, Stockwell T. Impacts of drinking-age legislation on alcohol-impaired driving crimes among young people in Canada, 2009–13. Addiction. 2016;111(6):994-1003.

9. Katikireddi SV, Molaodi OR, Gibson M, Dundas R, Craig P. Effects of restrictions to income support on health of lone mothers in the UK: a natural experiment study. The Lancet Public Health. 2018;3(7):e333-e40.

10. Wickham S, Bentley L, Rose T, Whitehead M, Taylor-Robinson D, Barr B. Effects on mental health of a UK welfare reform, Universal Credit: a longitudinal controlled study. The Lancet Public Health. 2020;5(3):e157-e64.

11. Brewer M, Dang T, Tominey E. Universal Credit: Welfare Reform and Mental Health. Bonn, Germany: IZA Institute of Labor Economics; 2022.

12. Nguyen QC, Rehkopf DH, Schmidt NM, Osypuk TL. Heterogeneous effects of housing vouchers on the mental health of US adolescents. American journal of public health. 2016;106(4):755-62.

13. Fetzer T, Graeber T. Measuring the scientific effectiveness of contact tracing: Evidence from a natural experiment. Proceedings of the National Academy of Sciences. 2021;118(33):e2100814118.

14. Fetzer T. Measuring the Epidemiological Impact of a False Negative: Evidence from a Natural Experiment. University of Warwick, Department of Economics; 2021.

15. Leviton LC, Khan LK, Rog D, Dawkins N, Cotton D. Evaluability assessment to improve public health policies, programs, and practices. Annual Review of Public Health. 2010;31(1):213-33.

16. Ogilvie D, Cummins S, Petticrew M, White M, Jones A, Wheeler K. Assessing the evaluability of complex public health interventions: five questions for researchers, funders, and policymakers. The Milbank Quarterly. 2011;89(2):206-25.

17. Alvarado M, Penney T, Clifford Astbury C, Forde H, White M, Adams J. Making integration foundational in population health intervention research: why we need ‘Work Package Zero’. 2022.

18. Sharma SV, Chuang R-J, Byrd-Williams C, Danho M, Upadhyaya M, Berens P, et al. Pilot evaluation of HEAL–A natural experiment to promote obesity prevention behaviors among low-income pregnant women. Preventive Medicine Reports. 2018;10:254-62.

19. Robinson M, Mackay D, Giles L, Lewsey J, Richardson E, Beeston C. Evaluating the impact of minimum unit pricing (MUP) on off-trade alcohol sales in Scotland: an interrupted time–series study. Addiction. 2021;116(10):2697-707.

20. Green J, Roberts H, Petticrew M, Steinbach R, Goodman A, Jones A, et al. Integrating quasi-experimental and inductive designs in evaluation: A case study of the impact of free bus travel on public health. Evaluation. 2015;21(4):391-406.

21. Barlow P, McKee M, Basu S, Stuckler D. Impact of the North American Free Trade Agreement on high-fructose corn syrup supply in Canada: a natural experiment using synthetic control methods. Cmaj. 2017;189(26):E881-E7.

22. Alvarado M, Penney TL, Unwin N, Murphy MM, Adams J. Evidence of a health risk ‘signalling effect’ following the introduction of a sugar-sweetened beverage tax. Food Policy. 2021;102:102104.

23. Hanckel B, Petticrew M, Thomas J, Green J. The use of Qualitative Comparative Analysis (QCA) to address causality in complex systems: a systematic review of research on public health interventions. BMC Public Health. 2021;21(1):877.

24. Beach D, Pedersen RB. Process-tracing methods: Foundations and guidelines: University of Michigan Press; 2019.

25. Byrne D. Evaluating complex social interventions in a complex world. Evaluation. 2013;19(3):217-28.

26. Summerskill W, Collingridge D, Frankish H. Protocols, probity, and publication. The Lancet. 2009;373(9668):992.

27. Marsden J, Cousijn J, Stapleton J. Pre-registration: not a daunting practice. Addiction. 2022;117(4):845-6.

28. Mahajan R, Burza S, Bouter LM, Sijtsma K, Knottnerus A, Kleijnen J, et al. Standardized protocol items recommendations for observational studies (SPIROS) for observational study protocol reporting guidelines: protocol for a Delphi study. JMIR Research Protocols. 2020;9(10):e17864.

29. Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ : British Medical Journal. 2013;346:e7586.

30. Baldwin JR, Pingault J-B, Schoeler T, Sallis HM, Munafò MR. Protecting against researcher bias in secondary data analysis: challenges and potential solutions. European Journal of Epidemiology. 2022;37(1):1-10.

**Guidance section: Quantitative methods**

|  |
| --- |
| **Quantitative methods summary points**   * Presenting a hierarchy of quantitative study designs for natural experimental evaluations is unhelpful * If individual level data are available, use an analytical method that uses data at that level (do not aggregate data to fit another approach) * As natural experimental evaluations use non-randomised designs, threats to internal validity are always a concern |

This chapter provides an overview of the main quantitative study designs and analytic methods available to the researcher when conducting a natural experimental evaluation (NEE).

The primary considerations for conducting a NEE should come from the nature of the research question(s). Following this, the design features of the evaluation will influence the choice of quantitative analytic method for the analysis of NEEs.(1) As part of this process, it can be useful to consider what the features of a randomised design would be were it feasible or ethical (i.e., emulate a ‘target trial’ (2)). An essential criterion for the appraisal of the strength of NEEs is the plausibility of as-if randomisation of the intervention.(3) This cannot be determined from the analytic method but is a subjective assessment based on characteristics of the methodology.(4) Although a helpful framework for highlighting key concerns on internal validity, it should be noted that some aspects will be impossible to emulate (e.g., blinding) and sometimes the natural experiment might have a superior characteristic to the equivalent in the hypothetical trial; for example, reduced risk of participation bias if using routinely collected data for the evaluation. In practice, however, sometimes the researcher will be forced to opt for a sub-optimal evaluation design if the data are not available; for example, when historic data are not available the researcher may need to rely on a cross-sectional evaluation instead of, for example, a repeated cross-sectional design. Similarly, where individual-level data are not available, the researcher might be able to make use of data aggregated to some grouped level, such as for example schools, geographical regions or countries. It might also be the case that the interest is in the impact on some aggregated unit rather than on the individual. Such decisions need to be clearly communicated.

As NEEs use non-randomised study designs, there will, in general, be a certain degree of self-selection into the exposed or unexposed group. This contrasts to randomised study designs where individuals (or groups of individuals) are actively randomised by the researcher. As a consequence, NEEs tend to estimate Average Treatment Effects on the Treated (ATTs) rather than Average Treatment Effects (ATEs). The exceptions are discontinuity study designs and the instrumental variable method which estimate local average treatment effects (LATEs). Further, with non-randomised study designs the degree of comparability between the intervention and control groups might affect internal validity. The methods outlined below take different approaches to address this, some by using measured variables (‘observables’) – i.e., matching, multivariable regression, propensity scores, with others accounting for unmeasured variables (‘unobservables’) – i.e., instrumental variables, discontinuity designs.

If individual level data are available, it is preferable to use an analytical method that uses data at that level rather than create aggregated units for analysis (this would in essence be ‘throwing information away’). It has also been shown that aggregated level analyses can be prone to bias if data at the individual level are missing at random (MAR).(5) Therefore, if only aggregated level data are available, the quality of the underlying data sources should be interrogated and the likelihood of missing data assessed. Please note that in the table we have summarised ITS and cITS designs assuming only aggregate level data are available as we consider this to be the most likely situation researchers are faced with.

In situations where individual level data or not available, or where the researcher made the conscious decision to use aggregated data, and the NEE aims to estimate effects for some aggregated units, the researcher might be able to use routinely collected quantitative data. In this situation, it might even be possible that these data are available as open access, in which case transparency can be improved by making both data and analytic coding scripts available.

In the table, we have treated exposure to the intervention as being uniform across the exposure group and it is the change from not being exposed (and comparisons to groups never exposed) that is the key feature of a natural experiment which sets them apart from other observational studies (where this change in exposure does not occur) but also from randomised designs (where the randomisation of exposure is controlled by the researcher), and thus also from the different methods used to analyse them. However, it could be that some groups within the intervention group receive different ‘doses’ of exposure and/or the timing of exposure varies (like what happens by design in a stepped wedge cluster randomised trial). These situations would lead to differences in how exposure is measured within the statistical approaches undertaken.

It should be stressed before outlining different NEE study designs that it is problematic to suggest a hierarchy in terms of methodological quality. This is partly because the design should be determined by the research questions but also because each design has its own strengths and weaknesses. For example, although a cITS is a superior design to ITS in theory, if the control group is poor (e.g., for example the parallel trends assumption (6) is not met) the comparison might be biased and it may be better to proceed without the control group. Similarly, a design that uses an instrumental variable will not be better than a study using only observables if the instrument itself is ‘weak’ (i.e., low correlation with exposure to intervention).(7)

We also highlight important new developments in the NEE field to complement this overview. Given the rise and prominence of machine learning it seems likely this will have a role in future research for predicting the counterfactual based on algorithms identifying patterns and relationships in pre-intervention data. For example, in a study evaluating the impact of social housing on mental health, machine learning methods were used to maximise precision of weights used in an inverse probability weighted (IPTW) approach.(8)

Additionally, rather than empirical NEE evidence, mathematical modelling can also be used to estimate the likely impact of new interventions before they are introduced. Examples of such modelling include evaluating effects of minimum unit pricing of alcohol(9) and municipal transportation policy.(10)

Finally, although the study designs and statistical approaches outlined in this overview aim to estimate causal effects of interventions, they remain ‘black-box’ effects. Although qualitative evidence is important for understanding causal mechanisms (<cross-reference to qualitative section>), a quantitative approach would be to use causal mediation analysis to understand how variables operate along the causal pathway between intervention exposure and outcome.(11)

The study designs below are presented in Table X.

*Study design 1:* An advantage of a **Cross-sectional design** is its simplicity. However, it is limited because the outcome is only measured after the intervention is started, but not measured before (neither in the intervention group or, if available, in a control group(s)). Even if the intervention group is matched to a control group(s) based on possible confounding variables, not knowing baseline/pre-intervention time trends in the outcome is a major weakness.

