NIHR | National Institute for Health and Care Research

Journals Library

Check for updates

Public Health Research

Volume 13 • Issue 3 • March 2025

ISSN 2050-439X

Using natural experiments to evaluate population health interventions: a framework for producers and users of evidence

Peter Craig, Mhairi Campbell, Manuela Deidda, Ruth Dundas, Judith Green, Srinivasa Vittal Katikireddi, Jim Lewsey, David Ogilvie, Frank de Vocht and Martin White

DOI 10.3310/JTYW6582

NIHR | National Institute for Health and Care Research

Public Health Research

Check for updates

Extended Research Article

Using natural experiments to evaluate population health interventions: a framework for producers and users of evidence

Peter Craig,[1*] Mhairi Campbell,[1] Manuela Deidda,[2] Ruth Dundas,[1] Judith Green,[3] Srinivasa Vittal Katikireddi,[1] Jim Lewsey,[2] David Ogilvie,[4] Frank de Vocht[5,6] and Martin White[4]

[1]MRC/CSO Social and Public Health Sciences Unit, School of Health and Wellbeing, University of Glasgow, Glasgow, UK
[2]Health Economics and Health Technology Assessment, School of Health and Wellbeing, University of Glasgow, Glasgow, UK
[3]Wellcome Centre for Cultures and Environments of Health, University of Exeter, Exeter, UK
[4]MRC Epidemiology Unit, University of Cambridge, Cambridge, UK
[5]Population Health Sciences, Bristol Medical School, University of Bristol, Bristol, UK
[6]NIHR Applied Research Collaboration West, Bristol, UK

[*]Corresponding author Peter.Craig@glasgow.ac.uk

Published March 2025
DOI: 10.3310/JTYW6582

This report should be referenced as follows:

Craig P, Campbell M, Deidda M, Dundas R, Green J, Katikireddi SV, *et al.* Using natural experiments to evaluate population health interventions: a framework for producers and users of evidence. *Public Health Res* 2025;**13**(3). https://doi.org/10.3310/JTYW6582

Public Health Research

PHR MRC-PHSG

ISSN 2050-439X (Online)

A list of Journals Library editors can be found on the NIHR Journals Library website

Public Health Research (PHR) was launched in 2013 and is indexed by Europe PMC, NCBI Bookshelf, DOAJ, INAHTA, Ulrichsweb™ (ProQuest LLC, Ann Arbor, MI, USA) and MEDLINE.

This journal is a member of and subscribes to the principles of the Committee on Publication Ethics (COPE) (www.publicationethics.org/).

Editorial contact: journals.library@nihr.ac.uk

The full PHR archive is freely available to view online at www.journalslibrary.nihr.ac.uk/phr.

Criteria for inclusion in the *Public Health Research* journal

Manuscripts are published in *Public Health Research* (PHR) if (1) they have resulted from work for the PHR programme or, commissioned/managed through the Methodology research programme (MRP), and (2) they are of a sufficiently high scientific quality as assessed by the reviewers and editors.

Reviews in *Public Health Research* are termed 'systematic' when the account of the search appraisal and synthesis methods (to minimise biases and random errors) would, in theory, permit the replication of the review by others.

PHR programme

The Public Health Research (PHR) programme, part of the National Institute for Health and Care Research (NIHR), is the leading UK funder of public health research, evaluating public health interventions, providing new knowledge on the benefits, costs, acceptability and wider impacts of non-NHS interventions intended to improve the health of the public and reduce inequalities in health. The scope of the programme is multi-disciplinary and broad, covering a range of interventions that improve public health.

For more information about the PHR programme please visit the website: https://www.nihr.ac.uk/explore-nihr/funding-programmes/public-health-research.htm

This article

This issue of the Public Health Research journal series contains a project commissioned by the Medical Research Council's (MRC) Population Health Sciences Group (PHSG). Jointly funded by the MRC and NIHR, the work updated and extended the MRC guidance on using natural experiments to evaluate population health interventions.

PHSG is responsible for developing the MRC's strategy for research to improve population health. NIHR's mission is to improve the health and wealth of the nation through research. As population level interventions in community and clinical settings become more important, and as science advances and innovates, both funding partners agreed that updating the existing framework was timely and needed.

The authors have been wholly responsible for all data collection, analysis and interpretation, and for writing up their work. The Public Health research (PHR) programme editors and publisher have tried to ensure the accuracy of the authors' report and would like to thank the reviewers for their constructive comments on the draft document. However, they do not accept liability for damages or losses arising from material published in this article.

This article presents independent research. The views and opinions expressed in this publication are those of the authors and do not necessarily reflect those of the MRC, the NIHR, NETSCC, the PHR programme or the Department of Health and Social Care. If there are verbatim quotations included in this publication the views and opinions expressed by the interviewees are those of the interviewees and do not necessarily reflect those of the authors, those of the NHS, the NIHR, the MRC, NETSCC, the PHR programme or the Department of Health and Social Care.

This article was published based on current knowledge at the time and date of publication. NIHR is committed to being inclusive and will continually monitor best practice and guidance in relation to terminology and language to ensure that we remain relevant to our stakeholders.

Published by the NIHR Journals Library (www.journalslibrary.nihr.ac.uk), produced by Newgen Digitalworks Pvt Ltd, Chennai, India (www.newgen.co).

Abstract

Background: There has been a substantial increase in the conduct of natural experimental evaluations in the last 10 years. This has been driven by advances in methodology, greater availability of large routinely collected datasets, and a rise in demand for evidence about the impacts of upstream population health interventions. It is important that researchers, practitioners, commissioners, and users of intervention research are aware of the recent developments. This new framework updates and extends existing Medical Research Council guidance for using natural experiments to evaluate population health interventions.

Methods: The framework was developed with input from three international workshops and an online consultation with researchers, journal editors, funding representatives, and individuals with experience of using and commissioning natural experimental evaluations. The project team comprised researchers with expertise in natural experimental evaluations. The project had a funder-assigned oversight group and an advisory group of independent experts.

Results: The framework defines key concepts and provides an overview of recent advances in designing and planning evaluations of natural experiments, including the relevance of a systems perspective, mixed methods and stakeholder involvement throughout the process. It provides an overview of the strengths, weaknesses, applicability and limitations of the range of methods now available, identifies issues of infrastructure and data governance, and provides good practice considerations.

Limitations: The framework does not provide detailed information for the substantial volume of themes and material covered, rather an overview of key issues to help the conduct and use of natural experimental evaluations.

Conclusion: This updated and extended framework provides an integrated guide to the use of natural experimental methods to evaluate population health interventions. The framework provides a range of tools to support its use and detailed, evidence-informed recommendations for researchers, funders, publishers, and users of evidence.

Study registration: This methodological project was not registered.

Funding: This project was jointly funded by the Medical Research Council (MRC) and National Institute for Health and Care Research (NIHR), with project reference MC_PC_21009. The work is published in full in *Public Health Research*; Vol. 13, No. 3.

Contents

List of tables

List of figures

List of boxes

List of abbreviations

ATE	average treatment effect
ATT	average treatment effect in the treated
CATE	complier average treatment effect
CBA	cost–benefit analysis
CCA	cost–consequence analysis
CEA	cost-effectiveness analysis
CIDACS	Centre for Data and Knowledge Integration for Health
cITS	controlled interrupted time series
CLD	causal loop diagram
CUA	cost–utility analysis
CVD	cardiovascular disease
DiD	difference in differences
GRADE	Grading of Recommendations, Assessment, Development and Evaluations
ITS	interrupted time series
LATE	local average treatment effect
MRC	Medical Research Council
QALYs	quality-adjusted life-years
QCA	qualitative comparative analysis
QoL	quality of life
RCT	randomised controlled trial
ROBINS-I	Risk Of Bias In Non-randomized Studies – of Interventions
TRE	trusted research environment

Plain language summary

Governments and other organisations often make changes, for example bringing in new laws, new taxes or changes in the way health care is organised. Changes like these, which are not made by researchers, can be called 'natural experiments'. As long as some people are affected by a change and some are not, researchers may be able to study the health effects of the changes anyway. We call this 'natural experimental evaluation'.

There is already some guidance on how to conduct this type of research, but methods are advancing constantly and this study needed to update the guidance in a new framework. The researchers formed a writing group to do this, made up of people with skills and experience in doing evaluations of natural experiments. The researchers also held workshops and consulted online with a wider group of experts, including people who use the findings of natural experimental evaluations to help decisions about making public policy. This wider group advised on what should be included in the framework. The writing group was assigned to write the final framework.

In this framework, the researchers explain key words and phrases. They also explain why it is important to have a broad definition of a natural experiment. The researchers outline key aspects to bear in mind when designing an evaluation. These include identifying the best opportunities for evaluations, understanding natural experiments within their real-world context, using a variety of research methods, obtaining data, involving stakeholders and various other practical issues. The researchers provide an overview of research methods than can be used, including quantitative, qualitative and economic methods and combinations of these methods. They also provide advice about combining evidence from more than one study. The framework will help people design and use evaluations of natural experiments so they can provide good scientific evidence, but also be as useful as possible for making decisions about how to protect and improve the health of populations.

Scientific summary

Background

Natural experiments, defined as events outside the control of researchers that divide populations into exposed and unexposed groups, are a valuable opportunity to evaluate population health and health system interventions. The conduct of evaluations of these natural experiments has increased substantially since guidance was first published by the UK Medical Research Council in 2012. This increase was due to advances in relevant methods, greater availability of large administrative or routinely collected datasets, and a rise in demand for evaluation of natural experimental interventions delivered at a population level. The original guidance and recent summaries of alternative designs for natural experimental studies have primarily focused on quantitative methods for measuring the size or effect of interventions. Therefore, there is a need for an updated and extended framework that provides additional information on designing and planning natural experimental evaluations, the role of qualitative, mixed methods and economic evaluation along with quantitative methods, and considerations for evidence synthesis and accessing and using routinely collected data.

The objective of this framework is to provide an integrated guide for the use of a natural experimental approach to evaluate population health interventions and to:

- raise awareness among researchers of the range of approaches available for evaluating interventions or other exposures as natural experiments;
- provide information to help intervention stakeholders, for example in local or central government, decide whether a natural experimental evaluation would be useful, and if so of what kind; and
- provide information to help journal editors, funders and peer-reviewers to understand the strengths and weaknesses of funding proposals for, and articles reporting, natural experimental evaluations.

Methods

To develop the framework, the researchers convened a writing group comprising population health researchers from a range of disciplines, including epidemiology, health economics, public health and sociology, and with methodological expertise in statistics, qualitative research and evidence synthesis. Firstly they developed a working draft of the framework, with each member of the group being responsible for drafting one or more chapter. The researchers then used online workshops and an online consultation to collect expert opinions on the content and coverage of the draft. Participants in the workshops (n = 21) and consultation (n = 44) comprised researchers and other relevant stakeholders in Europe, Africa, the Americas and Australasia, including members of research funding boards, journal editors, and representatives of national and local governments. The researchers asked participants in the consultation to review each section of the framework, with the additional feedback being used to further refine the content.

Results

This framework has a broader scope than the previously published guidance. Feedback from the workshops and online consultation was collated and used to revise the content by the writing group. This input helped guide the use of a broad definition of natural experiment, refine a framework for planning and conducting evaluations of natural experimental studies, and specify in detail the role and content of analytic methods in evaluations. The content of the framework presents information to consider when conducting or using natural experimental evaluations, as set out below.

Concepts and definitions

The study defines natural experiments as events or processes outside the control of a researcher that divide a population into groups with differing degrees of exposure. 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. The researchers argue that this broad definition is preferable to narrower definitions based on particular designs or methods, as such lists tend to be arbitrary, or on assignment being 'as-if randomised' which can be difficult to define or apply in practice. As methods originating from a range of disciplines are used in the evaluation of natural experiments, a glossary is provided to define key terms.

Design and planning

An adaptation of the MRC/NIHR framework, for the development and evaluation of complex interventions, provides a structure for planning and conducting a natural experimental evaluation and highlights the value of applying a complex systems perspective for the evaluation.

Important stages when scoping and planning a natural experimental evaluation are:

Identifying and theorising the natural experiment: opportunities include difference in time or place in the presence or level of exposure between otherwise similar subpopulations, policy eligibility criteria that identify some units within a population but not others as exposed, phased implementation of a policy, randomisation used to assign a policy, and flaws in policy implementation.

Assessing the evaluability of the natural experiment: using structured engagement with stakeholders to agree on a conceptual model of how the intervention is expected to achieve its impacts, access relevant data, and consider the costs and usefulness of the evidence. The assessment helps identify key uncertainties and limitations, increases the likelihood of obtaining intended results, and ensures that questions addressed are relevant for decision-making.

Conducting feasibility studies for the evaluation: assessing whether the evaluation is viable, and the practicalities of implementing the evaluation design, such as whether routinely collected data adequately captures the necessary exposures.

Natural experiments typically occur within complex systems that influence health. When evaluating a natural experiment, considering the implications of the context in which the natural experiment exists helps to understand why the intervention succeeds or fails to achieve the intended impact.

An evaluation design with a combination of methods, both qualitative and quantitative, is often needed to provide a comprehensive understanding of a natural experiment, with theory and planning required for how to bring different study designs, types of data and analyses can be brought together.

As with most evaluation designs, it is good practice to develop a protocol, or methods-appropriate advance study plan, and to place it in the public domain before analysis commences.

In evaluations of natural experiments there will be a diverse range of stakeholders involved at different stages of both the natural experiment intervention and the evaluation. Involvement of relevant stakeholders maximises the likelihood of findings being understood, taken up and used for decision-making.

Quantitative methods

The quantitative methods used will be defined by the research question being investigated and the design features of the evaluation, with a complex systems perspective likely to require a combination of qualitative methods alongside the quantitative methods. A variety of quantitative methods will often be required to address threats to internal validity of the non-randomised natural experiment. An overview of key quantitative methods is provided, avoiding a hierarchy as the choice of methods should be determined by the research questions and the availability of data, with each method having strengths and weaknesses and therefore appropriateness of use determined by the circumstances of the evaluation.

Economic evaluation

As there are usually resource constraints for implementations of policies, economic evaluations are valuable in conjunction with evaluations of effectiveness of the natural experiment. Designing, conducting and reporting economic evaluations generates specific challenges, including measuring and identifying costs and outcomes, selecting appropriate analytical methods, identifying the time horizon and considerations of equity.

Qualitative methods

Qualitative methods make an essential contribution to most natural experimental evaluations, with evaluations benefitting from a mixed-method design incorporating qualitative methods throughout. Key components of an evaluation in which qualitative methods should be integrated include describing the intervention, the system or context, developing a theory of change, informing the selection of populations and controls, characterising and selecting outcomes and indicators, generating data on outcomes, understanding mechanisms and mediators, explaining change and understanding stakeholders' perspectives. The evaluation should be planned to ensure that the qualitative components are incorporated into the evaluation at the appropriate stages of the evaluation to achieve maximum use of the qualitative data gathered.

Reporting, critical appraisal and evidence synthesis of natural experimental evaluations

Clear reporting of the natural experiment and the evaluation is crucial for the best use and understanding of these studies. Critical appraisal may be required to understand the rigour of an individual study or undertaken as part of a synthesis of evidence. There is no single tool that can fully assess the risk of bias of all natural experimental evaluation study designs. Therefore, the tools available should be considered in terms of their strengths and limitations for the purpose of a given review. Synthesising evidence from natural experimental evaluations requires consideration of how to manage the probable diversity in study design and characteristics. For some review topics it may be more valuable to examine effectiveness in a broader sense rather than an estimated effect size.

Data infrastructure and information governance

Natural experimental evaluations often use data that were 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. A potential solution to this is to establish secure research platforms, trusted research environments, designed to maintain security and confidentiality of the data and provide efficient access to researchers.

Recommendations

Good practice considerations

Good practice considerations are provided for different users of the framework, derived from the content of the updated framework. In condensed form, these recommendations are provided below.

All producers and users of natural experimental evaluations should:

- Understand the design and planning processes of an evaluation of a natural experiment, including how to identify opportunities for natural experimental evaluation, select the most appropriate evaluation approach and assess the feasibility of the evaluation.
- Consider the variety and importance of stakeholders.
- Recognise the respective strengths of quantitative, qualitative and integrated analytical approaches, incorporating perspectives from diverse disciplines, such as economics, social sciences, epidemiology, for investigating the impacts of natural experiments.

Researchers conducting natural experimental evaluations should:

- 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.

- Consider adopting a systems thinking approach to evaluating natural experiments.
- Consider using a combination of methods, including alternative methods of effect estimation, robustness checking and a mixture of qualitative and quantitative methods.
- Adopt open science practices, publishing a protocol or plan of the study in advance in open access journals or repositories.
- Clearly report the natural experiment event and all stages of the evaluation, including its planning, protocol, analyses and results, using established reporting standards where available ensuring key details are in plain language appropriate for the evidence users.
- Include a health equity perspective in the evaluation.
- Be aware that evaluation of the strength of evidence from natural experimental evaluations should be based on detailed appraisal of the strengths and limitations of the study methods, not on broad study labels.

Research funders and commissioners supporting and investing in natural experimental evaluations should:

- Encourage best practice when commissioning or funding natural experimental evaluations, requiring that a protocol or methods appropriate advance study plan is available prior to analysis commencing, findings are published in open access journals and the relevant reporting guidelines are followed.
- Establish processes within funding bodies to facilitate flexible and timely responses to prospective natural experimental evaluation opportunities.
- Support capacity building for natural experiments through investment in infrastructure and the workforce.
- Negotiate with data owners to make routinely collected data both available and linkable to other datasets for evaluations of policies and programmes.
- When commissioning natural experimental evaluations, be prepared to be flexible and pragmatic and accept that both the evaluability of the natural experiment and the feasibility of the evaluation require assessment, which may result in the full evaluation not being viable. Flexibility may also be required when considering the start date and timescale of the research, as policy interventions can be delayed, changed or withdrawn, the effects of each will require consideration in an evaluation.

Journal editors, policy-makers, practitioners and other decision-makers publishing and using evidence from natural experimental evaluations should:

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

Conclusion

The new framework has been developed in consultation with international experts in natural experimental evaluations. It aims to promote the conduct and application of methodologically robust studies where a policy or programme is amenable to evaluation as it is or has been implemented. The researchers hope it will raise awareness of the whole range of issues that need to be considered when planning an evaluation or using the results to influence policy. With the large number of topics covered by the framework, the study aimed to convey the key points with signposting to more detailed information provided throughout.

Funding

This project was jointly funded by the Medical Research Council (MRC) and National Institute for Health and Care Research (NIHR), with project reference MC_PC_21009. The work is published in full in *Public Health Research*; Vol. 13, No. 3.

Chapter 1 Introduction

Summary

- The value of natural experimental evaluations for assessing the impacts of interventions on population health is more widely recognised now than it was a decade ago when the UK Medical Research Council (MRC) guidance was first published. The use of natural experimental evaluations in population health has since proliferated.
- To make the most of the opportunities for natural experimental evaluations of interventions in population health, health services and health systems, issues of data infrastructure, information governance and study planning need to be resolved. Key definitions and concepts need clarification, and agreement is required on the circumstances in which natural experimental evaluations can provide trustworthy and useful evidence for decision-making.
- This updated and extended guidance discusses a wide range of research methods that can be applied to natural experiments and places them in the context of good practice in the design, planning, conduct and reporting of natural experimental evaluations.

Natural experiments, defined in the UK MRC's original guidance (2012) as events outside the control of researchers that divide populations into exposed and unexposed groups,[1] have long been utilised as opportunities to evaluate population health and health system interventions. For the last 50 years natural experimental evaluations have been overshadowed by the randomised controlled trial (RCT) as the method of choice for evaluating interventions to improve health, but in the decade or so since the MRC first published guidance on natural experiments, the balance has shifted. One possible reason for the greater appreciation of the scope and value of the opportunities, provided by natural experiments, is their successful use in other fields such as economics.[2] Another is the availability of a wider range of methods, again mostly developed by economists or by other social or political scientists working outside the health field, but then applied to good effect within it. A third possible reason is better availability of large administrative and other 'big data' datasets that link information on exposure to public policies with health and other outcomes, creating more opportunities to apply the methods. A fourth possible influence is growing demand by policy-makers for timely evidence about 'what works' to address public health problems that are too pervasive and complex to be tackled solely by improved medical treatments or other individual-level interventions, and where a RCT would be impractical or unethical, or would not provide the most useful evidence to guide decision-making. Such pervasive and complex problems result in calls for action to address the social determinants of health. Evaluations of natural experiments are often the most appropriate approach for evaluating these types of upstream interventions.[3–5]

These developments make it timely to revisit and renew the guidance. In this introduction we set out the rationale for doing so, explain the structure and coverage of the new framework and describe the methods we have used to update and extend the previous version.

Rationale

Over the past decade, there have been several attempts to summarise alternatives to RCTs for evaluating the impact of interventions on population health.[6–8] The focus of such overviews is primarily on quantitative methods for effect estimation. Other aspects of study design and conduct, such as how to identify a good opportunity for a natural experimental evaluation, or how to place effect estimates into a broader model of whether, how and in what circumstances an intervention achieves its effects, have largely been neglected. Some of these papers apply a hierarchy of study designs to natural experiments that unhelpfully restricts the range of methods that should be used to evaluate them.[9] There remains substantial uncertainty about key definitions and concepts, about how widely or narrowly natural experiments should be defined, and about the range of methods that should be considered as part of the methodological toolkit for evaluation. We take the view that, while unbiased estimates of effectiveness are an important goal of evaluations, these are not the only goal and may be unachievable in some circumstances.[10] If research

seeks to produce evidence that is useful for decision-making, a range of methods is likely to be required, including those that work in situations when a planned experiment, such as a RCT, would not be feasible or ethical.

In this new framework, the researchers therefore review a range of methods that can contribute to a natural experimental evaluation and situate them in the context of the diverse circumstances in which a natural experimental approach might be useful. The researchers intend the framework to provide a single, integrated, up-to-date guide to the use of a natural experimental approach to evaluate population health interventions and to:

- raise awareness among researchers of the range of approaches available for evaluating interventions as natural experiments and provide signposts to the more detailed methodological literature;
- provide clearly targeted messages for intervention stakeholders (e.g. in local or central government) to help them decide whether a natural experimental evaluation would be useful, and if so of what kind; and
- provide information to help journal editors, funders and peer-reviewers to understand the strengths and weaknesses of funding proposals for, and articles reporting, natural experimental evaluations.

Natural experimental evaluations have already proved their value across a wide and disparate range of health and non-health policy areas, providing otherwise unobtainable evidence about the effects on population health of clean air legislation, new transport infrastructure, food policy interventions, suicide prevention, tobacco and gun control, trade agreements, nonpharmaceutical pandemic control measures, and many other kinds of policies and programmes. The researchers believe they can contribute much to these and other areas. It is the researchers hope that this new framework, on how to identify opportunities for natural experimental evaluation, how to design, conduct, report and synthesise the evidence from such studies, and on what kinds of research infrastructure and governance processes are needed, will help to realise this potential.

Chapter 2 Methods for developing the framework

Summary

- Development of the framework involved convening a writing group and an advisory group, and holding expert workshops and an online consultation to gather expert opinions on the content of the framework.
- The advisory group included stakeholders with experience of using natural experimental evidence and with methodological expertise to provide input at key stages of the project.
- Three workshops were held online to enable participants from Europe, Africa, the Americas and Australasia to provide feedback on an initial draft of the framework.
- An online consultation invited participants to review each chapter of the revised framework. Invitations were distributed worldwide and received input from researchers, members of research funding boards, policy-makers, journal editors and representatives of national and local governments.
- Feedback from the workshops and consultation was used by the writing group to finalise the framework, with drafts of the framework provided to the advisory and oversight groups.

The researchers formed a project writing group, reporting to a funders' oversight group, and convened a stakeholder advisory group. After developing a first draft of the framework, the researchers undertook a consultation process using online workshops and an online consultation to seek recommendations on what content to include in the updated guidance for conducting and using natural experimental evaluations.

The project writing group comprised population health researchers representing a range of disciplines and methods, including epidemiology, statistics, health economics, public health, qualitative and evidence synthesis research. Sections of the framework were developed by subgroups of the writing group, with drafts shared with the full group for comment and additional input. Monthly meetings were held throughout the project to assess timeline appraisal and study goals and objectives. This included regular communication with Social and Public Health Sciences Unit staff who assisted with the practicalities and logistics of the online consultation questionnaire data collection.

An advisory group was convened to represent key stakeholders in natural experimental evaluations, including those with experience of using natural experimental evidence in addition to members with methodological expertise. The group met quarterly over the course of the project, with the responsibility to provide expert input and discuss progress against key milestones of the project. An oversight group was convened by the funders, comprising representatives of the MRC and NIHR, including the MRC-NIHR Methodology Advisory Group, and the UKRI MRC Population Health, Population Health Sciences, the Public Health Intervention Development funding boards and the NIHR Public Health Research Programme. This group met every 2 months with members of the project team to provide oversight of the content and progress of the project. Members of the advisory and oversight groups are listed in *Appendix 1*.

Parts of this text have been reproduced from Craig *et al.*[11] This is an Open Access article distributed in accordance with the terms of the Creative Commons Attribution (CC BY 4.0) license, which permits others to distribute, remix, adapt and build upon this work, for commercial use, provided the original work is properly cited. See: https://creativecommons.org/licenses/by/4.0/. The text below includes minor additions and formatting changes to the original text.

Expert workshops

The writing group developed a first draft of the framework, creating an outline of proposed sections for the framework and key issues for those sections to include. Online expert workshops were held to obtain expert opinions from international researchers, policy-makers, funders and journal editors, on the first draft of the framework, a key decision point in the development of the framework. The objectives were to identify additional content, unnecessary content, any topics of contention, and any required revisions to the structure of the framework.

There were three workshops, each with the same content, to enable participants from different time zones (North and South America, Europe and Africa and Australasia) to partake at a suitable time. In advance of each workshop, participants were provided with a briefing document summarising the project objective, questions for the workshop sessions based on the above objective, plus specific issues for the proposed framework chapters and a copy of an article by the project team highlighting issues relating to natural experimental evaluations.[12] The questions used to prompt discussion are presented in *Table 1*.

The briefing document and the workshop format were split into three sessions:

- Session A examined coverage of the framework, and concepts and definitions used in the framework guidance.
- Session B focused on processes relating to natural experimental evaluations: design and planning, pre registration of evaluations, infrastructure and information governance for the use of administrative data, and engaging policy-makers and other stakeholders.
- Session C focused on methods: quantitative, qualitative, mixed methods, economic evaluations, reporting and evidence synthesis.

TABLE 1 Workshop questions

Session A: Coverage, concepts and definitions
Coverage
1. Does the proposed list of sections provide a comprehensive and coherent structure for the guidance?
2. Are there any key issues that do not fit within the proposed structure?
Concepts and definitions
1. Is this broad definition of natural experimental evaluations appropriate? If not, how can it usefully be made more specific?
2. Would it be useful to include in the guidance a glossary along the lines of *Table 3*? If so, are there any key terms missing? Are brief summaries of the kind illustrated likely to be useful. Would links to references for more formal definitions be helpful?
Session B: Process
Design and planning
1. Does the framework provide a useful structure for planning and conducting a natural experimental evaluation? Are there other aspects of design and planning natural experimental evaluations that should be addressed in this guidance?
2. Have we identified the most important sets of circumstances that provide opportunities for natural experimental evaluations? Are there others we should add to the list?
Pre registration of natural experimental evaluations
1. How can trial registration systems, designed for prospective studies, be adapted to the requirements of retrospective studies?
2. How can the benefits of registration be aligned with need for flexibility in study protocols?
3. What additional safeguards, if any, are needed to ensure transparency in the conduct and reporting of natural experimental evaluations?
4. What can funders and journal editors do to support transparency in the conduct and reporting of natural experimental evaluations?
Infrastructure and information governance for the use of administrative data
1. What are the most important barriers to the efficient use of routinely collected data for natural experimental evaluations?
2. How can these barriers be removed or minimised, while still preserving data security?
3. Are there examples of good practice (e.g. effective infrastructures for making data available, or streamlined approaches to information governance) that we should highlight in the guidance?