*Study design 2:* The **Repeated cross-sectional design** is superior to a cross-sectional design as data is collected on the outcome before and after the intervention starts. However, a weakness is that these are not necessarily the same individuals, so results could be biased because of different populations in each cross-section. It would be possible to adjust for known (and measured) confounding variables in a regression model. An example of a NEE study with this design was used to assess the impact of an opening of a new franchise of a restaurant on young people’s eating behaviours, (12) with data collected via an online questionnaire at baseline (prior to restaurant opening) and 3 and 9 months post-opening.

*Study design 3:* The **Before and after design** is an improvement to the repeated cross-sectional design as there are repeated measurements available, before and after the intervention start, on the same individuals (e.g., individuals are acting as their own control(13)). An example of a NEE study with this design was used to assess the change in emotional response to the COVID-19 pandemic at two time points, during strict lockdown measures and later when vaccination programmes were being rolled out.(14) Note: some definitions of this design allow for individuals to be different before and after the intervention start(15) which equates to our study design 2.

*Study design 4:* A **Difference in Differences (DiD)** design is essentially a repeated cross-sectional design but with data collected in an intervention and control groups. Mutiple time points in the pre-intervention period allow to better estimate the counterfactual (what the outcome would be in the post-intervention period if the intervention did not occur) than just a single time point, or time points with uneven periods between them. Ikenwilo(16) employed a DiD design to evaluate the effects of free dental check-ups in Scotland using the rest of the UK as a control group (multiple time points were obtained from the annual British Household Panel Survey and the intervention started in 2006).

*Study design 5:* In contrast to the study designs introduced so far, a **Discontinuity design** can account for confounding effects from unmeasured variables. In a ‘sharp’ regression discontinuity design, an assignment variable determines whether an intervention is received deterministically, whereas in a ‘fuzzy’ design is probabilistically assigned at a threshold of the assignment variable.(17) Geography can be the key feature of such designs.(18) An example of a NE study using this type of design is a study that assessed the effect of increased primary schooling on adult women’s HIV status in Malawi and Uganda(19) where a new policy allowed girls aged 13 years and under to continue schooling but not those aged above 13 years (i.e., a ‘sharp’ design). A ‘fuzzy’ regression discontinuity design is equivalent to an **Instrumental variable design**.(20) It follows, therefore, that an instrumental variable design also accounts for unmeasured confounding. An instrument is a variable that is associated with exposure to the intervention, but conditional on that exposure there is no independent association with the outcome.(7) In an Indonesian study,(21) the amount of rainfall (which is strongly correlated with crop production) is used as an instrument to assess the impact of income on mental health outcomes (assuming that amount of rainfall is not directly associated with these outcomes).

If individual-level data are not available, the researcher can use time series data, often from routine data sources, to evaluate a NE using the following study designs.

*Study Design 6:* an **Interrupted Time Series (ITS) design** can be considered a repeated cross-sectional design, but is based on aggregated data (generally counts or rates) from more than two measurement occasions with equally spaced time intervals. An example of an interrupted time series design describes the evaluation for the impact of lockdown policies in India on COVID-19 incidence(22) which used daily new cases as the outcome and dates of stages of lockdown as the ‘interruptions’ that change exposure status.

*Study design 7:* a **Controlled interrupted time series (cITS)** is an ITS design but with a control group with measurements at the same time points as the intervention group. If a suitable control can be identified (this can be assessed by pre-intervention time trends and comparison of demographics / time varying confounding variables) it can be argued this provides a superior counterfactual to ITS alone. It is possible to incorporate more than one control group into the analytical design.(10) Further, different types of control other than a control group, but which should also not respond to the intervention (e.g., so called ‘control outcomes’(6)), can be used if appropriate. An example of a NEE with a cITS design is a US study into the effect of a safe opioid prescribing initiative on levels of opioid prescribing(23) where benzodiazepine prescribing was used as a control outcome. The Synthetic control design can be considered as a DiD or cITS with the control condition based on a weighted combination of control units so that it mimics the characteristics of the exposed group. This ‘synthetic control’ is then used to estimate what would have happened in the exposed group had the intervention not happened (the counterfactual), and the comparison between that estimate and the real data is considered as the causal effect. In a study from Ethiopia,(24) a synthetic control was used to assess a health extension program on maternal mortality with the control being a weighted linear combination of other countries in Sub-Saharan Africa that did not have the intervention being evaluated.

|  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Table X:   |  |  |  |  |  |  | | --- | --- | --- | --- | --- | --- | | Study design | Comparison | Data | Statistical approaches | Characteristics | Illustrative example – Cardiovascular disease (CVD) ‘health check’ delivered in primary care centres (i.e., screening for CVD risk factors). Outcome measure – incident CVD events (hospitalisations / deaths) | | 1) Cross-sectional | Individual level | Post-intervention; random sample (ideally); single time point of data collection; data potentially collected in control group(s) | Descriptive statistics for effect size – with representation of uncertainty; possible matching of intervention group(s) with control group(s) | No information on pre-intervention outcome, so effect size unclear | Study conducted in all / subset of primary care centres post-intervention; rate of incident CVD events compared to control group(s) (or compared to literature); possible matching of control group(s) to intervention group(s) before comparison | | 2) Repeated cross-sectional | Individual level | Pre- and post-intervention; random samples (ideally); data collection at unequally spaced time intervals; no data from control group(s) | Difference between pre- and post-intervention in means / proportions / rates (depending on nature of outcome measure variable) – with representation of uncertainty; regression models / propensity scores – to adjust for confounding variables / assess effect modification | Allows for comparison with pre-intervention outcome, but because pre/post groups include different people this might bias comparisons | Study conducted in all / subset of primary care centres pre- and post-intervention (two time points); difference in rate of incident CVD events compared pre- and post-intervention | | 3) Before and after | Individual level | Pre- and post-intervention; random sample (ideally); two time points of data collection (on same individuals – repeat measurements); no data from control group(s) | Average difference between pre- and post-intervention measurements of the outcome measure; regression models / propensity scores – to adjust for confounding variables / assess effect modification | Allows for comparison with pre-intervention outcome based on repeated measures of the same group, but does not have a control group. Issues of missing data might bias comparison | Not possible given nature of outcome measure (incident CVD events); would be possible if SBP was outcome measure | | 4) Difference in Differences | Individual level | Random samples (ideally); pre- and post-intervention; intervention and control group | Regression models; possible matching of intervention group with control group | Before-and-after design with a control group; can be difficult to identify comparable control unit(s) | Difference between intervention and control groups in difference in rate of incident CVD events compared pre- and post-intervention | | 5) Discontinuity designs | Individual level | Random samples (ideally); data collected either side of a ‘cut-off’ for a variable determines if an individual is eligible for intervention (and assignment to intervention or control group) / An ‘instrument’ is a variable that is associated with exposure to the intervention but not itself associated with outcome | Nonparametric methods; regression models; assess effect modification | Limited situations where a cut-off can be identified / Strong instruments are difficult to identify | Use one of the eligibility criteria of health check (systolic blood pressure (SBP) > 140mmHg) as ‘cut-off’ / Use distance to primary care centre where health check is being offered as ‘instrument’ | | 6) Interrupted time series | Aggregated data | Pre- and post-intervention; data collection on multiple occasions at equally spaced time intervals; ‘interruption’ is at time point when intervention starts; no data from control group(s) | Time series; (s)ARIMA / (panel) regression models; adjustment for confounding variables; assess effect modification | Allows for comparison with pre-intervention outcome based on multiple repeated measures of the same group, but does not have a control group | Study time series of rates of incident CVD events; single time series (data from primary care centres combined) or multiple time series (for each/sub-groups of primary care centres) | | 7) Controlled interrupted time series | Aggregated data | Pre- and post-intervention; multiple time points of data collection (evenly spaced intervals); ‘interruption’ is at time point when intervention starts; intervention and control group(s) | Time series; ARIMA / (panel) regression models; adjustment for confounding variables; assess effect modification / Use the pre-intervention data to create a ‘synthetic control’; a weighting procedure is applied using the outcome variable and possible confounding variables from the pool of control groups | Interrupted time series with control group. Pre-intervention time period differences between intervention and control groups may cast doubt on intervention effect estimates / If appropriate controls cannot be identified, a synthetic control can be developed to obtain counterfactual. Quality of synthetic control not always easy to establish. Communication to practitioners difficult | Study time series of rates of incident CVD events in intervention and control group(s) / Synthetic control group | |

**Q1.** Do you agree with the content of the proposed **Quantitative methods** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

[*freetext/comments box*]

References

1. de Vocht, F., et al., *Conceptualising natural and quasi experiments in public health.* BMC Med Res Methodol, 2021. **21**(1): p. 32.

2. Hernán, M.A. and J.M. Robins, *Using Big Data to Emulate a Target Trial When a Randomized Trial Is Not Available.* Am J Epidemiol, 2016. **183**(8): p. 758-64.

3. Dunning, T., *Natural experiments in the social sciences - a design based approach*. 2012: Cambridge University Press.

4. Craig, P., et al., *Using natural experiments to evaluate population health interventions: new Medical Research Council guidance.* J Epidemiol Community Health, 2012. **66**(12): p. 1182-6.

5. Bazo-Alvarez, J.C., et al., *Handling Missing Values in Interrupted Time Series Analysis of Longitudinal Individual-Level Data.* Clin Epidemiol, 2020. **12**: p. 1045-1057.

6. Lopez Bernal, J., S. Cummins, and A. Gasparrini, *The use of controls in interrupted time series studies of public health interventions.* Int J Epidemiol, 2018. **47**(6): p. 2082-2093.

7. Lousdal, M.L., *An introduction to instrumental variable assumptions, validation and estimation.* Emerg Themes Epidemiol, 2018. **15**: p. 1.

8. Bentley, R., et al., *The impact of social housing on mental health: longitudinal analyses using marginal structural models and machine learning-generated weights.* Int J Epidemiol, 2018. **47**(5): p. 1414-1422.

9. Brennan, A., et al., *Potential effects of minimum unit pricing at local authority level on alcohol-attributed harms in North West and North East England: a modelling study.* Public Health Research, 2021. **9**(4).