TABLE 1 Workshop questions (*continued*)

Engaging policy-makers and other stakeholders
1. Are there good examples of guidance on the use of natural experimental evaluations to inform policy-making from countries other than the UK?
2. What other approaches might help to improve the understanding and use of natural experimental evaluation evidence by decision-makers, especially those outside central government?
Session C: Methods
Quantitative methods
1. Are any important methods missing from the list above? Have we included any that do not belong (e.g. because they could not be used to identify a causal effect in any circumstances)?
2. Is there a hierarchy among the methods, or should they be seen as a toolkit?
3. Is a target trial framework a useful way of evaluating natural experimental evaluations study designs? Are there more straightforward alternatives?
Economic evaluation of natural experimental evaluations
1. The MRC/NIHR framework for developing and evaluating complex interventions recommends CCA and CBA as the most appropriate forms of economic evaluation, and the adoption of a broad (e.g. societal) rather than a narrow (e.g. provider) perspective for identifying costs and outcomes. Do these recommendations also apply to natural experimental evaluations?
2. Does the incorporation of economic evaluation into a natural experimental evaluation pose specific challenges over and above those associated with any evaluation of a complex intervention. Are there good examples of how these have been addressed?
Qualitative methods
1. The guidance currently focuses on practical issues and not on methodological theory around qualitative natural experimental evaluation design. Do you agree with this emphasis?
2. The section is aiming to be 'theory neutral' and does not compare different theoretical frameworks or designs, e.g. realist designs, normalisation process theory, etc. Do you agree?
3. Can a natural experimental evaluation rely entirely on qualitative methods, for example to understand change, mechanisms and mediators, or the perspectives and practices of intervention stakeholders, or should such methods always be used in conjunction with quantitative methods of effect estimation?
Integrated, mixed- and multimethod evaluations
1. Practicalities aside, are any of the methods listed clearly preferable to the others?
2. Taking practicalities into account, which method or methods are most appropriate to natural experimental evaluations?
Reporting
1. Are existing reporting guidelines adequate for reporting natural experimental evaluations, or is there value in extending them to cover issues specific to natural experimental evaluations?
Critical appraisal and evidence synthesis
1. What balance would users find most useful between conceptual and practical issues in guidance for appraising and synthesising natural experimental evaluation evidence?
2. To what extent are the challenges for critical appraisal and evidence synthesis of natural experimental evaluations particular to NEEs?
3. Is it best to use ROBINS-I and GRADE to assess risk of bias and certainty of evidence, despite the challenges involved, or are there other approaches that are sufficiently rigorous but more straightforward?

Respondents were recruited for the workshop through co-investigator and advisory group networks and individuals identified as the framework was developed. Authors of academic articles, reporting natural experimental evaluations in population health, were identified as the framework content was developed and invited to participate in the workshop. Relevant stakeholders in the research community were also recruited, representing funders from international funding boards, journal editors and peer reviewers of journal reporting on natural experimental evaluations. Knowledge within the project team and a project advisory group was used to add to the list of people with professional expertise of

natural experimental evaluations of population health interventions. A total of 72 individuals were invited to attend, 23 accepted the invitation and 21 participants attended. Each workshop lasted approximately 3 hours.

Results: key points emerging from the expert workshops

Concepts and definitions

The definition of 'natural experiment' needed to be revised. While there was some support for a narrow definition of a natural experiment that focuses on assignment being 'as-if randomised', others in the workshop supported the idea that this would rule out many studies that this framework is intended to cover. Therefore, the researchers aimed for a compromise, explaining that there is a spectrum of studies, keeping the definition broad to make the guidance as useful as possible while acknowledging the strength of a natural experimental evaluation involving 'as-if randomised' allocation. It was noted that the perspective of the definition is from the starting point of having an event or policy that is to be evaluated and exploring how that can be done.

Design and planning

There was support for adapting the framework for evaluating complex interventions for planning and conducting natural experimental studies. Use of the framework for complex interventions entailed some modifications to be made as some of the content (e.g. that parts relating to the development, refinement and implementation of the intervention) was not relevant to this natural experimental evaluations framework.

Quantitative methods

The researchers' original list of methods required revision to clarify between study designs, analytic methods, and additional design features. The workshop feedback suggested that the current list mixed study designs and methods. It was suggested that there be three sections, study designs, analytical methods and additional testing to establish causal inference.

Economic evaluation

There was need to clarify issues that are specific to natural experimental evaluations, for example uncertainty of effect estimates, difficulty of obtaining retrospective cost data, rather than those that are common to all complex interventions. One key challenge is that of obtaining estimates of costs. With evaluations involving natural experiment interventions there are more assumptions that must be made as in many cases the researcher cannot design their own data collection.

Qualitative methods

Further information was added on the consideration of the role of qualitative methods in evaluating natural experimental evaluations. There was general acceptance that qualitative methods are a useful adjunct to quantitative methods of effect estimation, but a strong view from some participants that studies relying exclusively on qualitative methods were beyond the scope of the framework.

Evidence synthesis

There was discussion of the extent to which all the issues can be covered. There was clarification that the framework will focus on signposting both the issues that arise when appraising or synthesising evaluations of natural experiments and signpost tools and guidance that will help.

Online consultation

Following revisions to the framework, in response to feedback from the workshops, an online consultation was developed to consolidate the content of the framework. A draft of the framework was provided in which each section had a bullet point introduction of key messages, giving the option to continue reading the full text of the section, answer questions and provide comments on the section, or move forward to the next section. Questions asked of the participants are provided in *Appendix 1*, *Table 8*. The respondents were invited to provide their details to be included in

acknowledgements (see *Appendix 1*, *Table 9*), or alternatively to complete the consultation anonymously. All responses were de-identified and no statements were linked to particular respondents.

The consultation was open for 7 weeks, from 21 September 2022 until 8 November 2022. An initial list of 95 individuals were invited to partake, including individuals involved in commissioning, conducting, and using evidence from natural experimental evaluations, that is researchers, local and national government representatives and practitioners, policy-makers, journal editors and funder representatives. Invitees were identified from among those invited to the previous workshops, recommendations from the project advisory group, and 'snowball' invitation suggestions from participants. Members of the NIHR School for Public Health Research (SPHR) Network for the use of Natural Experiments in Public Health were also invited. By the closing date 200 people had been invited to participate; the intention was to share the consultation widely and a high response rate was not anticipated.

Participants were invited by e-mail, outlining the purpose of the consultation and inviting them to reply indicating whether they would partake. When the participant accepted, their e-mail address was added to an access list for the consultation and they were provided a link to the online consultation questionnaire. The introductory page of the questionnaire provided a clear plain language statement explaining the consultation and a PDF document of full participant information details relating to the consultation.

Results from the online consultation

There were 44 completed responses. Participants provided details of their roles; many provided information about multiple roles. Most respondents were researchers (intervention *n* = 18, quantitative *n* = 39, qualitative *n* = 17), and several were members of funding boards (*n* = 6), representatives of a funding body (*n* = 1), journal editors (*n* = 8), policy-makers (*n* = 1), practitioners (*n* = 2) or clinicians (*n* = 5). Most respondents reported a university as their institution (*n* = 41) although some participants had dual or multiple roles, and some respondents reported their organisations as public sector (*n* = 5), not for profit (*n* = 2) or a for-profit organisation (*n* = 2). Summaries of feedback from the online consultation are outlined in *Table 2*.

TABLE 2 Summaries of online consultation feedback for each framework section

Framework section
Concepts and definitions
• Explanation of a natural experiment: a natural experiment includes an instance where exposure was changed as the result of some exogenous event, something that we can harness. Suggestion current first sentence of the explanation does not differentiate enough from observational data.
• Clarifications of the definitions of natural experimental evaluations and natural experimental studies, state explicitly if the terms are used interchangeably. Request to clarify whether the terms 'natural experiment' and 'natural experimental evaluation' are different.
• Queries on meaning of 'intervention' as used in this framework, whether this covers a natural experiment.
• Several suggestions for additional definitions including the concept of contamination of comparison groups and scale.
• Clarifications for theory of change, whether this includes the role of contextual circumstances, how does theory of change differ from programme theory, and whether a theory of change could theorise how and why a natural experiment (rather than an intervention) impacts on outcomes of interest.
• Request for more explanation of system boundary.
• Suggestion that the definition of positivity should be strengthened by amending 'possible' to 'all key', 'there are observations in all key…'.
• Suggestion that the definition of triangulation could be altered to reflect that this can exist across investigators triangulation, theory, data source, in addition to methods triangulation.

continued

TABLE 2 Summaries of online consultation feedback for each framework section (*continued*)

Framework section
Design and planning
• Stakeholders: include the importance of community stakeholders.
• Evaluability: suggestion to reduce this content.
• Suggestion that a key stage missing that is 'characterising the intervention'.
• Feasibility: suggestion to emphasise an audit to assess the basic feasibility of the intervention by assessing things like exposure, access and completion.
• Equity: several comments to expand content on equity. If economic evaluation is included also include information about impact on equity and on climate.
• Requests to clarify the role of qualitative methods in the framework.
• Suggestion to include information about integrating research methods.
• Subsection on discovering and theorising natural experiments has text in a box that should be in the main text.
Quantitative methods
• Query the accuracy of 'always be a certain degree of self-selection into the exposed or unexposed group'. A natural experiment where units of people are assigned to different policies may have less contamination than a randomised trial.
• Query of explanation of exposure 'we have treated exposure to the intervention as being uniform across the exposure group …' People change their exposure levels all the time in repeated measures studies 'naturally', and we harness this as researchers. A natural experiment though is where that variation in exposure is (to varying extents) exogenous.
• Suggestion to include directed acyclic graphs to inform model inputs.
• Query of where systems and systems thinking come into this section.
• Queries about the labelling and apparent hierarchy of study methods.
Economic evaluation
• Several suggestions to expand content on equity.
• Clarify meaning of 'dynamic microsimulation models'.
• Recommendation that this guidance needs to reach commissioners of evaluations who often request an economic evaluation when it is near-impossible due to study design or lack of data.
• Suggestion to highlight that the ideal outcomes for a natural experimental evaluation may be quite different from the outcomes or effects needed for a full cost-effectiveness evaluation.
Qualitative methods
• Queries on why the framework is recommending qualitative methods.
• Suggestion to include process tracing as a useful method. Suggestion to mention system mapping techniques being used to explore and describe the system in which an intervention operates.
• Suggestion to add policy process analysis as another potential way that qualitative analyses can support natural experimental evaluations.
• Suggestion to add that qualitative methods are important for understanding assignment mechanisms and how the implementation of an intervention or policy can affect treatment exposure in unplanned ways.
Critical appraisal and evidence synthesis
• Provide more guidance on searching, involve an information specialist and consult with experts in the field.
• Emphasise that all studies, including randomised controlled trials and non-randomised studies, are at risk of bias.
• Clarify what is meant by 'depending on how tightly natural experimental evaluations are being defined'.
• Improve the structure of the synthesising results section.

TABLE 2 Summaries of online consultation feedback for each framework section (*continued*)

Framework section
Infrastructure and information governance
• Suggestion for additional text expanding on the population basis of data linkage. Some trusted research environments link data between two or more datasets while others use a more population-based spine. It is important that the limitations of the population base are reported for trusted research environments. • Recognise that some types of existing data can be expensive, e.g. sales and purchasing data, datasets developed for commercial purposes. • Examples of data linkage from Denmark, which includes school outcomes, personal medical trajectories, personal income information, UK Open SAFELY, Brazil CIDACS, Australia The Provincial Health Data Centre of the Western Cape Province, the Multi Agency Data Integration Project (MADIP) and the (in development) National Disability Data Asset (NDDA), Canada – The Data Liberation Initiative (DLI), New Zealand SNZ IDI.
Reporting and good practice considerations
• Reporting resources are weighted towards quantitative methods. Suggestions for considerations for all framework users: • Emphasise the role of multiple actors in the co-creation of a rigorous evaluation. • Encourage more co-creation in policy and or practice partnerships. • Raise awareness of the unpredictability in researching natural experiments which means not all scoped opportunities will result in evaluations. Therefore, it is good practice to stop when no meaningful evaluation is possible; research environments accommodate the uncertainty of these studies and facilitate publication of altered or discontinued evaluations. • Research funding bodies to increase flexibility and speed to respond to prospective natural experimental evaluation opportunities. • Natural experimental evaluations are likely to have some methodological imperfections and limitations that may be more obvious than those found in a well-designed randomised controlled trial. However, funding research into many important policies and interventions which are not amenable to trials requires a willingness to tolerate these imperfections. Suggestions for considerations for researchers: • Stress the importance of flexibility. • Suggest there is a duty on the part of researchers to educate the consumers of their research on the nuances and assumptions that underpin the analysis. This includes: • In many studies there can be vagueness about 'the effect'. Ensure the estimand, the treatment effect, is clearly defined, explaining the quantity being measured and the relevant subpopulation. • There is often a lack of a clear definition of the counterfactual. • There is a need for clear language, lay language descriptions of assumptions that are made during analysis. Suggestion for commissioners of natural experimental evaluations: • Avoid being over-prescriptive in the type of evaluation commissioned, to allow the researchers to make the best use of the data available, e.g. to avoid specifying an economic evaluation unless certain the data are available.

IDI, Integrated Data Infrastructure.

In addition to revisions made by the writing group in response to the feedback summarised above, decisions made by the writing group after discussion and further feedback from the advisory group, include those listed below.

Terminology

Clarification was required in *Concepts and definitions* that the framework focuses on natural experimental evaluations; however, the terms 'natural experimental evaluations' and 'natural experimental studies' were on occasion used interchangeably. Clarification was also required in the *Introduction* and *Concepts and definitions* section that 'intervention' is used in a broad sense to include policies, etc. Sometimes 'intervention or event' is used. It was recommended that abbreviations be kept to a minimum.

There was recommendation for an introduction to provide background to the framework, including a statement of the objectives, differences from original guidance, very brief methods and a summary paragraph of the sections.

Equity

There were comments from the consultation to expand content on equity, if economic evaluation is included then there should be inclusion of information about impact on equity and on climate. The researchers will ensure they address equity in the relevant sections and contexts, particularly in *Design and planning*, emphasising the importance of considering impact on equity but not mandatory. Considering impact on climate will also be mentioned in *Design and planning*, although not all natural experimental evaluations will have a direct climate impact.

Stakeholders

Feedback relating to stakeholders resulted in the following updates to the framework. Emphasis on the importance of working with various stakeholders and that it is clear this term covers a broad scope of individuals/organisations at different stages of the natural experiment and the evaluation. Acknowledgement of the possible complexities of stakeholder relationships, where there may be conflicts of interests or political sensitivities in relation to the evaluation. In the *Design and planning* section, recommendation to establish clear terms of engagement with stakeholders and note that having a published protocol for the evaluation will provide transparency and information of the independence of the evaluation from the original natural experiment.

Mixed/integrated methods

Feedback recommended that this information should not get lost when moved from being a standalone section to *Design and planning*. This included checking that the information was adequately referred to in *Introduction*, *Concepts*, *Design and planning*, *Quantitative methods*, and *Qualitative methods* sections, with signposting sentences signalling the importance of mixed methods and appropriate cross referencing to the main content in the *Design and planning* section. An example of process tracing, combining both quantitative and qualitative research, was added as an example in a box.

Glossary

We checked for additional key terms from other sections, for example from the *Economic evaluation* section.

Design and planning

As this section covered too many topics, it was appropriate to split this into two parts. More information was included about stakeholders, the importance of stakeholders and collaborative work, the wide range of individuals or organisations the term ‘stakeholder’ may cover, that different stakeholder groups will be relevant at different stages of the natural experimental intervention and the evaluation, and issues of conflict of interest for the evaluation.

Quantitative methods

There was a suggestion to restructure the study design table to have two parts: analytical methods (simple before/after, interrupted time series (ITS) matching, difference-in-difference, regression discontinuity, and synthetic control) and data structures: cross-sectional, repeated cross-sectional, panel data, pseudo-panel data and different units of analysis (individual or aggregated). After discussion, it was decided that the distinctions between analytical methods and study designs becomes complicated and less clear and therefore the preamble, before the descriptions of the quantitative methods, will be adjusted to explain the purpose of the descriptions provided. It was decided to add a more detailed introduction to the table describing the methods, explaining the structure and why this includes both ways of gathering data and analytical methods.

Economic evaluation

More information was added to highlight that ideal outcomes for a natural experimental evaluation may be quite different from the outcomes needed for a full cost-effectiveness analysis (CEA). Further suggested changes were to discuss that the costs and benefits may be borne by different jurisdictions, thus not directly ‘cost savings’ and there was inclusion of further information about the use of microsimulation models.

Infrastructure and information governance

A box describing Centre for Data and Knowledge Integration for Health (CIDACS), as an example from a LMIC, was included. Another example suggested was the potential of the Malawi Epidemiology and Intervention Research Unit (MEIRU).

Finalising the framework

De-identified feedback from the online consultation was summarised and collated to gather advice on the content of the framework. The writing team used the feedback from the consultation to revise the guidance accordingly. Throughout the project stages, drafts of the framework were provided to the advisory group and the oversight group for feedback.

The good practice considerations were developed by the writing team as summaries of key messages contained in the framework for different users of the framework. The good practice considerations were included in the online consultation to give participants the opportunity to comment and make further suggestions.

Limitations

The framework provides an overview and signposts to available information, rather than covering all topics in detail. A limitation of the framework is that methods are constantly being developed and enhanced. This study raises awareness of continually improving methods; however, in future, there will be further developments requiring updates of the framework. There are some issues relating to conducting or using evidence from natural experimental evaluations that remain to be resolved, for example how best to conduct economic evaluations, or how to assess the risk of bias of some types of natural experimental evaluations. The researchers acknowledge that no participants in the project were recruited specifically for their experience as patients or as having been exposed to particular interventions. However, the aim of the study was to develop methodological guidance with input required from methods experts, and individuals with experience of developing, commissioning, and using evidence from natural experimental evaluations. Given the nature of interventions studied in natural experimental evaluations, most participants also had lived experience of such interventions, although we recognise that professional participants are not representative of all public groups. The researchers also acknowledge that despite getting wide input from across geographical regions, representation from participants with expertise in low-income country settings was limited.

Research recommendations

Recommendations resulting from the development of the framework were incorporated into the framework, with *Good practice considerations* providing good practice considerations when planning, commissioning, conducting, reporting, and using evidence from natural experimental evaluations. These recommendations are designed for specific audiences. In addition to recommendations for all producers or users of natural experimental evidence, there are specific considerations for those conducting, supporting and investing in, and publishing and using evidence from natural experimental evaluations. As noted in the limitations paragraph above, methods will continue to be developed that will be useful to apply in relation to this framework. There is a need for further development of methods to conduct economic evaluations, assess risk of bias, and synthesise natural experimental evaluations.

Chapter 3 Framework for using natural experiments to evaluate population health interventions

Structure of the framework

Concepts and definitions discusses how broadly or narrowly natural experiments should be defined[13] and whether it is useful to distinguish sharply between natural experimental evaluations and other observational studies that attempt to identify causal relationships using changes or variation in exposure. The researchers argue for a broad definition and explain why attempts to narrow the definition to include only studies that use one of a prescribed range of methods, such as those that can address unobserved confounding[9] or that satisfy some other criterion such as 'as-if randomisation',[14] are unsatisfactory. Given the diverse disciplinary origins of the methods used to evaluate natural experiments, the researchers also provide a glossary of key terms.

Design and planning: concepts suggests a framework for planning natural experimental evaluations and discusses the sets of circumstances that are likely to generate useful opportunities for studies of this kind. A number of such situations recur in the literature and provide useful pointers for the design of future studies. This section also reviews the use of mixed and integrated methods. It discusses assessing the evaluability of the natural experiment and the feasibility of the evaluation design, how best to involve stakeholders and highlights the importance of study registration and independent oversight.

Researchers planning a natural experimental evaluation can now draw on an extensive methodological toolkit. Whereas previous guidance and overviews have tended to focus on quantitative methods for effect estimation, we argue for a wider scope. Including qualitative methods can provide vital information about threats to validity, possible causal mechanisms and how the assignment process for an intervention actually operates in practice. *Qualitative methods* and *Economic evaluation* review the main quantitative methods available to researchers and issues arising in the use of economic evaluation methods within natural experimental evaluations, and *Qualitative methods* provides an overview of where qualitative methods contribute.

Reporting, critical appraisal and evidence synthesis signposts readers to relevant reporting guidance. The researchers review issues arising in the synthesis of evidence from natural experimental evaluations, which remains challenging[15,16] despite recent improvements in the tools available for assessing risk of bias[17] and combining evidence from different study designs.[18,19]

Many natural experimental evaluations are conducted retrospectively, so good-quality routinely collected data – either from administrative systems or other sources of 'real-world data', population surveys or combinations of such sources – is critical, as are streamlined and proportionate systems of information governance. As researchers are typically not in control of these data-gathering processes, a good understanding of the strengths and weaknesses of such sources is also vital, and in *Infrastructure and information governance*, the study reviews recent improvements in data infrastructure.

Finally, in *Good practice considerations*, the study presents good practice considerations synthesised from the preceding chapters and picks out issues particularly pertinent to researchers, publishers and key users of natural experimental evidence, such as policy-makers.

Concepts and definitions

Summary

- The framework defines natural experiments as events or processes outside the control of a researcher that divide a population into groups with differing degrees of exposure.
- 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 evaluations 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 research 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; therefore, the framework provides a glossary defining key terms as they are used in this guidance.

The value of natural experiments is that they widen the range of interventions and outcomes that can be evaluated beyond those that are amenable to experimental manipulation. These experiments provide opportunities to evaluate outcomes that are by-products of, rather than the main purpose of, the intervention and which would not provide a persuasive or ethically sound basis for experimental manipulation.[20] 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 or financial commitments can make experimentation unattractive to decision-makers. A widely used example of a natural experiment, relevant to population health, might be the raising of the minimum school leaving age.[21] As the motivation is to raise educational standards rather than to improve health, conducting a trial to identify health impacts, which may take many years to emerge, is unlikely to be acceptable, either to policy-makers or to the young people who would be affected. Natural experimental evaluations have been conducted comparing the health in adulthood of cohorts who reach the minimum school leaving age before and after the change is made, using data that is routinely collected on health outcomes, for example via surveys or vital events recording.[22–25]

For this updated framework, as in the first version of the MRC guidance on natural experiments,[1] the researchers adopted a broad definition of natural experiments as events, processes outside the control of a researcher that divide a population into exposed or unexposed subpopulations, or into groups with differing degrees of exposure. A natural experimental *evaluation* uses data emerging from the introduction, delivery or (less commonly) 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 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,[9] or satisfy some other criterion such as 'as-if randomisation'.[14] The researchers 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 assessed by comparing exposed and unexposed groups in terms of observed confounders, whereas the key question 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 one spectrum along which natural experimental studies lie. It captures one

important characteristic of such studies rather than providing a criterion that distinguishes them from other kinds of observational study. We recommend the use of a target trial approach (see definition in *Table 3*) for a full assessment of risk of bias in a natural experimental study, de Vocht *et al.*[13] 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.

TABLE 3 Glossary

Term	Usage
As-if randomisation	A process that determines exposure to an intervention that is effectively random, such that the exposed and unexposed groups that result are balanced in terms of both observed and unobserved confounders, even if it does not involve the kind of formal randomisation process that would be used in a randomised controlled trial.
Assignment, allocation	The process that determines whether units (which may be individual people, groups of people, geographical areas, or other aggregates) are exposed to the intervention that is being studied. Also referred to as the data generating process.
Causal estimand	See treatment effect.
Concurrent interventions (or co-interventions)	A common source of confounding in natural experimental evaluations is when one or more interventions, other than the intervention being studied, are implemented at around the same time. Depending on the extent of overlap in exposure, it may be possible to disentangle the effects – e.g. by focusing on outcomes specific to the intervention of interest, or on populations exposed to one intervention but not the other(s).
Confounding, confounder (observed, unobserved)	Confounding refers to the mixing of the effects of the intervention (or exposure of interest) with the effects of characteristics associated with the intervention. 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 also address unobserved confounders.
Control, comparator, counterfactual	Controls (often also referred to as comparators) 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 (see *Quantitative methods*), the pre-intervention trend can be extrapolated beyond the point at which the intervention occurs in order to estimate the counterfactual.
Covariate	Variables that measure characteristics of exposed and unexposed units. Characteristics that are either direct or indirect causes of both exposure and outcomes may cause confounding.
Exposure	A factor that may explain or predict an outcome in an observational study; in the context of a natural experimental evaluation, this term refers to receipt of an intervention and is effectively synonymous with treatment. 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 an 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 estimate 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 that cause 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, while they are in place (using knowledge of an assignment process that determines exposure at an individual unit level within a population) or when they are withdrawn.
Natural experiment	An event, process, or intervention outside the control of a researcher that can be used to divide a population into exposed and unexposed subpopulations, or into subpopulations with differing levels of exposure. The division may be spatial, temporal or based on the characteristics of individual units, such as age or income.
Natural experimental evaluation	A natural experimental evaluation uses the differences in exposure generated by an event or process to identify, measure or understand the effects of the intervention.
Negative control outcomes, 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.

TABLE 3 Glossary (*continued*)

Term	Usage
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 inferred by measuring outcomes for a group that is not exposed but is intended to be 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'.
Population health intervention	Policies, programmes, services, organisational changes, infrastructure investments and other activities that affect population health directly or indirectly, including those primarily intended to influence some other outcome such as living standards, educational attainment, travel behaviour and other determinants of health. The definition includes changes in the organisation and delivery of health services, and interventions originating in other sectors of social and economic policy.
Process tracing	Process tracing is used to infer the best explanation of an outcome within a case. It uses a structured framework to examine hypothesised causal mechanisms by applying a series of empirical tests, collecting and assessing evidence including the sequence of events, to confirm or disconfirm a hypothesis about how an outcome occurred.
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 conducted by researchers. 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, as in the case of vaccination) or when the exposure is poorly defined.
Target trial	A hypothetical (and not necessarily feasible) 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 change	Theory of how and why the intervention impacts on outcomes of interest. One of a range of closely related terms used to refer to a conceptual model of how change occurs as a result of an intervention. Others include logic model and programme theory. A comprehensive theory of change should include aspects of context that may moderate the effects of the intervention as well as characteristics of the intervention itself.
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.
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 average treatment effect in the treated (ATT), also called the per protocol effect, is the ATE in the subpopulation who are actually exposed. The local or complier average treatment effect (LATE or CATE) is the ATE among compliers, i.e. units whose exposure status is determined by the assignment process being used for identification of the causal effect. Different methods estimate different causal effects (see *Quantitative methods*).
Triangulation	In statistical analyses, triangulation refers to the comparison of effects obtained using different methods or sources of data 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 to study the same phenomena, without necessarily seeking convergence.

CATE, complier average treatment effect.

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. *Box 1* and *Table 3* define key terms within the natural experimental evaluation literature as they are used in this guidance.

BOX 1 Natural experiments and natural experimental evaluations key points

We define natural experiments as events or processes (often referred to as 'data-generating processes') that are outside the control of researchers and that divide populations into exposed and unexposed groups, or into groups with differing levels of exposure. Comparisons of the exposed and unexposed groups, or of groups with differing levels of exposure, are used to make inference about the effects of exposure.

In a natural experimental evaluation, the data-generating process is the implementation or, more rarely, the withdrawal of an intervention, such as a policy, programme, service, piece of infrastructure, etc.

Natural experimental evaluations are a subset of all natural experimental studies, which in turn are a subset of all observational studies.

An important aim of the guidance is to encourage wider use of the opportunities provided by natural experiments to generate evidence that is useful for population health decision-making. For that reason, we do not restrict the definition of natural experiments to those involving an 'as-if random' data generating process. As-if randomisation is a strong basis for causal inference and a useful way of defining one end of the spectrum of natural experimental studies, but there is a range of methods for dealing with departures from randomness that allow useful evidence to be extracted from natural experiments that fall short of as-if randomisation.