10. Arnason, T., et al., *Municipal transportation policy as a population health intervention: estimating the impact of the City of Ottawa Transportation Master Plan on diabetes incidence.* Can J Public Health, 2019. **110**(3): p. 285-293.

11. Imai, K., L. Keele, and D. Tingley, *A general approach to causal mediation analysis.* Psychol Methods, 2010. **15**(4): p. 309-34.

12. al, H.M.C.O.M.S.L.e., *A natural experiment using repeat cross‐sectional data measuring the impact of the opening of a new franchise of a multi‐national restaurant on young people's eating behaviours and their perceptions*, in *European and International Congress on Obesity 2020*. 2020. p. 219.

13. Kraska, M., *Repeated Measures Design*, in *Encyclopedia of Research Design*, N. Salkind, Editor. 2012, Sage.

14. Mozes, M., I. van der Vegt, and B. Kleinberg, *A repeated-measures study on emotional responses after a year in the pandemic.* Scientific Reports, 2021. **11**.

15. Sedgwick, P., *Before and after study designs.* Bmj, 2014. **349**: p. g5074.

16. Ikenwilo, D., *A difference-in-differences analysis of the effect of free dental check-ups in Scotland.* Soc Sci Med, 2013. **83**: p. 10-8.

17. Oldenburg, C.E., E. Moscoe, and T. Bärnighausen, *Regression Discontinuity for Causal Effect Estimation in Epidemiology.* Curr Epidemiol Rep, 2016. **3**: p. 233-241.

18. Keele, L.J. and R. Titiunik, *Geographic Boundaries as Regression Discontinuities.* Political Analysis, 2015. **23**.

19. Behrman, J.A., *The effect of increased primary schooling on adult women's HIV status in Malawi and Uganda: Universal Primary Education as a natural experiment.* Social Science & Medicine, 2015. **127**: p. 108-115.

20. Lee, D.S. and T. Lemieux, *Regression Discontinuity Designs in Economics.* Journal of Economic Literature, 2010. **48**(2): p. 281-355.

21. Hanandita, W. and G. Tampubolon, *Does poverty reduce mental health? An instrumental variable analysis.* Social Science & Medicine, 2014. **113**: p. 59-67.

**Guidance section: Economic evaluation**

|  |
| --- |
| **Economic evaluation summary points**   * There is a lack of guidance on the design and conduct of economic evaluations alongside natural experimental evaluations. * Designing, conducting, and reporting economic evaluations pose specific challenges when conducted alongside a natural experimental evaluation. These include the measurement and identification of costs and outcomes, selecting appropriate analytical methods, identifying the time horizon, and equity considerations. |

**Economic evaluation**

1. **Introduction: economic evaluation**

Economic evaluations compare two or more courses of actions in terms of their costs and consequences,(1) by identifying, measuring and valuing costs and outcomes and summarizing the result in an incremental cost-effectiveness ratio (ICER) which is compared with a willingness to pay (WTP) threshold. The comparative analysis involved in the economic evaluation provides a systematic framework to support transparent and evidence-based decision-making about funding of cost-effective public health interventions (PHIs). In consideration of the resource constraints faced by government, comparing costs and benefits of alternative interventions on population health using an objective framework is essential to ensure well-informed decisions on the allocation of scarce resources.(2)

Depending on the outcome of interest, economic evaluations can take the form of: cost-effectiveness analysis (CEA), where outcomes are expressed as a single natural unit (e.g. life-years gained); cost-utility analysis (CUA), measuring outcomes in terms of quality-adjusted life years (QALY); cost-benefit analysis (CBA), where health benefits are measured in monetary terms; cost-consequence analysis (CCA), where multiple outcomes are reported in a summary table and multiple outcomes of interest are reported.

1. **Designing and conducting economic evaluations alongside natural experiments**

Recently, natural experimental (NE) design have emerged as a viable and valuable framework to assess the cost-effectiveness of PHIs, especially when a randomised design is not feasible because of ethical or practical reasons.(3) While consensus exists on the methodological challenges posed by PHIs (attribution of effects; measuring and valuing outcomes; identification of intersectoral costs and consequences; equity considerations),(4) and on the need for specific guidance for economic evaluation of PHIs,[5] there is paucity of literature on the design and conduct of economic evaluations of PHI alongside NE design. Indeed, very few examples of full economic evaluations of PHIs alongside NEs exist (e.g. (6, 7)). Existing literature on NEs focuses on effectiveness only,(8) does not explore the effect of the policy on population health or, even though considering some cost components, fails to evaluate costs and benefits in a systematic cost-effectiveness framework (e.g.(9)). On the other hand, the literature considering full economic evaluation frameworks is limited to statistical methods to address the selection bias arising from non-randomised studies in a cost-effectiveness framework,(10, 11) or focuses on the challenges of conducting economic evaluations of PHIs, without considering the NE framework.(12, 13)

As shown in a recent methodological paper, NE frameworks pose specific challenges to economic evaluations, thus requiring a specific guidance.(3) Indeed, the available economic evaluation guidance is tailored on the most common randomised controlled trial (RCT) framework(1) and does not cover the challenges inherent to NE designs. However, designing, conducting, and reporting economic evaluations without specifically considering the NE framework can lead to sub-optimal design, data collection and analysis for NEs, leading to bias in the estimated effectiveness and cost-effectiveness of the PHI.(3)

The following sections outlines the key challenges posed by NE design in the economic evaluation of interventions to improve public health.

***Measurement and identification of costs and outcomes in NE***

As linked, administrative data used in NE evaluation usually do not include preference-based outcome measures (e.g. EQ5D), the health economics researcher interested in estimating the effect of a public health policy on quality of life (QoL), should rely on ‘intermediate’ outcome, which will subsequently be mapped into QALY using decision-analytic models.(15)

However, as a societal perspective is advocated for the economic evaluation of PHIs, (5, 16) broader economic evaluation frameworks incorporating multiple, multisectoral outcomes, such as CCA or CBA(16) should be preferable to CUA. Multi-criteria decision analysis (MCDA),(17) weighting and valuing multiple outcomes using an explicit method (e.g. discrete choice experiment), is also a recommended tool to aid decision making in the economic evaluation of PHIs. Similarly, multiple costs and cost savings falling beyond the health sector should be included when possible. As the health economist cannot rely on a bespoke data collection instrument, this may be challenging. It is recommended that the design of economic evaluations alongside NE design should include a thorough investigation of routine data, to identify the availability of data sources to identify relevant, multisectoral costs and outcomes.

Furthermore, in the absence of a bespoke data collection, proxy unit costs may be used to evaluate available resource use (e.g. average cost/bed-day unit cost). It is thus recommended that sensitivity around the assumptions made in relation to unit costs is done.

***Time horizon***

The availability of routinely collected data for a long observational period allows the health economics researcher to capture the impact of the intervention on costs and consequences over a long time horizon. This is a key advantage for the economic evaluation of PHIs, which might exert their impact only on a medium-long term or may have effects that ‘carry over’ after the end of the intervention. Decision-analytic models, usually populated with parameters derived from the literature in RCT contexts, can be populated using routine data specific to the target population, following advice specific to PHIs.(18)

***Analytical methods***

When conducting economic evaluations of PHIs alongside a NE design, quantitative methods to deal with observed and unobserved confounding need to be embedded in economic evaluation frameworks. This implies dealing with the additional complexity generated by correlated costs and outcomes,(19) skewed and non-normal distributions of costs and outcomes,(20, 21) potentially correlated multiple outcomes,(22) clustering.(23)

Most methods to deal with observed confounding in non-randomised setting allow the estimation of the Average treatment effect (ATE), which can be interpreted as the incremental cost and incremental effectiveness parameters in the general population, whereas methods to deal with unobserved confounding (e.g. IV and RDD) allow the estimation of LATE (incremental cost and incremental effectiveness parameters among the compliers population).

Sensitivity analysis should explore multiple designs (and multiple control groups) as to explore the sensitivity of economic evaluation to multiple sources of bias and strengthen the credibility of results.

Moving beyond traditional cost effectiveness framework, dynamic microsimulation models also offer a valid and flexible tool for the evaluation of costs and outcomes associated with policies to improve population health.(24)

***Equity considerations***

In consideration of the potential effect of PHIs on health inequalities, methods exploring equity, besides efficiency, such as distributional cost-effectiveness analysis(25) and extended cost-effectiveness analysis(26) should be also considered. However, as these methods are more demanding in terms of data requirements, they may not be feasible to be implemented using observational data.

**Q1.** Do you agree with the content of the proposed **Economic evaluation** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

[*freetext/comments box*]

References

1. Drummond, M.F., M.J. Sculpher, K. Claxton, G.L. Stoddart, and G.W. Torrance, *Methods for the economic evaluation of health care programmes*. 2015: Oxford university press.

2. McIntosh, E. and R. Luengo-Fernandez, *Economic evaluation. Part 1: Introduction to the concepts of economic evaluation in health care.* BMJ Sexual & Reproductive Health, 2006. **32**(2): p. 107.

3. Deidda, M., C. Geue, N. Kreif, R. Dundas, and E. McIntosh, *A framework for conducting economic evaluations alongside natural experiments.* Social science & medicine, 2019. **220**: p. 353-361.

4. Weatherly, H., M. Drummond, K. Claxton, R. Cookson, B. Ferguson, C. Godfrey, N. Rice, M. Sculpher, and A. Sowden, *Methods for assessing the cost-effectiveness of public health interventions: key challenges and recommendations.* Health policy, 2009. **93**(2-3): p. 85-92.

5. NICE, *Methods for the Development of NICE Public Health Guidance: Process and Methods Guides.* 2012.

6. Alfonso, Y.N., D. Bishai, J. Bua, A. Mutebi, C. Mayora, and E. Ekirapa-Kiracho, *Cost-effectiveness analysis of a voucher scheme combined with obstetrical quality improvements: quasi experimental results from Uganda.* Health policy and planning, 2015. **30**(1): p. 88-99.

7. Leyland, A.H., S. Ouédraogo, J. Nam, L. Bond, A.H. Briggs, R. Gray, R. Wood, and R. Dundas, *Evaluation of health in pregnancy grants in Scotland: a natural experiment using routine data.* Public Health Research, 2017. **5**(6).