Understanding how exposure is determined and how departures from randomness can be dealt with is a key feature of the design and conduct of a natural experimental evaluation.

Design and planning: concepts

Summary

- 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.
- Three important phases in the scoping and planning of natural experiment evaluations are identifying and theorising natural experiments, assessing their evaluability and conducting feasibility studies for a future evaluation.
- Mixed-methods designs are often needed to provide a comprehensive understanding of a natural experiment. Theory and planning are needed to bring different study designs, types of data and analyses together.

The MRC/NIHR framework for the development and evaluation of complex interventions[10] 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 by researchers, and they are often evaluated retrospectively rather than prior to full scale implementation. This implies a process of evidence generation ('practice-based evidence') that works in the opposite direction to the conventional translational research pipeline (*Figure 1*).

Otherwise, the phases and the core elements of the MRC/NIHR complex interventions 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. Methods for the evaluation of natural experiments are covered in detail in *Quantitative methods*, *Economic evaluation* and *Qualitative methods*. In this section, the study focuses on the early stages of identifying and appraising opportunities for a natural experimental study and working out a feasible and appropriate design (*Figure 2*). For guidance on other, more general aspects of evaluation research that are not unique to natural experimental studies, such as the importance of characterising and describing the intervention, the researchers refer readers to the complex interventions framework.

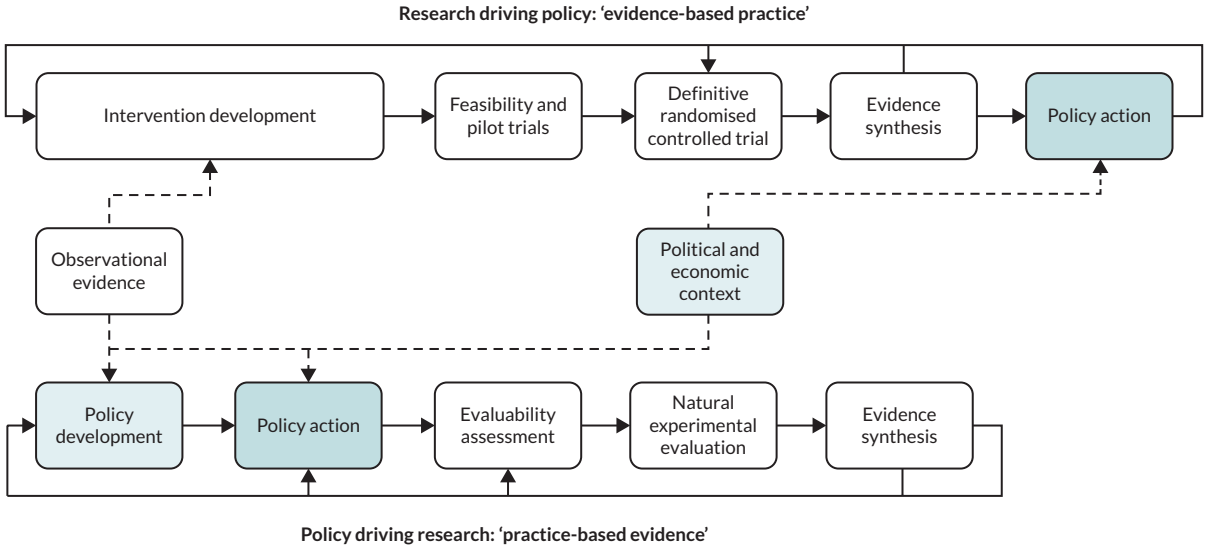

FIGURE 1 Two complementary modes of evidence. Source: Ogilvie *et al.*[26]

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 (see *Table 4*). One is where there is a clear division in presence, level or type of exposure between two or more otherwise similar subpopulations 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 USA or the minimum unit price at which alcohol can be sold in some UK nations 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,[38] or means tests that define entitlement to social security benefits. A third is where a policy is implemented in a phased way across a population, so that observations in different people or at different times or places can be categorised as exposed or unexposed. A related scenario is where the intervention is implemented all at once, but timing of data gathering, for example via a general population survey based on a random probability sample, is unrelated to the timing of policy implementation.[39,40] A fourth is where randomisation is built into the policy, as in the case of the Vietnam draft lottery[34] 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.[41]

Taking a systems perspective in evaluations of natural experiments

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 they are implemented. Natural experiments typically occur within complex systems that influence health. Such systems comprise several interacting elements and feedback loops 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.[42] 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 finance ministry as elements of the cigarette taxation system, whose behaviours 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 conditions might be necessary 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 the next section, we suggest ways in which the design of a natural experimental study can take account of these interactions.

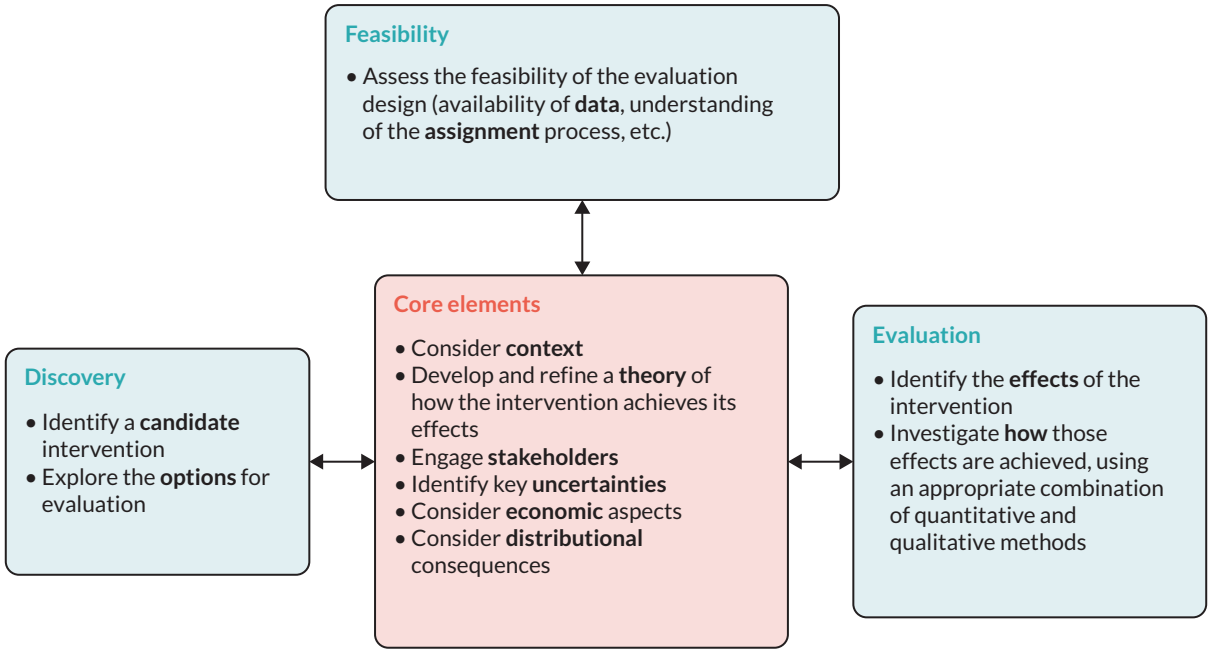

FIGURE 2 A framework for planning natural experimental evaluations: an adaptation of the MRC/NIHR framework for developing and evaluating complex interventions.[10]

TABLE 4 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;[27] English Teenage Pregnancy Strategy[28]
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;[29] eligibility rules within social security systems[30,31]
Phased implementation of a policy across a population	Rollout of Universal Credit[32,33]
Randomisation used as an assignment mechanism within a policy	Vietnam Draft Lottery;[34] housing vouchers[35]
Flaws or shortcomings in policy implementation	Database errors[36,37] and false-negative test results in the UK's Test and Trace programme

Mixed and integrated methods

A diverse range of methods can be considered when planning a natural experimental evaluation (synthesis of evidence from multiple natural experimental studies is discussed in *Reporting, critical appraisal and evidence synthesis*). To be most useful, evaluations should seek to explain how an intervention achieves its effects as well as to estimate effect sizes, and so a mixed-methods design that draws upon both quantitative and qualitative data will be needed. There are a number of ways of conceptualising how qualitative and quantitative methods can be integrated within an evaluation[43,44] (*Box 2*).

Evaluations involving multiple analyses, often using diverse methods, require the evidence to be collated. Although this integration is often considered at the end of the project, it is more likely to be successful if it is planned in advance,[49] for example by including an integration work package in the project plan (*Figure 3*) and explicitly earmarking resources for it.[50] For example, sequential designs will require careful consideration of the timing of components of the study to ensure that findings from one component can feed into the next steps of analysis for another; in parallel convergent designs, sufficient time must be allocated to integrate all analyses in the final logic model.

Process tracing[51] is one approach to integrating different types of evidence within a single study, with the aim of strengthening causal inference (see *Box 3*). The available evidence is categorised in terms of its uniqueness and definitiveness in relation to a causal hypothesis. Evidence is unique if observing it supports only the proposed hypothesis (i.e. specific, because it would be unlikely if some alternative hypothesis were true) and definitive (i.e. sensitive) if failure to observe it disconfirms the hypothesis. Evidence can therefore fall into four classes or 'test types'. It may be neither unique nor definitive (a 'straw in the wind'), unique but not definitive (a 'smoking gun'), definitive but not unique (a 'hoop test'), or both unique and definitive ('doubly decisive'). Evidence from a well-designed experiment (natural or planned) can be doubly decisive because alternatives to the preferred hypothesis can be ruled out by design. An advantage of process tracing over other methods of integration is that it makes the procedure for evaluating a causal hypothesis transparent (see *Box 3*).

BOX 2 Possible designs for mixed-methods studies

Sequential exploratory design: one method is first used to explore and understand the natural experiment and its context, and then another method is used 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. These identified both anticipated positive impacts and also potential unintended adverse impacts, which were then evaluated in further quantitative and qualitative studies.[45]

***Sequential 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 were used to explore the reasons for no observable reduction in distances walked, despite an increase in the number of bus trips made.[46]

***Parallel convergent design*:** a design which entails studying the same phenomenon using different data sources, methods or theoretical perspectives throughout the study. Triangulation can be used to bring together qualitative and quantitative analysis at the end, or to combine evidence from a range of qualitative or a range of quantitative sources. For example, 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 methods were used to assess the effects of this natural experiment.[47]

***Integrated design*:** drawing on multiple analyses to answer different but highly inter-related research questions. For example, Alvarado *et al.* (*Box 3*) 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 the method of 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.[48]

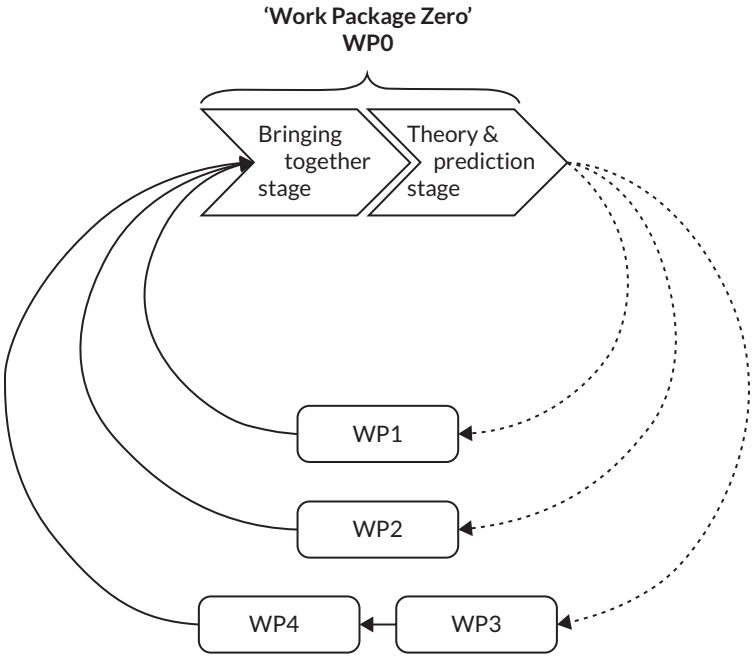

FIGURE 3 Incorporating an integration work package within a project plan. Source: Alvarado *et al.*[50]

BOX 3 Using process tracing to integrate qualitative and quantitative data

Alvarado and colleagues used process tracing in a study that combined qualitative content analysis with an interrupted time series analysis to examine the effect of a sugar sweetened beverage tax on the consumption of sodas and sugar sweetened juice drinks in Barbados.[48] The researchers theorised that risk signalling might explain the impact of the tax, if its introduction changed public awareness of the risks associated with consuming sugary drinks and perceptions of the types of drinks affected by the tax, leading to fewer purchases of those drinks. For each step in the causal pathway, they set out their expectations of what the evidence from the relevant component of the study (qualitative interviews, content analysis of media coverage and a time series analysis of sales data) would show if the hypothesis were true or false, and what kind of test the evidence would represent. For example, evidence from point of sale data that showed the volume of purchases declined was classed as a hoop test, because a fall could reflect a response to increased price rather than a change in risk perception, whereas a lack of change would rule out the possibility that changes in perceptions had caused a fall in purchases. Reports from participants in the qualitative interviews that their perception of risks associated with sugary drinks had increased were classed as a smoking gun test – such reports would support the hypothesis that the tax was changing perceptions, but a lack of them would not disconfirm the hypothesis as respondents could be unaware that their beliefs had changed. The study found that perceptions of sodas changed as a result of the tax and associated media coverage, and sales fell, but perceptions of juice drinks did not change and consumption increased slightly – possibly due to changes in producers' marketing practices. The different effects for sodas and juice drinks suggests that price changes were not the sole cause of the changes in sales.

Natural experimental studies that draw mainly or exclusively on qualitative data, using methods such as qualitative comparative analysis (QCA)[52,53] or case study designs,[54] are also valuable but are beyond the scope of this guidance.