8. Craig, C. Cooper, D. Gunnell, S. Haw, K. Lawson, S. Macintyre, D. Ogilvie, M. Petticrew, B. Reeves, and M. Sutton, *Using natural experiments to evaluate population health interventions.* Glasgow, UK: Medical Research Council, 2011.

9. Angrist, J., E. Bettinger, E. Bloom, E. King, and M. Kremer, *Vouchers for Private Schooling in Colombia: Evidence from a Randomized Natural Experiment.* The American economic review, 2002. **92**(5): p. 1535-1558.

10. Manca, A. and P.C. Austin, *Using propensity score methods to analyse individual patient level cost effectiveness data from observational studies.* The University of York: Health Economics and Data Group Working Paper, 2008. **8**: p. 20.

11. Kreif, Grieve, and Sadique, *Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice.* Health Econ, 2013. **22**(4): p. 486-500.

12. Smith, R.D. and M. Petticrew, *Public health evaluation in the twenty-first century: time to see the wood as well as the trees.* Journal of Public Health, 2010. **32**(1): p. 2-7.

13. Petticrew, M., S. Cummins, C. Ferrell, A. Findlay, C. Higgins, C. Hoy, A. Kearns, and L. Sparks, *Natural experiments: an underused tool for public health?* Public health, 2005. **119**(9): p. 751-757.

15. Briggs, A., M. Sculpher, and K. Claxton, *Decision modelling for health economic evaluation*. 2006: Oup Oxford.

16. Edwards, R.T. and E. McIntosh, *Applied health economics for public health practice and research*. 2019: Oxford University Press.

17. Marsh, K., M. IJzerman, P. Thokala, R. Baltussen, M. Boysen, Z. Kaló, T. Lönngren, F. Mussen, S. Peacock, and J. Watkins, *Multiple criteria decision analysis for health care decision making—emerging good practices: report 2 of the ISPOR MCDA Emerging Good Practices Task Force.* Value in health, 2016. **19**(2): p. 125-137.

18. Squires, H., J. Chilcott, R. Akehurst, J. Burr, and M.P. Kelly, *A framework for developing the structure of public health economic models.* Value in Health, 2016. **19**(5): p. 588-601.

19. Nixon, R.M. and S.G. Thompson, *Methods for incorporating covariate adjustment, subgroup analysis and between‐centre differences into cost‐effectiveness evaluations.* Health economics, 2005. **14**(12): p. 1217-1229.

20. Jones, A.M., J. Lomas, P. Moore, and N. Rice, *A quasi-Monte Carlo comparison of developments in parametric and semi-parametric regression methods for heavy-tailed and non-normal data: with an application to healthcare costs.* Journal of the Royal Statistical Society: Series A (Statistics in Society), 2015: p. 951-974.

21. Basu, A., *ESTIMATING PERSON-CENTERED TREATMENT (PeT) EFFECTS USING INSTRUMENTAL VARIABLES: AN APPLICATION TO EVALUATING PROSTATE CANCER TREATMENTS.* Journal of Applied Econometrics, 2014. **29**(4): p. 671-691.

22. Teixeira‐Pinto, A. and S.L.T. Normand, *Correlated bivariate continuous and binary outcomes: issues and applications.* Statistics in medicine, 2009. **28**(13): p. 1753-1773.

23. Gomes, M., E.S.-W. Ng, R. Grieve, R. Nixon, J. Carpenter, and S.G. Thompson, *Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials.* Medical Decision Making, 2012. **32**(2): p. 350-361.

24. Zucchelli, E., A. Jones, and N. Rice, *The evaluation of health policies through dynamic microsimulation methods.* International Journal of Microsimulation, 2012. **5**(1): p. 2-20.

25. Asaria, M., S. Griffin, and R. Cookson, *Distributional cost-effectiveness analysis: a tutorial.* Medical Decision Making, 2016. **36**(1): p. 8-19.

26. Verguet, S., R. Laxminarayan, and D.T. Jamison, *Universal public Finance of tuberculosis treatment in India: an extended cost‐effectiveness analysis.* Health economics, 2015. **24**(3): p. 318-332.

**Guidance section: Qualitative methods**

|  |
| --- |
| **Qualitative methods summary points**   * Most natural experimental evaluations will require a mixed-method design incorporating qualitative methods, which importantly enhance the explanatory potential of a quantitative evaluation. * Qualitative methods are necessary for adequately characterising system, context and intervention. * Qualitative evidence is typically required for better characterisation of exposed/non-exposed groups; understanding causal mechanisms; selection and measurement of meaningful indicators for outcomes and stakeholder perspectives. * Ideally, qualitative analysis is also drawn on to help make credible attributions of causality. |

**Qualitative methods**

There is existing guidance on many primarily qualitative designs that can contribute to natural experimental evaluations (NEEs), such as case study(1, 2) and realist synthesis.(3) Here, we focus on the role of qualitative methods within those NEEs that use differences in exposure to estimate effect size, which typically employ qualitative methods in tandem with statistical analysis of quantitative data within mixed-method designs. Guidance for process evaluations for complex interventions(4) is orientated primarily at trials, in which both development of the intervention and assessing fidelity are key issues. However, the principles of this guidance also apply to NEEs. Ideally, qualitative evidence should be integrated throughout evaluations. Most NEEs will require qualitative evidence to ensure that: the system, context and intervention are well characterised; exposed/non-exposed groups are appropriately defined and cases assigned to those groups; causal mechanisms are understood; outcomes selected are meaningful; indicators for these are valid and reliably measured; and stakeholder perspectives on the intervention and its effects are understood.

In addition, qualitative methods can add considerable value by explicating links between intervention and outcomes in context, and thus strengthening attributions of causality and transferability. This guidance does not address the selection of appropriate methods of data generation and analysis. They are likely to include both existing documentary sources, and also primary data, from a range of methods including individual and focus group interviews, ethnographic observation and photo elicitation, depending on the specific needs of the study and its theoretical framing. Rather, this guidance outlines the key points at which qualitative methods should be integrated within NEEs to maximise the opportunities for robust and generalizable evaluations.

1. **Characterising the intervention, context and system**

***Describing the intervention***

The intervention itself requires detailed description and delineation, including, where appropriate, details of how the intervention was implemented in practice. Hanckel et al(5) for instance, used ethnographic observation and interviews to describe the Daily Mile intervention (which encourages children to run for 15 minutes each school day), comparing the intervention as described in principle in promotional material, and how it was implemented in practice, using the PHP-TIDieR checklist.(6)

***Describing the system/context***

A strength of NEEs is that they can explore relationships between the context/system and intervention, and help characterise the complexity of the system. This also relies on detailed ‘thick’ description, drawing typically on documentary analysis, observations and interviews as appropriate to describe the relevant demographic, geographical, historical, political, and organisational aspects of context,(7) including any co-occurring interventions in the setting.

***Developing theories of change***

Qualitative methods such as interviews and workshops with stakeholders should be used to develop an initial and (at project end) a revised theory of change for whether, how and why the intervention had particular effects, within a system, including mapping all potential positive and negative outcomes of importance to stakeholders. Christie et al(8), for instance, undertook a qualitative study to inform the theory of change for an NEE of the introduction of Graduated Driving Licencing in Northern Ireland. This characterised the key components of context (such as widespread reliance on car transport for accessing goods and services; car sharing within local economies) to inform choice of plausible comparator settings; hypothesised links between intervention and potential public health impacts; and identified an important co-occurring intervention in the system (the widespread adoption of telematics insurance products). Penney et al(9) used expert workshops, a Delphi exercise and interviews with key stakeholders to describe the complex system in which a proposed soft drinks industry levy in the UK would intervene, and to identify likely impacts.

1. **Informing selection of populations, controls, and sub-groups for analysis**

Initial qualitative work to characterise the intervention, context and broader system informs and sense-checks the selection of, and assignment of units to, appropriate exposed/non-exposed populations, and helps identify important sub-groups for analysis. For example, Ogilvie et al(10) used media analysis and ethnographic observations to characterise the Cambridgeshire Guided Busway, finding that it was used before being officially open, and that prior experience of different modes influenced initial perceptions. This information was used to select appropriate dates for measuring change resulting from the intervention, and for identifying important sub-groups for analysis of effects.(10, 11) De Vocht et al(12) used local practitioners’ insights on appropriate comparator areas to contribute to creating synthetic controls for the counterfactuals in small-area studies of the public health impact of alcohol licensing decisions.

During the evaluation, as a nested or concurrent study, detailed qualitative analysis can help unpack how the intervention effects differ across sub groups. Gibson et al(13), for instance, as part of a mixed-method evaluation of social prescribing for people with diabetes,(14) analysed four individual cases to understand how ‘health capital’ and structural conditions shaped participants’ capacity to interact with and benefit from the intervention.

1. **Characterising outcomes and indicators**

The choice of appropriate primary and secondary outcomes, and indicators for these, should be informed by qualitative evidence at the outset. This can help ensure that indicators are capturing what is intended. For instance, in a study of the public health impacts of free bus travel, Jones et al(15) used observational and interview data to identify that travelling by bus entailed considerable physical activity for young people: using ‘bus trips’ as an indicator of passive travel would therefore underestimate ‘active travel’. Interviews may be insufficient as a standalone method, given the challenges in articulating rationales for routine practices. Ogilvie et al(10) used photo-elicitation with commuters to gain insights into the factors that might influence mode and route choice, but which may not be articulated in interviews.

Qualitative evidence will also inform the interpretation of statistical findings, by shedding light on the meaning of indicators in context, the potential limitations in survey or other measures, and the interpretation of relationships between indicators. Guell et al(16), for instance, in exploring unexpected findings that some commuters used active modes despite reporting negative perceptions of the environments for walking and cycling on questionnaires. In interviews, some reported that their questionnaire response reflected a desire to make a political point about facilities for cycling and walking, rather than necessarily representing their own perceptions.

1. **Generating data on outcomes**

Although quantitative measures are needed for estimating effects on outcomes, qualitative evidence plays a role in providing additional sources of data for triangulation and comparisons (to aid analytical inference) or, on occasion, (where, for instance, quantitative data do not exist or would be difficult to access) providing the main source of evidence on change resulting from an intervention. For instance, in a NEE of the impact of street lighting reductions on the public health, although change in road injury and crime could be identified through routine data sets, there was limited available data on other important outcomes, such as public views of impact on wellbeing and on changes in behaviour as a result of lighting reduction: a range of qualitative methods, including intercept interviews, focus groups and media analysis was used to document these.(17)

1. **Understanding change, mechanisms, and mediators**

Whilst sophisticated quantitative analyses can explore points on causal pathways, qualitative methods have particular strengths in exploring processes through: observing causal mechanisms in action; understanding processes of change; and making analytical inferences from comparisons within the case, through techniques such as process tracing or analytic induction. For instance, Alvarado et al(18) used process tracing to test the plausibility of the ‘signalling effect’ as a mechanism through which taxes on sugar-sweetened drinks in Barbados reduced sales; and Green et al(19) in an evaluation of the introduction of free bus travel for young people in London, used comparisons within the qualitative data (such as deviant cases who reported lack of access to a pass or lack of ability to use buses easily) to help understand the mechanisms of change in context, suggesting that it was essential that the intervention was universal, and implemented in the context of improvements in the public transport system.

Qualitative evidence is often invoked as particularly necessary for explaining unexpected findings or policy failures. In a NEE which found no evidence that Scotland’s introduction of a lower drink-drive limit had reduced RTAs, for instance, Lewsey et al(20) suggested that qualitative evaluations of public views would have been useful to understand why expected reductions did not occur. Benton et al(21) drew on evidence from walking interviews and photo-elicitation(22) to explain why their evaluation found no evidence that low-cost improvements to local urban amenities improved older adults’ health behaviours or use of green space. However, if there is sufficient uncertainty about likely effects to justify undertaking an evaluation, both positive and negative findings require explanation: qualitative evidence should be sought to explain the mechanisms of expected effects or policy success as well as failure.

1. **Stakeholders’ perspectives on the intervention and impact**

Whilst all NE evaluations should include appropriate involvement of patients, publics, practitioners and policy makers affected by the intervention, qualitative research is also essential for understanding the perspectives and practices of all stakeholders, and understanding the role of these actors as an integral part of the complex system in which the intervention is implemented.

Qualitative methods have strengths for documenting and learning from the experiential knowledge of publics directly affected by the intervention, which may not reflect expectations or theory. Cummins et al,(23) for instance, identified that residents near the Olympic site in London felt safer from, rather than marginalised by, enhanced security in their neighbourhoods. Qualitative methods also help understand knowledge, perspectives, and practices of those implementing policies or programmes. This is essential data to identify the alterations that affect how interventions are implemented on the ground, and any unanticipated effects. Stakeholders’ perspectives and practices may change across the process of implementation and evaluation, and may interact with the intervention (and its evaluation) in more or less predictable ways.

Understanding public, policy and practitioner perspectives is also essential for informing future roll out, scale up, or transferability of interventions. In a NEE of the potential for home energy interventions to improve the public health, for instance, Armstrong et al(24) used household interviews with those who had received the interventions to understand their motivations for installation: they found that policy framings around ‘environmental sustainability’ resonated poorly with householders’ motivations, suggesting that this would not encourage further take up of similar schemes.

1. **Maximising value from qualitative components**

Whilst descriptive content analysis of data may be sufficient for some of the above aims, such as documenting key aspects of context or stakeholders’ perspectives, maximising the value from qualitative components of an NEE usually entails more detailed analysis, appropriately integrated into the whole evaluation, from design to final report. This requires: planning of workstreams such that data generation and analyses can be incorporated at key points (e.g., before final selection of outcome indicators, to inform sub-group analyses); sufficient time for integrating the final analysis; design and management of the qualitative components by an investigator with appropriate skills and training; and a sound theoretical framing appropriate to the evaluation.

There may be under-exploited potential for utilising existing qualitative data in NEEs. On some common outcome indictors (e.g. around the meanings of travel mode, physical activity, smoking cessation) there may be existing datasets of transcripts from comparable settings which can be interrogated to enhance understanding prior to finalising evaluation design or outcome selection. Existing qualitative studies may also provide evidence to support hypothesised mechanisms. Archiving newly generated data is good practice, as these data sets can be used for comparisons in future follow-ups, or for comparative studies.

**Key points for guidance:**

* Ideally, most NEEs will include a substantial qualitative component, that is led by a senior investigator with expertise in qualitative design, and is integrated throughout the evaluation and in the report. A justification should be provided if no qualitative research is included.
* Qualitative data generated in a NEE should be archived where possible to facilitate further follow up, comparative studies and future evaluations.
* Final reports and theories of change should account for the qualitative as well as quantitative findings.
* Well-analysed qualitative data can enhance ability to make credible claims about causal relationships between intervention and outcomes; understand how the intervention affected the system; and generalise beyond the case.

**Q1.** Do you agree with the content of the proposed **Qualitative methods** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

**Q2.** Are there good examples for the use of qualitative methods within NEEs in Low/Middle Income countries which you would recommend? Please provide details in the comments section below.

**Q3.** Are there good examples of the use of qualitative methods within NEEs in health services/systems research which you would recommend? Please provide details in the comments section below.

[*free text box]*

References

1. Paparini S, Green J, Papoutsi C, Murdoch J, Petticrew M, Greenhalgh T, et al. Case study research for better evaluations of complex interventions: rationale and challenges. BMC Medicine. 2020;18(1):301.

2. Shaw S, Green J, et al. TRIPLE C Reporting Principles for Case study evaluations of the role of Context in Complex interventions [Manuscript submitted for publication]. forthcoming.

3. Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. BMC Medicine. 2016;14(1):96.

4. Moore GF, Audrey S, Barker M, Bond L, Bonell C, Hardeman W, et al. Process evaluation of complex interventions: Medical Research Council guidance. BMJ. 2015;350.

5. Hanckel B, Ruta D, Scott G, Peacock JL, Green J. The Daily Mile as a public health intervention: a rapid ethnographic assessment of uptake and implementation in South London, UK. BMC Public Health. 2019;19(1):1-14.

6. Campbell M, Katikireddi SV, Hoffmann T, Armstrong R, Waters E, Craig P. TIDieR-PHP: a reporting guideline for population health and policy interventions. Bmj. 2018;361.

7. Craig P, Di Ruggiero E, Frolich KL, Mykhalovskiy E, White M, Campbell R, et al. Taking account of context in population health intervention research: guidance for producers, users and funders of research. Southampton: National Institute for Health Research; 2018.

8. Christie N, Steinbach R, Green J, Mullan MP, Prior L. Pathways linking car transport for young adults and the public health in Northern Ireland: a qualitative study to inform the evaluation of graduated driver licensing. BMC Public Health. 2017;17(1):1-14.

9. Penney T, Adams J, Briggs A, Cummins S, Harrington R, Monsivais P, et al. Evaluation of the impacts on health of the proposed UK industry levy on sugar sweetened beverages: developing a systems map and data platform, and collection of baseline and early impact data. National Institute for Health Research. 2018.

10. Ogilvie D, Panter J, Guell C, Jones A, Mackett R, Griffin S. Health impacts of the Cambridgeshire Guided Busway: a natural experimental study. Public Health Research. 2016;4(1).

11. Jones CH, Cohn S, Ogilvie D. Making sense of a new transport system: an ethnographic study of the Cambridgeshire Guided Busway. PLoS One. 2013;8(7):e69254.

12. de Vocht F, McQuire C, Brennan A, Egan M, Angus C, Kaner E, et al. Evaluating the causal impact of individual alcohol licensing decisions on local health and crime using natural experiments with synthetic controls. Addiction. 2020;115(11):2021-31.

13. Gibson K, Pollard TM, Moffatt S. Social prescribing and classed inequality: A journey of upward health mobility? Social Science & Medicine. 2021;280:114037.

14. Moffatt S, Wildman J, Pollard TM, Penn L, O’Brien N, Pearce MS, et al. Evaluating the impact of a community-based social prescribing intervention on people with type 2 diabetes in North East England: mixed-methods study protocol. BMJ Open. 2019;9(1):e026826.

15. Jones A, Steinbach R, Roberts H, Goodman A, Green J. Rethinking passive transport: Bus fare exemptions and young people's wellbeing. Health & Place. 2012;18(3):605-12.

16. Guell C, Panter J, Ogilvie D. Walking and cycling to work despite reporting an unsupportive environment: insights from a mixed-method exploration of counterintuitive findings. BMC Public Health. 2013;13(1):1-10.

17. Perkins C, Steinbach R, Tompson L, Green J, Johnson S, Grundy C, et al. What is the effect of reduced street lighting on crime and road traffic injuries at night? A mixed-methods study. NIHR Journals Library. 2015;3(11).

18. Alvarado M, Penney TL, Unwin N, Murphy MM, Adams J. Evidence of a health risk ‘signalling effect’ following the introduction of a sugar-sweetened beverage tax. Food Policy. 2021;102:102104.

19. Green J, Roberts H, Petticrew M, Steinbach R, Goodman A, Jones A, et al. Integrating quasi-experimental and inductive designs in evaluation: A case study of the impact of free bus travel on public health. Evaluation. 2015;21(4):391-406.

20. Lewsey J, Haghpanahan H, Mackay D, McIntosh E, Pell J, Jones A. Impact of legislation to reduce the drink-drive limit on road traffic accidents and alcohol consumption in Scotland: a natural experiment study. Public Health Research. 2019;7(12).

21. Benton JS, Cotterill S, Anderson J, Macintyre VG, Gittins M, Dennis M, et al. Impact of a low-cost urban green space intervention on wellbeing behaviours in older adults: a natural experimental study. Wellbeing, Space and Society. 2021;2:100029.

22. Macintyre VG, Cotterill S, Anderson J, Phillipson C, Benton JS, French DP. “I Would Never Come Here Because I’ve Got My Own Garden”: Older Adults’ Perceptions of Small Urban Green Spaces. International Journal of Environmental Research and Public Health. 2019;16(11):1994.

23. Cummins S, Clark C, Lewis D, Smith N, Thompson C, Smuk M, et al. The effects of the London 2012 Olympics and related urban regeneration on physical and mental health: the ORiEL mixed-methods evaluation of a natural experiment. Public Health Research. 2018;6(12).