Design and planning: practicalities

Summary

- An evaluability assessment can help to ensure that a natural experimental evaluation addresses questions that are relevant for decision-making and that stakeholders understand its strengths and limitations.
- Feasibility testing for a natural experimental evaluation should focus on the practicalities of implementing the evaluation design, such as availability of data on the assignment process, outcomes and confounders.
- Publication or registration of a study protocol enables other researchers and users of the evidence to understand what aspects of study design were decided in advance of obtaining data, and whether and how plans were modified in light of the data.
- Stakeholders should be engaged throughout the evaluation. This will help to ensure that findings will be understood, taken up and used for decision-making. Engagement should be based on a clear, mutual understanding of the roles of researchers and stakeholders, to prevent conflicts of interest from influencing the conduct and reporting of the evaluation.

This section addresses practical aspects of designing and planning an evaluation of a natural experiment, including assessment of the evaluability of the natural experiment, testing the feasibility of the evaluation design, the importance of publishing a study protocol, and the involvement of stakeholders. Incorporating these aspects into the process of designing an evaluation contributes to minimising bias in the evaluation. Other sections of the framework address more specific issues such as dealing with bias in quantitative analysis (see *Quantitative methods*), how qualitative evidence strengthens evaluations and how a multimethod design can offset design weaknesses (see *Qualitative methods*), and assessing bias in an evaluation (see *Reporting, critical appraisal and evidence synthesis*).

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.[55] It often involves structured engagement with community, policy and practice stakeholders to develop an agreed conceptual model of how the intervention is expected to achieve its impacts, 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 overall and distributional (i.e. equity) effects is it plausible to expect given what is already known, both from previous studies of this type of intervention and from information about the feasibility of the current implementation; and how might the results of an evaluation contribute to the wider evidence base?[26,56]

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. The direct involvement of stakeholders in the evaluation planning process helps ensure a shared understanding of what an evaluation can and cannot deliver given resources and other constraints on evaluation design. If the study has been initiated by researchers rather than by policy or other decision-makers, a collaborative 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. Whether by a formal collaborative evaluability assessment or some other process of data gathering, establishing a clear understanding of the assignment process for the intervention is central to robust natural experimental evaluation and will often depend on communication with stakeholders involved in intervention design or implementation.

Feasibility studies for natural experimental evaluations

Once the basic design options for the evaluation have been identified, a thorough assessment of their feasibility should be carried out.[56] Funding may have to incorporate contingency in case the proposed evaluation is unviable, for example by including an explicit breakpoint in the award at which a formal decision whether to continue would be made. Whereas feasibility studies for a RCT need to consider both the practicalities of delivering the trial intervention and the feasibility of implementing the trial procedures, the focus of feasibility assessment 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; and if there is no comprehensive source of routinely collected data, are there feasible and timely ways of collecting primary data? Practical consideration of resources, the amount of funding, staff and time available, is important to understanding what can be achieved in the evaluation.

Study and protocol registration

A natural experimental evaluation will often use multiple datasets and methods of analysis. This makes it challenging to know whether all analyses and results have been reported, rather than a subset which show effects in line with some prior expectation about the pattern of benefits or harms. Because evaluations of natural experiments are often conducted retrospectively using data collected for other purposes, 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 such prior knowledge. For the sake of transparency, it is good practice for a study protocol, covering both qualitative and quantitative methods, to be available in the public domain before analysis commences. Some research funders require publication of a protocol on their webpages as a condition of funding, registration of study protocols is now a condition of publication in some leading journals,[57] and many journals publish protocol papers. 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 natural experimental evaluation protocol, although recommendations have been made[58] and a guideline is in development for reporting protocols for observational studies.[59] In the meantime, the *Standard Protocol Items: Recommendations for Interventional Trials* (SPIRIT 2013) guideline for reporting clinical trial protocols may be useful for identifying key features of the study to include in a protocol.[60] We recommend that, as a minimum, protocols for a natural experimental evaluation 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 and the context in which it is implemented; a description of the assignment process and definitions of the different exposure groups; the proposed design, methods and analysis plan, including sensitivity analyses, robustness checks, and

the approach that will be used to integrate findings from the different study components; and information about ethical approval and 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 access to datasets and variables, the distributions of key variables and the extent of missingness (among other issues) are known. Protocols may also require amendments as new information emerges about the assignment process, how the intervention works or, in a prospective natural experimental evaluation, 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, or vice versa. Registration should allow such amendments, but the process must be transparent. Baldwin *et al.*[61] recommend a 'decision-tree' approach, noting and time-stamping any changes to the analyses as the study progresses.

Pre registration 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.

Engaging stakeholders

In an evaluation, the perspectives of a diverse range of stakeholders may be required to understand the natural experiment and its associated context as well as to contribute to ensuring a viable evaluation design. Stakeholders may include legislators, policy-makers, organisations and individuals responsible for implementation,[62] institutions enabling access to necessary datasets, advocacy groups, representatives of affected communities and recipients of the event or intervention, as well as the funder(s) of the evaluation. Involving all of the diverse interests in the natural experiment and its evaluation requires a thorough and inclusive approach to citizen engagement.[63] The individuals and organisations who are stakeholders in a retrospective evaluation may differ from those involved in the natural experiment, and perspectives will be required from both groups of stakeholders.

It is important to secure good stakeholder buy-in to the evaluation without jeopardising scientific integrity. To maintain the independence of the evaluation, a clear, mutually agreed understanding of each stakeholder's remit is needed. Individuals or organisations whose policy or programme is being evaluated may have a vested interest in the results of the evaluation, and strong prior expectations about what an evaluation will show. This applies equally to commercial and non-commercial stakeholders, and evaluation researchers themselves may become invested in the evaluation outcome as their working relationship with stakeholders develops. The risk of conflicts of interest[64] can be mitigated by setting explicit terms of reference for stakeholder involvement in advisory or steering groups, which participants are asked to sign, and by registering or publishing a protocol of the planned evaluation.

Quantitative methods

Summary

- A variety of study designs and quantitative analytical methods can be used in natural experimental evaluations, depending on the nature of the research question(s) and the data available to researchers.
- As natural experimental evaluations use non-randomised designs, threats to internal validity are always a concern. Often, a combination of methods will be needed to address the threats.
- Presenting quantitative analytical methods as though they form a hierarchy is unhelpful. The choice of method or methods should be determined by the research questions and may be constrained by the availability of data. Each has its own strengths and weaknesses and may be the best available in some circumstances but not others.

This section provides an overview of the main quantitative analytic methods available to the researcher when conducting a natural experimental evaluation. Elsewhere in this guidance, we use *design* to refer to the overarching

approach to an evaluation (which may include multiple types of data and analytical method) and *method* to refer to a specific type of analysis. However, many quantitative analytical methods are often referred to as 'study designs' in other literature, and in this section we have adopted that terminology in places to avoid confusion.

Considerations when selecting quantitative methods

The primary considerations for conducting a natural experimental evaluation 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 or methods.[13] A useful starting point, known as target trial emulation, is to consider what the features of a randomised trial would be, were it feasible or ethical to conduct such a trial to answer the research questions.[65] An important criterion for appraising the strength of a natural experimental evaluation is the plausibility of 'as-if randomisation' of the intervention – that is, how closely the process determining exposure emulates random allocation in a trial.[14] Whether the assignment process is as-if random cannot be fully determined by any type of quantitative analysis, but the judgement can be aided by a qualitative assessment of how the assignment process works.[1] Although the target trial framework is useful for highlighting possible threats to internal validity, it should be noted that some aspects, such as blinding, may be impossible to emulate. Sometimes the natural experiment might avoid a risk of bias incurred by the target trial, for example if there is a reduced risk of participation bias when using routinely collected data for the evaluation, or if there is a lower risk of confounding when trial procedures aimed at maintaining compliance do not apply.

Availability of data is a key constraint in evaluation design. Often a researcher will have to opt for a suboptimal evaluation design because the ideal data are not available, particularly when a natural experimental evaluation relies on routinely collected outcome data. For example, where individual-level data are not available, the researcher might be able to make use of data aggregated to some grouped level such as schools, geographical regions or countries. Likewise, when historic (i.e. pre intervention) data are not available, the researcher may need to rely on a cross-sectional study instead of a repeated cross-sectional design. Even where pre- and post-intervention data are available, the time-series may be too short to deal with temporal patterns in the data other than the effect of the intervention. Decisions about evaluation design and choice of analytical methods made in the light of such constraints should be clearly documented in the study protocol (see *Design and planning: practicalities*).

As natural experimental evaluations use non-randomised study designs, there will usually be a degree of selection into the exposed or unexposed group, for example as a result of self-selection or because the intervention is targeted at specific areas or individuals with particular characteristics. This contrasts with randomised study designs where individuals, groups of individuals or some other kind of cluster are randomised by the researcher. As a consequence, natural experimental evaluations tend to estimate average treatment effects on the treated (ATTs) rather than average treatment effects (ATEs). The exceptions are regression discontinuity and other instrumental variable methods which estimate local average treatment effects (LATEs). Just as for randomised study designs, it is important to be clear about (1) what treatment effect is being evaluated [by specifying the causal contrast or estimand(s)][13] and (2) whether, for users of evidence, the main interest is in conditional (i.e. the average effect of the intervention on an individual) or marginal (i.e. the average effect of the intervention on the population) treatment effects.[66]

In any natural experimental evaluation, causal inference requires the use of a counterfactual which represents the outcome that would be observed in the absence of the intervention. The degree of comparability (or 'exchangeability') between the exposed and unexposed groups (or between groups with different degrees of exposure) determines how far differences in outcomes can be interpreted as effects of the intervention. The methods outlined below take different approaches to achieving exchangeability, with some (e.g. matching, multivariable regression, propensity scores) using measured variables ('observables') – and others [e.g. difference in differences (DiD), instrumental variables, regression discontinuity] accounting for unmeasured variables ('unobservables'). 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 been shown that aggregated level analyses can be prone to bias if data at the individual level are missing at random.[67] If only aggregated level data are available, the quality of the underlying data sources should be interrogated and the likelihood and nature of missing data and other sources of bias (such as changes over time in sampling or measurement of outcomes) assessed as far as possible.

Before outlining different natural experimental study methods it is important to emphasise that the researchers are not proposing a hierarchy in terms of methodological quality. This is partly because the design should preferably be determined by the research questions and by the nature of the data available, but also because each method has its own strengths and weaknesses. The importance of these strengths and weaknesses will vary from case to case. For example, a controlled interrupted time series (cITS) might be regarded as superior to an ITS because it can address confounding by trends or events that affect both the exposed and the control series. But if the control group is poor (e.g. because the parallel trends assumption[68] 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 assumptions of the method are not met or if the instrument itself is 'weak' (i.e. is only weakly correlated with exposure to the intervention).[69] Each of the methods we describe should be implemented with appropriate tests of assumptions, robustness checks and sensitivity analyses.

In the next section of this chapter, we summarise the methods most widely used for natural experimental evaluations. Some, such as the synthetic control methods, are relatively new, and the more well-established methods are being continuously refined, so it is important to remain open to continuing developments (*Box 4*). Given the rise and prominence of machine learning, it seems likely that this will have a role in future research for predicting the counterfactual based on algorithms identifying patterns and relationships in pre-intervention data.[74] For example, in a study evaluating the impact of social housing on mental health, machine learning methods were used to maximise the precision of weights used in an inverse probability treatment weighted approach.[75]

Finally, although the study designs and statistical approaches outlined in this overview aim to estimate causal effects of interventions, how these effects come about may remain unclear. While qualitative evidence is important for understanding causal mechanisms (see *Qualitative methods*), quantitative data can also be used to explore such mechanisms – for example, causal mediation analysis can be used to understand how variables operate along the pathway between intervention exposure and outcome.[76]

Overview of key quantitative methods

In *Table 5*, the rows provide summary information on the methods described below. The final column illustrates how each study design might be employed to evaluate a hypothetical cardiovascular disease (CVD) 'health check' in primary care that aims to reduce incident CVD events. As the researchers emphasised in the paragraphs preceding, each design has its own strengths, weaknesses and assumptions so the ordering of rows should not be taken to represent a methodological quality hierarchy. In the table, the researchers have treated exposure to the intervention as being uniform across the exposure group. However, it could be that some groups within the intervention group receive different 'doses' of exposure and/or the timing of exposure varies, similar to 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.

Cross-sectional design

Uses data at a single point in time or over a brief period of time. An advantage of this design is its simplicity. However, it is limited because the outcome is measured only after the intervention has started, but not before (either in the intervention group or in any control group). Even if the intervention group is matched to (1) control group(s) based on possible confounding variables, not knowing baseline/pre-intervention time trends in the outcome is a major weakness

BOX 4 Recent developments in quantitative analytic methods

Causal inference from the methods described previously can be strengthened conceptually through the use of Directed Acyclic Graphs (DAGs), for example as used in an evaluation of the impact of crime maps on house prices.[70] Causal inference can be strengthened analytically by conducting the analyses within a Bayesian framework, for example Bayesian synthetic controls[71] which have been used to evaluate the impact of alcohol licensing on public health[72] by including weighted Bayesian model averaging to mitigate issues of arbitrary selection of covariates and overfitting. Outside of a Bayesian framework such weighted model averaging methods can be achieved through, for example, ensemble models.[73] Further strengthening of causal inference can be achieved through triangulation of results based on different quantitative data sources or different analytic methods, falsification and sensitivity tests,[13] and triangulation of inferences from qualitative and quantitative analyses (see *Design and planning: concepts*).