24. Armstrong B, Bonnington O, Chalabi Z, Davies M, Doyle Y, Goodwin J, et al. The impact of home energy efficiency interventions and winter fuel payments on winter-and cold-related mortality and morbidity in England: a natural equipment mixed-methods study. Public Health Research. 2018;6(11):1-110.

**Guidance section: Critical appraisal and evidence synthesis**

|  |
| --- |
| **Critical appraisal and evidence synthesis summary points**   * Clearly defining the purpose of a given synthesis is essential in order to make the best use of the available studies and the most appropriate methods to apply. * Critical appraisal of the studies involves selecting the tool or tools most suitable for the included natural experimental study designs. * While diversity within and between natural experimental studies presents opportunities to produce valuable explanatory findings, that diversity also presents challenges for finding, managing, and interpreting the studies during the synthesis process. |

**Critical appraisal and evidence synthesis**

**Issues specific to identifying, appraising, and synthesising evidence from natural experimental evaluations**

Bringing together evidence from a set of natural experimental evaluations (NEEs) requires consideration of issues throughout the critical appraisal and evidence synthesis process. Defining the overarching research question(s) for a given evidence synthesis, i.e. why a given set of studies is being synthesised, will inform decisions on how best to do it.(1) Each element of the scope of a conventional effectiveness synthesis, i.e., the populations, interventions, comparators, and outcomes (PICO), may be difficult to define, e.g. because of a lack of standardisation of intervention or exposure and control conditions, or multiplicity of outcomes in various formats, across the studies. Assessing effectiveness in a narrow sense, for example by producing a pooled effect size estimate using meta-analysis, may not be the most useful question or even an answerable one. It may be more valuable to examine effectiveness in a broader sense,(2) and in some cases the focus of a synthesis may be, for example, to establish causality, describe lived experiences, or explore mechanisms of action. Investigating health equity requires consideration of potentially disadvantaged populations from conception of the review.(3) Often these research questions may be most effectively addressed by incorporating multiple study designs in the review, for example using a mixed-method design incorporating a combination of quantitative and qualitative evidence.(4, 5) Any evidence synthesis should have a protocol to demonstrate that the results are not selectively chosen. With a synthesis of NEEs, the protocol may require to be dynamic and flexible to transparently record necessary alterations to the analysis plan as details of the included studies emerge.

1. **Identifying NEE evidence**

Natural experimental evaluations include various study designs and analytic approaches, and are not consistently labelled as natural experimental evaluations in titles, abstracts or keywords. There is not the simplicity associated with identifying RCTs, i.e., there is not the ease of using formalised bibliographic search (e.g. MeSH) terms or familiar text strings to filter a literature search by study design. Relevant studies in a synthesis of natural experimental evaluations are likely to comprise multiple study designs. As study design labelling is generally inconsistent, it is useful to focus on study design features rather than labels.(6)

Evidence may be published across disciplines and across databases, requiring searching databases across differing disciplines which exacerbates issues such as differing study design terminology between disciplines, e.g., epidemiology and economics. In some disciplines, e.g., economics, it is common for evidence to be published as reports, ‘grey literature’, rather than in peer-reviewed journals. Designing the literature search strategy for a review of NEEs may be further complicated by the multifaceted nature of many NEEs, e.g. the reporting of multiple analyses or mixed methods from a single study. Defining relevant time periods may be difficult as NEEs are often examining a NE that occurred at a previous time, possibly decades prior, to that of the evaluation. These issues have implications for the literature search strategy, achieving a balance between the sensitivity and the specificity of the search strategy.

1. **Critical appraisal of NEEs**

When critically appraising a NEE, whether to understand the rigour of an individual study or as part of a synthesis of evidence, reviewers may wish to use tools for assessing risk of bias. However, it should be noted that these typically assume a focus on a minimally-biased primary effect size estimate as the main objective of each study. With that caveat, the Risk Of Bias In Non-Randomized Studies - of Interventions (ROBINS-I)(7) is currently the most thorough of the critical appraisal tools currently available, although applying it requires expertise or significant time to learn the process.(8, 9) However, some features of NEEs can be problematic to fit into this type of appraisal process. When using ROBINS-I, few NEEs are likely to be assessed as at any less than ‘moderate’ risk of bias, despite some study designs going to considerable lengths to account for issues such as selection and confounding. A tool that produces an insufficient spread of studies from stronger to weaker may have limited value for discriminating between studies in a given set. Depending on how tightly NEEs are being defined, there may be large differences in the level of internal validity of the NEEs. Alternatively, if they are more methodologically homogeneous there may be questions about their overall external validity in terms of any ‘evaluative bias’ in the exclusion of studies of interventions that fail to meet narrowly defined design criteria.(10)

Where studies are multifaceted, tools such as ROBINS-I and the target trial framework on which it is based have no obvious way of dealing with the different components of a study.(11, 12) While ROBINS-I may be the most obvious off-the-shelf tool available, therefore, the tools available should be considered in terms of their strengths and limitations for the purpose of a given review. Adaptations of ROBINS-I are being developed to better capture features of methodological approaches commonly used in NEEs such as regression discontinuity or instrumental variables designs.(13) For some types of NEEs the alternative ROBINS-E, which is designed to assess the risk of bias in observational epidemiological studies, may be more appropriate.(14) Other tools that may be useful include the Effective Public Health Practice Project (EPHPP) quality assessment tool,(15) and the Cochrane Risk of Bias 2 with Cochrane Effective Practice and Organisation of Care (EPOC) guidance notes.(16) For some types of NEE, and some systematic reviews, an appraisal framework more like those used for qualitative research may be more appropriate.(17, 18)

1. **Synthesising results from NEEs**

Synthesising NEEs contributes to making best use of available studies. The likely diversity in these studies provides opportunities for original research such as identifying and exploring intervention mechanisms and investigating and refining intervention theories. However, a diverse body of studies also presents challenges in synthesising the range of characteristics of the NEE designs, diversity of the NEE content, context, and outcomes studied. Some NEEs may report multiple effect estimates for the same outcome, multiple outcomes, or both, for example, in transport studies there are often trade-offs between competing ‘good’ outcomes, e.g. more cycling at the expense of walking.(19) The included NEEs may report different types of estimands, outcomes of treatment effect, that relate to different types of effect estimate (e.g. LATE vs ATT vs ATE) and to different population subgroups and different outcome measures, often dependent on data availability. There may be different lengths of follow-up for assessing an outcome, and outcome measurements may be presented in different configurations across studies of the same design. Where replication of highly context-specific interventions is unlikely, it may be more realistic to make inferences about general causal mechanisms or functions of interventions (e.g. Ogilvie et al(20)). Investigating health inequity requires study data to be able to be grouped to make appropriate comparisons, which should be clearly reported.(21)

A particular challenge with data extraction relates to the diversity of the NEE studies included in many syntheses. It may be difficult to determine which effect estimate to extract from the multiplicity reported in a given study.(22) Details of the natural experiments and important information about the context need to be captured uniformly to enable synthesis.(23, 24)

While meta-analysis is potentially useful in some reviews, included studies may differ widely in terms of the nature of their potential biases. Particularly when a meta-analysis may be dominated by a few studies based on very large sample sizes (e.g. drawn from large-scale population administrative data) it may be at least as important to triangulate findings across studies that are at risk of differing biases than to rely on a single pooled effect size. Synthesis approaches that do not use meta-analysis, such as guidance provided by Cochrane(25, 26) and the Realist and Meta-narrative Evidence Syntheses: Evolving Standards (RAMESES) group(27) may often be useful when synthesising NEEs. The specifics of the evidence synthesis questions and studies involved will determine the appropriate methods to use, with a mixed methods approach often useful.(4, 5) On some occasions the focus of the review may be to combine effect estimates with simulation models(28) or to conduct a synthesis of economic evidence.(29)

Exploring heterogeneity between the NEE results can result in valuable insights from the synthesis, developing potential hypotheses about similarities and differences in mechanisms and study characteristics between the NEEs. The exploration of heterogeneity may be through qualitative methods, triangulation methods, or quantitative methods when these are statistically feasible and substantively meaningful.(30)

1. **Assessing certainty of evidence**

Depending on the focus of the evidence synthesis, it may be appropriate to formally assess and summarise the overall certainty of the findings, (i.e. how confident we are in estimating an effect), although in other situations a more appropriate objective may be to better understand the problem.(1) When assessing certainty of the evidence is appropriate, the framework most often used is Grading of Recommendations, Assessment, Development and Evaluations (GRADE).(31) Applying this tool with NEEs raises some issues similar to those noted above when assessing risk of bias. These include ensuring evidence from NEEs is appropriately acknowledged in GRADE by appropriately assessing risk of bias across the evidence (paying attention to study design features), incorporating a variety of differing study designs, and selecting outcomes for synthesis.(32, 33) The lack of scale available within GRADE can make it difficult to differentiate between certainty assessments.(32)

**Q1.** Do you agree with the content of the proposed **Critical appraisal and evidence synthesis** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

[free text box]

References

1. Ogilvie D, Adams J, Bauman A, Gregg EW, Panter J, Siegel KR, et al. Using natural experimental studies to guide public health action: turning the evidence-based medicine paradigm on its head. Journal of Epidemiology and Community Health. 2020;74(2):203-8.

2. Petticrew M, Rehfuess E, Noyes J, Higgins JP, Mayhew A, Pantoja T, et al. Synthesizing evidence on complex interventions: how meta-analytical, qualitative, and mixed-method approaches can contribute. Journal of clinical epidemiology. 2013;66(11):1230-43.

3. O'Neill J, Tabish H, Welch V, Petticrew M, Pottie K, Clarke M, et al. Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health. Journal of clinical epidemiology. 2014;67(1):56-64.

4. Sandelowski M, Voils CI, Barroso J. Defining and designing mixed research synthesis studies. Research in Schools. 2006;13(1):29.

5. Heyvaert M, Hannes K, Onghena P. Using mixed methods research synthesis for literature reviews: the mixed methods research synthesis approach. London: Sage Publications; 2016.

6. Reeves BC, Wells GA, Waddington H. Quasi-experimental study designs series—paper 5: a checklist for classifying studies evaluating the effects on health interventions—a taxonomy without labels. Journal of Clinical Epidemiology. 2017;89:30-42.

7. Sterne JA, Hernán MA, Reeves BC, Savović J, Berkman ND, Viswanathan M, et al. ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions. BMJ. 2016;355:i4919.

8. Thomson H, Craig P, Hilton-Boon M, Campbell M, Katikireddi SV. Applying the ROBINS-I tool to natural experiments: an example from public health. Systematic Reviews. 2018;7(1):15.

9. Igelström E, Campbell M, Craig P, Katikireddi SV. Cochrane's risk of bias tool for non-randomized studies (ROBINS-I) is frequently misapplied: A methodological systematic review. Journal of Clinical Epidemiology. 2021;140:22-32.

10. Ogilvie D, Egan M, Hamilton V, Petticrew M. Systematic reviews of health effects of social interventions: 2. Best available evidence: how low should you go? Journal of Epidemiology & Community Health. 2005;59(10):886-92.

11. Waddington H, Aloe AM, Becker BJ, Djimeu EW, Hombrados JG, Tugwell P, et al. Quasi-experimental study designs series—paper 6: risk of bias assessment. Journal of Clinical Epidemiology. 2017;89:43-52.

12. Hilton Boon M, Burns J, Craig P, Griebler U, Heise TL, Vittal Katikireddi S, et al. Value and Challenges of Using Observational Studies in Systematic Reviews of Public Health Interventions. American Journal of Public Health. 2022;112(4):548-52.

13. Villar PF, Waddington H. Within study comparisons and risk of bias in international development: Systematic review and critical appraisal. Campbell Systematic Reviews. 2019;15(1-2):e1027.

14. ROBINS-E Development Group. Risk Of Bias In Non-randomized Studies - of Exposure (ROBINS-E) 2022 [Available from: https://www.riskofbias.info/welcome/robins-e-tool.

15. Effective Public Health Practice Project. Quality assessment tool for quantitative studies. Hamilton, Canada [Available from: http://www.ephpp.ca/PDF/Quality%20Assessment%20Tool\_2010\_2.pdf.

16. Cochrane Effective Practice and Organisation of Care (EPOC). Suggested risk of bias criteria for EPOC reviews. EPOC resources for review authors. 2017 [Available from: https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/suggested\_risk\_of\_bias\_criteria\_for\_epoc\_reviews.pdf.

17. Noyes J, Booth A, Flemming K, Garside R, Harden A, Lewin S, et al. Cochrane Qualitative and Implementation Methods Group guidance series—paper 3: methods for assessing methodological limitations, data extraction and synthesis, and confidence in synthesized qualitative findings. Journal of Clinical Epidemiology. 2018;97:49-58.

18. Critical Appraisal Skills Programme (CASP). Qualitative Checklist. [online] 2018 [cited 2022 July 2022]. Available from: https://casp-uk.b-cdn.net/wp-content/uploads/2018/03/CASP-Qualitative-Checklist-2018\_fillable\_form.pdf.

19. Ogilvie D, Egan M, Hamilton V, Petticrew M. Promoting walking and cycling as an alternative to using cars: systematic review. BMJ. 2004;329(7469):763.

20. Ogilvie D, Bauman A, Foley L, Guell C, Humphreys D, Panter J. Making sense of the evidence in population health intervention research: building a dry stone wall. BMJ Global Health. 2020;5(12):e004017.

21. Welch V, Petticrew M, Petkovic J, Moher D, Waters E, White H, et al. Extending the PRISMA statement to equity-focused systematic reviews (PRISMA-E 2012): explanation and elaboration. Journal of Clinical Epidemiology. 2016;70:68-89.

22. López‐López JA, Page MJ, Lipsey MW, Higgins JP. Dealing with effect size multiplicity in systematic reviews and meta‐analyses. Research Synthesis Methods. 2018;9(3):336-51.

23. Campbell M, Katikireddi SV, Hoffmann T, Armstrong R, Waters E, Craig P. TIDieR-PHP: a reporting guideline for population health and policy interventions. Bmj. 2018;361.

24. Craig P, Di Ruggiero E, Frolich KL, Mykhalovskiy E, White M, Campbell R, et al. Taking account of context in population health intervention research: guidance for producers, users and funders of research. Southampton: National Institute for Health Research; 2018.

25. McKenzie JE, Brennan SE. Synthesizing and presenting findings using other methods. Cochrane handbook for systematic reviews of interventions. 2019:321-47.

26. Noyes J, Booth A, Cargo M, Flemming K, Harden A, Harris J, et al. Qualitative evidence. Cochrane Handbook for systematic reviews of interventions. 2019:525-45.

27. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review-a new method of systematic review designed for complex policy interventions. Journal of Health Services Research & Policy. 2005;10(1\_suppl):21-34.

28. Brozek JL, Canelo-Aybar C, Akl EA, Bowen JM, Bucher J, Chiu WA, et al. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence—An overview in the context of health decision-making. Journal of Clinical Epidemiology. 2021;129:138-50.

29. Shemilt I, Mugford M, Vale L, Marsh K, Donaldson C. Evidence-based decisions and economics: health care, social welfare, education and criminal justice: John Wiley & Sons; 2011.

30. Deeks JJ, Higgins JP, Altman DG, Group CSM. Analysing data and undertaking meta‐analyses. Cochrane handbook for systematic reviews of interventions. 2019:241-84.

31. Guyatt G, Oxman AD, Akl EA, Kunz R, Vist G, Brozek J, et al. GRADE guidelines: 1. Introduction—GRADE evidence profiles and summary of findings tables. Journal of Clinical Epidemiology. 2011;64(4):383-94.

32. Hilton Boon M, Thomson H, Shaw B, Akl EA, Lhachimi SK, López-Alcalde J, et al. Challenges in applying the GRADE approach in public health guidelines and systematic reviews: a concept article from the GRADE Public Health Group. Journal of Clinical Epidemiology. 2021;135:42-53.

33. Cuello-Garcia CA, Santesso N, Morgan RL, Verbeek J, Thayer K, Ansari MT, et al. GRADE guidance 24 optimizing the integration of randomized and non-randomized studies of interventions in evidence syntheses and health guidelines. Journal of Clinical Epidemiology. 2022;142:200-8.

**Guidance section:** **Infrastructure and information governance**

|  |
| --- |
| **Infrastructure and information governance summary points**   * Natural experimental studies often use data that was originally collected for other purposes. * Negotiating access to such datasets can be a time-consuming, costly and uncertain process, especially if the research involves the linkage of data from multiple sources. * The Trusted Research Environment model of curation and access of routinely collected data, of which there are already a number of good examples, is a potentially much more efficient solution than ad hoc linkages initiated by research teams. |

Natural experimental evaluations are often conducted as retrospective studies, using data originally collected for another purpose, such as vital events registration,(1) population surveys,(2) administrative datasets,(3) sales and purchasing data(4) and, increasingly, information gathered via mobile phones, fitness apps, and other forms of ‘crowd-sourced’ data.(5, 6) Such studies can be conducted when a prospective trial would no longer be practical (or would never have been possible) and can be highly efficient, where large datasets are available at a tiny fraction of the cost of primary data gathering. There has been extensive investment in the infrastructure for making such data available to researchers (Box 1), but outside of these settings, acquisition of data can be complicated, uncertain and time consuming.

Some of the most valuable datasets for natural experimental evaluations are those that link exposure and outcome data from different sources – for example, information on receipt of welfare benefits from the social security system with information on use of services from the health and social care system. But linking data from different sources can multiply the time and effort needed to obtain a research-ready dataset if research culture, information governance procedures and methods of identifying records differ from data owner to data owner. The consequences are to restrict the opportunities to conduct research using cross-sectoral linked datasets to a relatively small number of researchers with the time, resources and experience to deal with the extra complexities, and ultimately to limit the amount of research that is done using such datasets.

**Box 1 The SAIL Databank**

|  |
| --- |
| The Secure Anonymised Information Linkage (SAIL) Databank (<https://saildatabank.com/>) is a repository of anonymised health and social data based at Swansea University that seeks to help ‘research communities to access, link and analyse routinely collected health and administrative data within a safe and secure remote access environment.’ To access SAIL data researchers must follow a two stage application process that takes about 12 weeks. The first stage involves a scoping discussion to establish the viability of the project and what is required in terms of support from the Databank. At this stage the researcher must also undertake Safe Researcher Training, or provide evidence that they have has the training within the past two years, and show that they have secured funding for their project. At the second stage, the researcher submits a project scoping document and an Information Governance Review Panel application form. Following approval of the project, members of the research team can apply for approval to access the data remotely, to conduct analysis using a range of pre-installed statistical packages. A separate procedure is available to organisations wishing to add datasets to the databank. SAIL’s governance makes extensive use of public involvement, including involvement of members of the public in the review of applications. |

**Box 1b New Zealand’s Integrated Data Infrastructure (IDI)**

The IDI (<https://www.stats.govt.nz/integrated-data/integrated-data-infrastructure/>) is a research database that links together and holds individual-level de-identified data on New Zealand’s population from surveys, government agencies, the Census and non-government organisations on health, education and training, income and work, benefits and social services, justice, vital events, border movements, housing and transport. Researchers can apply to use the data for research ‘that is likely to have a wide public benefit.’ They must undergo a series of checks, undertake confidentiality training and sign a declaration of secrecy and an agreement to follow IDI rules. The data can only be accessed through a secure virtual environment, via facilities approved by Statistics New Zealand, but researchers can apply to set up a secure research facility in their own workplace, subject to meeting specified security conditions. Outputs are subject to disclosure control ‘to ensure information is grouped in a way that makes it impossible to identify individuals’ before they are released from the secure virtual environment. Statistics New Zealand supports applicants through the review process by advising on draft applications before submission and also provides a separate application pathway for organisations that lack the technical expertise to work in a secure virtual environment. As with SAIL there is a separate procedure for organisations that wish to add datasets to the IDI.

SAIL and the IDI are examples of ‘Trusted Research Environments’ (TREs), research platforms that seek to resolve the tension between maintaining the security and confidentiality of the data and providing efficient access to researchers. They do this by holding the data in a safe setting which researchers can access in order to conduct their analyses, rather than by distributing the data to researchers who then store and analyse it within their own organisation. The perceived risk of disclosure associated with the latter approach was found by a recent review of the use of health data for research in the UK to have led to a burdensome and inefficient system for providing access that frustrates researchers while failing to adequately reassure a substantial minority of clinicians and patients that the process is trustworthy.(7)

The TRE model has a number of potential advantages, beyond greater security, over the distributed model. A system comprising a small number of substantial well-funded TREs can deliver economies of scale, so that data can be curated to a high standard at low cost and allows for a concentration of expertise that can help to drive technical and methodological advances. An established TRE should also be in a strong position to negotiate the acquisition of new datasets. Such a system should therefore be an attractive investment for research funders, with more favourable returns than grants for research projects involving bespoke data linkages.