TABLE 5 Quantitative methods for evaluating natural experiments

Study design	'Level' of data collection	Data	Statistical approaches	Overview	Illustrative example – CVD 'health check' delivered in primary care centres (i.e. screening for CVD risk factors). Outcome measure – incident CVD events (hospitalisations/ deaths)
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)	Allows for estimation of effect, assuming it is 'known' what outcomes were pre intervention and/or outcomes are same in intervention and control groups pre intervention	Study conducted in all, or 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
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, or rates (depending on nature of outcome measure variable) – with representation of uncertainty; regression models or propensity scores to adjust for confounding variables and/or 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, or subset of, primary care centres pre and post intervention (two time points); difference in rate of incident CVD events compared pre and post intervention
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 or propensity scores – to adjust for confounding variables and/or 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	Not possible given nature of outcome measure (incident CVD events); would be possible if, e.g. systolic blood pressure, SBP was outcome measure
Regression discontinuity	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	Non-parametric methods; regression models; assess effect modification.	Can help minimise bias due to unmeasured confounding. 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) > 140 millimetres of mercury, mmHg] as 'cut-off'. Use distance to primary care centre where health check is being offered as 'instrument'

continued

TABLE 5 Quantitative methods for evaluating natural experiments (*continued*)

Study design	'Level' of data collection	Data	Statistical approaches	Overview	Illustrative example – CVD 'health check' delivered in primary care centres (i.e. screening for CVD risk factors). Outcome measure – incident CVD events (hospitalisations/ deaths)
Difference in differences	Individual or aggregate 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
Interrupted time series	Aggregate level	Pre and post intervention; data collection on multiple occasions, generally at equally spaced time intervals; 'interruption' is at time point when intervention starts; no data from control group(s)	Time series; (s)ARIMA or (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, or subgroups, of primary care centres)
Controlled interrupted time series	Aggregate level	Pre and post intervention; multiple time points of data collection (generally at evenly spaced intervals); 'interruption' is at time point when intervention starts; intervention and control group(s)	Time series; ARIMA or (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	Study time series of rates of incident CVD events in intervention and control group(s) or synthetic control group

ARIMA, autoregressive integrated moving average.

because the effect of the intervention cannot be distinguished from the effect of pre-intervention trends or the effect of unmeasured confounders.

Repeated cross-sectional design

Is a cross-sectional design with data collected on the outcome before and after the intervention starts, so that change in the outcome can be directly measured. However, a weakness is that these measurements are not necessarily made in the same individuals, so results could be biased by changes in the composition of the population in which outcomes are measured in each cross-section. It may be possible to adjust for known (and measured) confounding variables in a regression model, but as with the simple cross-sectional design, the effect of the intervention cannot be distinguished from the effect of pre-intervention trends or the effect of unmeasured confounders. An example of a natural experimental 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,[77] with data collected via an online questionnaire at baseline (prior to restaurant opening) and 3 and 9 months after opening.

Before and after design

Is similar to a repeated cross-sectional design but with repeated measurements available, before and after the start of the intervention, on the same individuals (i.e. individuals are acting as their own control[78]). A natural experimental 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.[79] An existing cohort can be used for such a study design, as was done in an evaluation of the impact of an earthquake.[80] Some definitions of this design allow for different individuals to be sampled in the 'before' and 'after' periods,[81] but we believe such a design is more helpfully categorised as repeat cross-sectional (see *study design* above).

Regression discontinuity design

In contrast to the methods introduced so far, a regression discontinuity design can account for confounding effects from unmeasured as well as measured confounding variables. Regression discontinuity methods can be used where assignment of participants to an intervention is determined by a threshold value of a continuous 'forcing variable', such as age or income. If assignment is not open to manipulation by recipients or administrators of the intervention, recipients close to the threshold value of the forcing variable are effectively randomised, enabling estimation of causal effects of the intervention. In a 'sharp' regression discontinuity design, an assignment variable determines completely whether an intervention is received, whereas in a 'fuzzy' design exposure is probabilistically assigned at a threshold of the assignment variable.[82] Geography can be the key feature of such designs.[83] An example of a natural experimental study using this type of design is one that assessed the effect of increased primary schooling on adult women's HIV status in Malawi and Uganda[84] 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.**[85] 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.[69] In an Indonesian study,[86] the amount of rainfall (which is strongly correlated with crop production) was used as an instrument to assess the impact of income on mental health outcomes (assuming that amount of rainfall was not directly associated with these outcomes).[87]

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

Difference in differences design

Essentially a controlled before and after design that can use either repeated cross-sectional or longitudinal data that is collected in both intervention and control groups. Multiple time points in the pre-intervention period allow a better estimate of 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[88] 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, obtaining data for multiple time points from the annual British Household Panel Survey. A DiD study can also be used on aggregated group data, in which case it is closely aligned with the cITS (see below). An example of this is a study by Dimitrovová *et al.*[89] which estimated the effect of national primary care reform on avoidable hospital admissions.

Interrupted time series design

Can be considered a repeated cross-sectional design that uses aggregated data (generally counts or rates) from more than two measurement occasions with equally spaced time intervals. Because temporal patterns are explicitly modelled in an ITS, the number of measurement occasions should be sufficient to capture the temporal pattern (e.g. for monthly data, at least 12 measurements pre intervention would be required to capture the annual pattern). An example of an interrupted time series design is the evaluation of the impact of lockdown policies in India on COVID-19 incidence[90] which used daily counts of new cases as the outcome and dates of stages of lockdown as the 'interruptions' that changed exposure status.

A controlled interrupted time series

An ITS design but with a control group in which measurements are made at the same time points as in the intervention group. If a suitable control can be identified (this can be assessed by pre-intervention time trends and comparison of

demographics and time-varying confounding variables) it can be argued that this provides a superior counterfactual to ITS alone. It is possible to incorporate more than one control group into the analytical design.[91] Another useful robustness check is to use outcomes which should not respond to the intervention (known as negative controls, control outcomes or non-equivalent dependent variables).[68] Changes in the level or trend in such outcomes associated with introduction of the intervention is indicative of residual confounding. An example of a natural experimental evaluation with a cITS design is a US study of the effect of a safe opioid prescribing initiative on levels of opioid prescribing[92] 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 interpreted as the causal effect. In a study from Ethiopia,[93] a synthetic control was used to assess a health extension programme 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.

The above list describes the main analytic methods that have been used in natural experimental evaluations but should not be considered exhaustive given the continuous development within the field (see *Box 4*).

Economic evaluation

Summary

- 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 in this context. These include the measurement and identification of costs and outcomes, selecting appropriate analytical methods, identifying the time horizon, and equity considerations.

For interventions that incur significant continuing costs, evaluations that focus only on effectiveness provide insufficient evidence to support decision-making. Economic evaluations compare two or more courses of action in terms of their costs and consequences,[94] by identifying, measuring and valuing costs and outcomes. Depending on the outcome of interest, economic evaluations can take the form of: 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 (QALYs); cost–benefit analysis (CBA), where benefits are valued in monetary terms; or cost–consequence analysis (CCA), where costs and outcomes are presented separately in a summary table.[95] In CUA and CEA frameworks, the results are usually summarised as an incremental cost-effectiveness ratio, which, in CUA analysis, is compared with a willingness-to-pay threshold (e.g. the NICE threshold of £20,000–30,000 per QALY in the UK[96]). The comparative analysis involved in the economic evaluation provides a systematic framework to support transparent and evidence-based decision-making about funding population health interventions that offer good value for money. Comparing costs and benefits of alternative interventions using an objective framework is essential to ensure well-informed decisions on the allocation of scarce resources and represents a key component in the evaluation of interventions.[97,98] However, good-quality economic evaluations are not routinely incorporated in natural experimental studies.[99]

Designing and conducting economic evaluations alongside natural experimental evaluations

There is a paucity of literature on the design and conduct of economic evaluations of population health interventions alongside natural experimental studies, and few examples of completed studies.[100,101] Most existing literature on natural experimental evaluations focuses on effectiveness only[102] or fails to evaluate costs and benefits in a systematic economic evaluation framework (e.g. Angrist *et al.*[103]). Most 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,[104,105] or addresses the more general challenges of conducting economic evaluations of interventions without considering the specific challenges posed by natural experiments.[106,107] Current available economic evaluation guidance is tailored to the most common RCT (control) framework and does not cover the challenges inherent in natural experimental study designs.

Key challenges for economic evaluations with natural experimental evaluations

Natural experimental evaluations pose specific challenges for economic evaluations.[108] The following paragraphs outline the key challenges and suggest ways in which they can be addressed.

Measurement of costs and outcomes and choice of perspective

The QALY is generally used as a health outcome measure in CUA economic evaluations, as it provides a summary measure of health and well-being which allows for comparison of interventions with different outcomes.

Linked administrative data used in natural experimental evaluations usually do not include preference-based outcome measures such as the EuroQol-5 Dimensions instrument to assess quality of life (QoL). Thus, the estimation of the effect of an intervention on QoL may have to rely on 'intermediate' outcomes which are subsequently mapped onto health economics outcomes (e.g. QALY or disability-adjusted life-year) using decision-analytic models, for example matching the increase of physical activity resulting from urban greenways to improvements in QoL.[109]

However, a cost/QALY analysis under a health and social care perspective is often too limited for the economic evaluation of population health interventions,[95,96] where a societal perspective is advocated.[95,96] Natural experiments are often used to evaluate intervention in policy sectors other than health and for which health impacts are by-product. Broader economic evaluation frameworks such as CCA or CBA[95] are more appropriate for incorporating multiple, multisectoral costs and outcomes (e.g. those falling in the education, environment or judicial sectors) besides health. Useful tools to consider include frameworks such as multicriteria decision analysis (MCDA),[110] which weighs and values multiple outcomes using an explicit method (e.g. a discrete choice experiment), or the framework proposed by Walker *et al.*[111] which explicitly considers and values opportunity costs across decision-makers belonging to different sectors.

Without a bespoke data collection instrument, it may be challenging to include multiple outcomes and costs/cost savings beyond the health sector. Therefore, the design of economic evaluations alongside natural experimental evaluation should include a thorough investigation of routine data, to identify the availability of data sources to identify relevant, multisectoral costs and outcomes. In the absence of bespoke data collection, proxy unit costs may be used to evaluate available resource use (e.g. average cost/bed-day unit cost), preferably including sensitivity analysis around the assumptions made in relation to unit costs.

Time horizon

The availability of routinely collected data for a long observational period allows an economic evaluation 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 population health interventions, which might exert their impact only in the medium-to-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 population health interventions.[112]

Analytical methods

When conducting economic evaluations alongside a natural experimental evaluation, 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,[113] skewed and non-normal distributions of costs and outcomes,[114,115] potentially correlated multiple outcomes,[116] and clustering.[117]

Depending on the method used to deal with observed and unobserved confounding, the ATE, ATT or the LATE will be estimated. These will be interpreted as the incremental cost-effectiveness parameter in the corresponding population, that is ATE: incremental effectiveness parameters in the general population; LATE: incremental cost and incremental effectiveness parameters among the population of compliers. Sensitivity analysis should be conducted using multiple designs and multiple control groups to explore the robustness of economic evaluation to multiple sources of bias and strengthen the credibility of results.

Decision analytic and microsimulation models

Decision analytic models (i.e. decision trees, Markov models) compare 'the expected costs and consequences of decision options by synthesising information from multiple sources and applying mathematical techniques'[118] (page 1). By making use of additional evidence, decision models have the potential to overcome some of the limitations of data availability and quality in natural experimental evaluations. Indeed, a decision analytic model can bridge the primary outcome of the study with a measure of relevance for the economic evaluation analysis (e.g. QALY) when preference-based instruments are not in the linked data or a longer-term outcome in presence of a data-constrained time horizon.

A number of approaches are available for economic evaluations that investigate the impact of a natural experiment as an event within a complex system.[119] Moving beyond a traditional cost-effectiveness framework, complex systems models include, among others, systems dynamics models, computational general equilibrium models, microsimulation models, agent-based models and partial differential equation models. Guidance has been developed to identify appropriate use and application of these models in economic evaluations.[119]

For example, dynamic microsimulation models (models that simulate the behaviour of micro-units over time) offer a valid and flexible tool for the evaluation of costs and outcomes associated with policies to improve population health, as well as allowing extrapolation of longer term health outcomes.[120] Specifically, these models can incorporate multisectoral costs and outcomes, avoiding using an additional method to weight and 'prioritise' multiple costs and consequences. By explicitly considering the dynamic interaction between individual level outcomes, they also go beyond the estimation of population-level average cost-effectiveness and allow the estimation of costs and effects at population subgroup level, therefore facilitating distributional CEA (see below).

Equity considerations

In *Design and planning*, we recommended that the distributional consequence of interventions should be a core consideration in the planning of natural experimental evaluations. Subgroup analysis or the inclusion of equity-relevant variables (e.g. poverty status, urban/rural classification) could provide useful insights on the equity impact of the interventions being evaluated in terms of cost-effectiveness.[121,122] More advanced methods such as distributional CEA[123] and extended cost-effectiveness analysis[124] should be also considered. However, as these methods are more demanding in terms of data requirements, they may not be feasible using observational data.

Qualitative methods

Summary

- Most natural experimental evaluations will benefit from a mixed-method design incorporating qualitative methods throughout the evaluation.
- Adequately characterising system, context and intervention will usually require qualitative evidence.
- Qualitative evidence contributes to better characterisation of exposed and non-exposed groups; understanding of the assignment process; understanding causal mechanisms; selection and measurement of meaningful indicators for outcomes; identifying potential costs of the intervention; and understanding stakeholder perspectives.
- Qualitative analysis can also be used to enhance the explanatory potential of evaluations.

Designs drawing wholly or primarily on qualitative methods can be used to evaluate natural experiments; existing guidance covers approaches such as realist evaluation,[125] and case studies.[126] Natural experiments can also be evaluated through designs such as QCA, which draws on a series of qualitative case studies to identify the configurations of conditions that are causally related to the absence or presence of an outcome.[52,54] Here, we focus on the role of qualitative methods in one particular set of designs: those evaluations that have a primary aim of estimating effect sizes of interventions through using differences in exposure. Typically, these employ qualitative methods in tandem with statistical analysis of quantitative data within mixed-method designs. In such designs, qualitative work packages are often incorporated into process evaluations. Guidance for process evaluations for complex interventions[127] is mainly aimed at trials, in which both development of the intervention and assessing fidelity are key issues. The principles of this guidance also apply to natural experimental evaluations, but qualitative evidence can and should do more than

contribute to process evaluation. For natural experimental evaluations, qualitative methods are essential, both to exploit the strengths of natural experimental designs for studying interventions in complex systems, and to minimise their design weaknesses in assessing 'as-if randomisation'.[14] To fully benefit the evaluation, qualitative methods should be integrated throughout. Most evaluations of natural experiments will require qualitative evidence to ensure that: the system, context and intervention are well characterised; groups are appropriately defined in terms of their degree of exposure, the assignment process is well understood and cases are appropriately assigned to those groups; causal mechanisms are understood; outcomes selected are meaningful; indicators for these outcomes are valid and reliably measured; and stakeholder perspectives on the intervention and its anticipated and actual effects are understood. In addition, when supported by analysis that goes beyond the descriptive, qualitative methods can add considerable value by explicating links between intervention and outcomes in context, and strengthening attributions of causality and transferability.