Under the TRE model, researchers stand to benefit from a streamlined process of applying for access, with less risk of complicated or inconsistent requirements for training, accreditation and storage of data imposed by different data providers. Other key characteristics of a system that works well from a researcher’s perspective are good quality metadata, including information on flaws in the data such as missingness, changes in variable definitions over time, etc.; support through the application process from TRE staff; a transparent pricing structure, with prices set on a cost recovery basis; an efficient system of disclosure control; and the ability to access a TRE and run analyses from within their own workplace. In turn, researchers working with TRE data should be required to adopt open working practices, and in particular to share the code used in their analyses.

**Q1.** Do you agree with the content of the proposed **Infrastructure and information governance** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

[free text box]

References

1. Degli Esposti M, Wiebe DJ, Gasparrini A, Humphreys DK. Analysis of “Stand Your Ground” Self-defense Laws and Statewide Rates of Homicides and Firearm Homicides. JAMA Network Open. 2022;5(2):e220077-e.

2. Goodman A, van Sluijs EMF, Ogilvie D. Impact of offering cycle training in schools upon cycling behaviour: a natural experimental study. International Journal of Behavioral Nutrition and Physical Activity. 2016;13(1):34.

3. Grundy C, Steinbach R, Edwards P, Green J, Armstrong B, Wilkinson P. Effect of 20 mph traffic speed zones on road injuries in London, 1986-2006: controlled interrupted time series analysis. BMJ. 2009;339.

4. Robinson M, Mackay D, Giles L, Lewsey J, Richardson E, Beeston C. Evaluating the impact of minimum unit pricing (MUP) on off-trade alcohol sales in Scotland: an interrupted time–series study. Addiction. 2021;116(10):2697-707.

5. Hong J, McArthur DP, Livingston M. The evaluation of large cycling infrastructure investments in Glasgow using crowdsourced cycle data. Transportation. 2020;47(6):2859-72.

6. Hong J, McArthur DP, Stewart JL. Can providing safe cycling infrastructure encourage people to cycle more when it rains? The use of crowdsourced cycling data (Strava). Transportation research part A: policy and practice. 2020;133:109-21.

7. Goldacre B, Morley J. Better, broader, safer: using health data for research and analysis. Department of Health and Social Care; 2022.

**Guidance section: Good practice considerations**

This section is a draft that summarises our main messages for planning, commissioning, conducting, reporting and using evidence from natural experimental evaluations. We have grouped the recommendations according to their key audience, but many will be relevant to more than one group of producers or users of natural experimental evidence.

**Good practice considerations for all producers and users of natural experimental evaluations**

Understand the design and planning processes of an evaluation of a natural experiment, including how to identify opportunities for natural experimental evaluation, select the best evaluation approach and assess the feasibility of the evaluation.

Consider stakeholders: for natural experimental evaluations, there may be stakeholders in the natural experiment who differ from those with a stake the evaluation of that natural experiment; some stakeholders may have a conflict of interest if involved in both the natural experiment and the corresponding evaluation.

Recognise the respective strengths of quantitative, qualitative, and integrated analysis approaches for investigating the impacts of natural experiments.

**Conducting natural experimental evaluations (mainly for researchers)**

Be aware of the circumstances that are likely to give rise to good opportunities for a natural experimental approach. Adopt methods that are appropriate to the data available and to the processes that determine exposure to the intervention of interest. There is no single method that is best in all circumstances.

Natural experimental evaluations will usually be stronger if they use a combination of methods, including alternative methods of effect estimation, robustness checking and a mix of qualitative and quantitative methods to understand how effects occur and how they are influenced by context.

Adopt open science practices, including publication of a study protocol, analysis code and data (or an indication of where the data can be obtained). Publish in open access journals or on other open platforms.

Clearly report the natural experiment event and all stages of the evaluation, including planning, protocol, analyses, and results, using established reporting standards where available (Table 1). Report the results of all the planned analyses or explain why any of those indicated in the protocol were not progressed.

Evaluation of the strength of evidence from natural experimental evaluations should be based on detailed assessment of risk of bias, not on broad study labels.

...

**Supporting and investing in natural experimental evaluations (mainly for research funders and commissioners)**

Encourage best practice when commissioning or funding natural experimental evaluations, for example by requiring that a protocol is available prior to analysis commencing, findings are published in open access journals, and the relevant reporting guidelines are followed.

As well as funding individual studies, support capacity building for natural experiments through investment in infrastructure (e.g., trusted research environments) and the workforce (e.g., training in evaluation methods, data science and research software engineering).

Negotiate with data owners to make routinely collected data available for natural experimental evaluations of policies and programmes.

...

**Publishing and using evidence from natural experimental evaluations (mainly for journal editors, policymakers, practitioners and other decision-makers)**

Provide guidance for authors and reviewers on requirements for reports of natural experimental studies.

Use evidence from high quality natural experimental evaluations when this is the most appropriate or available form of evidence, being aware of any limitations of the evaluation.

Incorporate evaluation plans into the implementation of new policies and programmes where there is uncertainty about impacts and cost-effectiveness, and make data generated through implementation available to evaluators.

...

**Table 1: Reporting guidance**

|  |  |
| --- | --- |
| **Reporting guidance** | **Focus** |
| SPIRIT 2013: Standard Protocol Items: Recommendations for Interventional Trials(1) | This guidance for reporting clinical trials protocols may be useful for identifying key features of the NEE to include in a protocol. A reporting guideline is in development for reporting protocols for observational studies (SPIROS)(2) |
| STROBE: Strengthening the Reporting of Observational Studies in Epidemiology(3) | A list of 22 items to aid the reporting of observational studies - cohort, case–control, or cross-sectional study design. |
| TREND: Transparent Reporting of Evaluations with Nonrandomized Designs(4) | Checklist for reporting nonrandomised behavioural and public health intervention evaluations. |
| RECORD: REporting of studies Conducted using Observational Routinely-collected health Data(5) | An extension to STOBE, this guidance provides a checklist (13 items) for studies using routinely collected data |
| TIDieR-PHP: Template for Intervention Description and Replication - population health and policy(6) | An adaptation of the TIDieR guidance,(7) TIDieR-PHP is a reporting guideline for evaluation studies of population health and policy interventions, such as legal, fiscal, structural, organisational, environmental, or policy interventions |
| CHEERS 2022: Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022)(8) | Guidance and 28 item checklist for reporting economic evaluations of health interventions |
| SRQR: Standards for Reporting Qualitative Research(9) | Guidance for reporting qualitative research that is intended to be relevant across differing paradigms and methods |
| PRISMA-P: Preferred Reporting Items for Systematic reviews and Meta-Analyses for Protocols 2015(10) | Guidance for reporting systematic review protocols |
| PRISMA 2020(11) | Guideline and checklist to promote complete reporting of systematic reviews |
| SWiM: Synthesis without Meta-analysis(12) | Guidance for reporting synthesis without meta-analysis in systematic reviews |

**Q1.** Do you agree with the content of the proposed **Good practice considerations** section?

* Agree
* Agree, but some additional content or explanation could be provided (please explain)
* Disagree (please explain)
* Don't know

Please explain any agreement, disagreement, or additional comments you have about the content of this section

[free text box]

**Q2.** Please suggest amendments and/or further recommendations that you think we should emphasise, based on the content of the earlier sections.

[free text box]

**References**

1. Chan A-W, Tetzlaff JM, Gøtzsche PC, Altman DG, Mann H, Berlin JA, et al. SPIRIT 2013 explanation and elaboration: guidance for protocols of clinical trials. BMJ : British Medical Journal. 2013;346:e7586.

2. Mahajan R, Burza S, Bouter LM, Sijtsma K, Knottnerus A, Kleijnen J, et al. Standardized protocol items recommendations for observational studies (SPIROS) for observational study protocol reporting guidelines: protocol for a Delphi study. JMIR Research Protocols. 2020;9(10):e17864.

3. Elm Ev, Altman DG, Egger M, Pocock SJ, Gøtzsche PC, Vandenbroucke JP. Strengthening the reporting of observational studies in epidemiology (STROBE) statement: guidelines for reporting observational studies. BMJ. 2007;335(7624):806-8.

4. Des Jarlais DC, Lyles C, Crepaz N, Group T. Improving the reporting quality of nonrandomized evaluations of behavioral and public health interventions: the TREND statement. American Journal of Public Health. 2004;94(3):361-6.

5. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, et al. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. PLoS Medicine. 2015;12(10):e1001885.

6. Campbell M, Katikireddi SV, Hoffmann T, Armstrong R, Waters E, Craig P. TIDieR-PHP: a reporting guideline for population health and policy interventions. Bmj. 2018;361.

7. Hoffmann TC, Glasziou PP, Boutron I, Milne R, Perera R, Moher D, et al. Better reporting of interventions: template for intervention description and replication (TIDieR) checklist and guide. BMJ. 2014;348.

8. Husereau D, Drummond M, Augustovski F, de Bekker-Grob E, Briggs AH, Carswell C, et al. Consolidated Health Economic Evaluation Reporting Standards 2022 (CHEERS 2022) statement: updated reporting guidance for health economic evaluations. International Journal of Technology Assessment in Health Care. 2022;38(1).

9. O’Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. Academic Medicine. 2014;89(9):1245-51.

10. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. BMJ. 2015;349.

11. Page MJ, Moher D, Bossuyt PM, Boutron I, Hoffmann TC, Mulrow CD, et al. PRISMA 2020 explanation and elaboration: updated guidance and exemplars for reporting systematic reviews. BMJ. 2021;372:n160.

12. Campbell M, McKenzie JE, Sowden A, Katikireddi SV, Brennan SE, Ellis S, et al. Synthesis without meta-analysis (SWiM) in systematic reviews: reporting guideline. BMJ. 2020;368.