This guidance does not address in detail the selection of appropriate methods of data gathering and analysis for these tasks. Depending on the specific needs of a study and its theoretical framing, data might be drawn from: existing documentary data; individual and focus group interviews; participant or researcher diaries; ethnographic observation; photo elicitation; or a range of other sources. Rather, this guidance outlines the key components of an evaluation in which qualitative methods should be integrated to maximise the opportunities for robust and generalisable findings. Examples for each point are provided in *Table 6*.

TABLE 6 Contributions of qualitative methods to natural experimental evaluations, with examples

Evaluation component	Role of qualitative methods	Data generation and/or analysis methods used
Characterising the intervention, context and system		
Describing the intervention	Characterised the Daily Mile intervention, which encourages children to run for 15 minutes each school day, by comparing the intervention as described in principle in promotional material, with how it was implemented in practice.[61,62,128] Within a process evaluation, identified important factors that enhanced and hindered the implementation and normalisation of a complex intervention in maternity services to reduce smoking rates in pregnancy.[129]	Ethnographic observation, document review, interviews. Observations, semi-structured and group interviews, analysed using normalisation process theory
Describing the system/context	Used system-mapping to describe the complex adaptive system in which a proposed sugar sweetened beverage levy in the UK would be introduced, and identified key stakeholders' perspectives on its likely impacts.[130]	Expert workshop, Delphi exercise, and qualitative interviews
Developing theories of change	Informed the logic model for an evaluation of a proposed Graduated Driving Licensing in Northern Ireland using focus group analysis to: inform choice of plausible comparator settings, hypothesise links between intervention and potential public health impacts, and identify adoption of telematics insurance products as a co-occurring intervention in the system.[131]	Group interviews analysed using thematic content analysis
Informing selection of populations, controls, and subgroups for analysis	Informed appropriate dates for measuring change resulting from the intervention, and identified important subgroups for analysis of effects in evaluation of Cambridgeshire Guided Busway (a new bus network, traffic-free walking and cycling route), by identifying that the intervention was used before being officially open, and that prior experience of different modes influenced initial perceptions.[132] Used local practitioners' insights on appropriate comparator areas to contribute to creating synthetic controls as the counterfactuals for an evaluation of the impact of alcohol licensing decisions on local health and crime.[72]	Interviews, media analysis, photo elicitation and participant observation. Consultation with local practitioners
Characterising and selecting outcomes and indicators	Refined outcome indicators in an evaluation of free bus travel, by identifying that travelling by bus entails considerable physical activity for young people, so 'bus trips' as an indicator of passive travel would underestimate 'active travel'.[133] Refined interpretation of outcomes measured in a survey questionnaire of commuters through interview data, which suggested that suggested that some survey respondents' negative reports of walking and cycling environments reflected a desire to make a political point about poor facilities, rather than necessarily representing their own perceptions.[134]	Ethnographic observation; Photo-elicitation Content analysis of post-questionnaire interviews

continued

TABLE 6 Contributions of qualitative methods to natural experimental evaluations, with examples (*continued*)

Evaluation component	Role of qualitative methods	Data generation and/or analysis methods used
Generating data on outcomes	Gathered evidence of outcomes not identifiable through routine datasets for an evaluation of the impact of reduced street lighting at night, such as experiential and behavioural impacts on well-being of light and dark at night.[135] Identified important and unanticipated outcomes of dance mats to increase physical activity in secondary schools, including improved reaction times and co-ordination skills, and acceptability to older girls.[136]	Individual and group interviews, media analysis. Before and after interviews and focus groups
Understanding mechanisms and mediators	Identified possible reasons for lack of effect on health behaviours or physical activity in older adults of low-cost improvements to local urban green spaces in Manchester, UK: that small green spaces were seen as belonging to others, and that residents preferred larger parks.[137,138] Developed understanding of limited efficacy of Smoke-Free Schools policies in reducing adolescent smoking through analysis of discourses of school children in seven European countries, which identified that policies were associated solely with 'the school', and that they displaced smoking to other spaces.[139] Explained the impact of health system context on outcomes in an evaluation of mergers of urology departments in Denmark, in which readmission rates went down, but length of stay increased after restructuring; qualitative analysis suggested that the expected efficiency gains of centralisation are undermined in contexts of cost constraint and external pressure.[140]	Walk along interviews and photo-elicitation. Critical discourse analysis of focus group data Interviews with service providers and institutional theory
Explaining change	Tested the plausibility of the 'signalling effect' (i.e. the policy debate itself draws attention to an issue among the public and triggers behaviour change) as a mechanism through which taxes on sugar-sweetened drinks in Barbados reduced sales.[48] Identified the mechanisms of change and necessary components for transferability of an intervention to provide free bus travel, though 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).[46] Explained subgroup differences in outcomes of a social prescribing intervention for people with diabetes by analysing how 'health capital' and structural conditions shaped participants' capacity to interact with and benefit from the intervention.[141]	Process tracing, using television news archives, interviews, the public, and point of sale data. Inductive qualitative analysis of focus group, interview and observational data. Ethnography involving interviews, photo-elicitation and participant observation
Understanding stakeholders' perspectives	Identified unexpected perceptions of residents in evaluation of impact of 2012 Olympics in London, who felt safer with, rather than marginalised by, enhanced security in their neighbourhoods.[142] Produced evidence to explore the unexpected effects of the COVID-19 pandemic on the implementation of a mass transit cable car in Bogotá, Colombia, through exploring residents' and policy stakeholders' perspectives on the likely impacts and the historical context of the intervention.[143] Identified limitations in likely effectiveness of future transferability of home energy efficiency interventions in England, through analysis of householders' motivations for installation, which identified that current policy framings around 'environmental sustainability' resonated poorly.[144]	Family narrative interviews, go-along interviews, focus group workshops. Citizen science methods, involving public volunteers in planning, conducting and analysing evidence. Household interviews

Characterising the intervention, context and system

Describing the intervention

The interventions in natural experimental evaluations are often context-specific: programmes, laws or organisational changes that have emerged for particular jurisdictions or populations. To ensure evaluations are useful for those in other contexts, the intervention should be well described.[62] This will include components such as characterising the rationale for introduction; details of any underpinning legislation; any organisation(s) that have a role in delivering the intervention; how it was delivered; where costs and benefits from the intervention might be incurred; and what flexibility there was or is in delivery. Although reviewing documentary sources may be sufficient for describing some components, other evidence (e.g. from ethnographic observation or interviews with individuals delivering or receiving the intervention) may be needed to describe how and why the intervention was implemented,[145] how it was delivered in practice and with what costs. Qualitative evidence on intervention acceptability and fidelity is also valuable for describing an intervention.

Describing the system/context

A strength of natural experimental evaluations is that they can explore relationships between the system/context and intervention, help characterise the complexity of the system, and account for mechanisms relating to the intervention as an interruption or emergent change in that system.[146] Multiple sources of data will be needed to understand this complexity, particularly for identifying what aspects of the context/system interact with the intervention and how. Qualitative contributions might draw on documentary analysis, observations, interviews and other sources as appropriate to describe the relevant demographic, geographical, historical, economic, political and organisational aspects of context.[147] An important element of qualitative work here is identifying any relevant co-occurring interventions in the setting that are likely to affect changes in outcomes over time or across exposed/non-exposed populations. System mapping techniques can be used to identify the components, and scales, that are important to include in the evaluation, and to aid in assessing the evaluability of an intervention[56] before undertaking an evaluation.[148] A system map, for example a causal loop diagram (CLD), can also be used to provide a robust theory of change for the intervention, and to develop a model to simulate intervention impacts.

Developing theories of change

Qualitative methods such as interviews and workshops with stakeholders help develop an initial theory of change for the intervention. Participatory methods can help ensure that all important preconditions, assumptions, positive and negative outcomes, potential mechanisms and aspects of the system are considered in the evaluation.[149] In natural experimental evaluations, this theory of change can be a dynamic model of how the intervention is theorised to impact on outcomes (e.g. a CLD), and can be revisited throughout the evaluation. Theories of change can hypothesise the structure and mechanisms of the system, and identify the leverage points for system change such as reinforcing feedback loops. At the end of the project, a good theory of change can assist with mapping all evidence for causal pathways investigated. It should ensure that findings from both qualitative and other work packages are accounted for, synthesised, and (in systems-orientated evaluations) inform modelling and visualisation of how the system has changed as a result.

Informing selection of populations, controls and subgroups for analysis

Initial qualitative work to characterise the intervention, context and broader system will inform and sense-check the assignment of units (such as individuals, or areas) to exposed and non-exposed populations, and the processes through which this assignment happened. For example, qualitative work will enable better characterisation of precisely how, when and where an intervention was implemented, and how it was taken up or responded to in initial phases.[40] This can help with selection of appropriate time frames for measuring changes that can be plausibly attributed to the intervention. Similarly, qualitative evidence on population interactions with interventions can help define geographical areas of exposure. Qualitative evidence on delivery of, uptake of, attitudes to or experiences of interventions and stakeholder priorities can also be used to identify important population subgroups for analysis, or appropriate geographical areas for comparison.

Characterising and selecting outcomes and indicators

The choice of appropriate primary and secondary outcomes, and indicators for these, should be informed by qualitative evidence. This can help ensure that the quantitative indicators selected capture what is intended, as well covering key outcomes that are important for stakeholders. Existing qualitative evidence might furnish useful information about the relevance, validity, limitations and credibility of existing quantitative measures of health and behavioural outcomes. Primary qualitative data are often also needed to understand better the strengths and limitations of quantitative measures that are selected, particularly if these are from routine secondary datasets generated for other purposes. These insights help make credible inferences from quantitative findings. Qualitative evidence helps strengthen the interpretation of statistical findings through shedding light on the meaning of indicators in context, the potential limitations in survey or other measures, and the interpretation of relationships between indicators.

Generating data on outcomes

Estimating the effects of the intervention on primary outcomes usually requires quantitative data, but qualitative evidence strengthens analysis in providing sources of data for triangulation. This can aid the credibility of claims made through the convergence or divergence of findings, and aid analytical inference through comparisons. There is also an important role in identifying secondary outcomes that might not have been captured in quantitative datasets or were

not anticipated at the outset.[150] Where quantitative data for outcomes or population subgroups do not exist or would be difficult to access, qualitative methods (such as interviews) are sometimes necessary for generating the primary source of evidence on changes in knowledge, understanding or practices that are associated with an intervention.

Understanding mechanisms and mediators

Incorporating qualitative evidence is essential for maximising the potential for a natural experimental evaluation to produce transferable, credible, and robust findings. While sophisticated quantitative analyses can explore relationships between points on hypothesised causal pathways, qualitative methods have particular strengths in exploring processes. In general, they do this by means such as observing parts of causal processes in action, analysing documentary evidence, or understanding processes of change from participants' perspectives. During the evaluation, as a nested or concurrent study, detailed qualitative analysis can also help unpack variations in impact, such as how and why the intervention effects differ across subgroups or over time, and can surface unintended or indirect outcomes of the intervention in a system.

Explaining change

Beyond these descriptions of causal mechanisms, qualitative analysis can contribute to making analytical inferences about causal processes from comparisons within the case.[151] This draws on approaches such as process tracing[152] or analytic induction.[153] Appropriate qualitative analysis, in the light of theories of change, thus strengthens causal inferences and claims about transferability.

Qualitative evidence is often invoked as particularly useful for explaining unexpected findings or policy failures. For instance, qualitative evaluations might be recommended to identify barriers to adoption or unanticipated outcomes.[154] 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 unanticipated effects and 'failure'.

Understanding stakeholders' perspectives

Natural experimental evaluations should include appropriate engagement of relevant stakeholders (such as the patients, publics, practitioners and policy-makers affected by the intervention) in selecting, planning and undertaking the evaluation. However, it is important that this is not the only source of evidence on stakeholders' perspectives on the intervention and its impact. Formal qualitative research is also needed not only for mapping the perspectives and practices of all stakeholders, but also for analysing these 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 populations directly affected by the intervention, which may not reflect expectations or theory. Qualitative methods also help understand the knowledge, perspectives and practices of those planning and implementing policies or programmes, which might include public health practitioners, industry actors, or policy actors. This is essential data to identify the alterations that affect how interventions are implemented on the ground, and any direct and indirect effects arising. 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 claims about future roll out, scale up, or transferability of interventions.

Maximising the value of qualitative components

To ensure findings from the qualitative components of a study are fully exploited, they should be appropriately integrated into the whole evaluation, from design to final report. This requires, first, planning of the qualitative work for timely incorporation of data generation and analyses, for instance, at key points such as final selection of outcome indicators, or to inform subgroup analyses. Second, sufficient time needs to be allocated for integrating the final analysis in revised theories of change.[50] Third, this requires an investigator with appropriate skills and training to design and manage the qualitative components and, importantly, to integrate a sound theoretical underpinning for the evaluation. A robust theoretical framework is needed to ensure that relationships between indicators (of, for instance, behavioural changes) are related conceptually to the capacities or functions[155] of the intervention, such that the likely transferability

(or not) of findings can be judged. Appropriate analytic approaches are also needed for all qualitative datasets. Descriptive or thematic content analysis may be sufficient for some of the above aims, such as documenting key aspects of context or stakeholders' perspectives. Others aims (e.g. contributing to causal inference) may require more analytical approaches, such as analytical induction or process tracing.

There is (often under-exploited) potential for utilising existing qualitative datasets in evaluations of natural experiments. Documentary sources (policy reports, records of meetings, social media captures) are invaluable as evidence for the chronology of intervention implementation, but also potentially useful for documenting change over time. On some common outcome indictors (e.g. around the meanings of travel mode, physical activity, smoking cessation) there may be existing published studies or datasets of transcripts from comparable settings which can be interrogated to enhance understanding prior to finalising evaluation design or outcome selection. Reviews of existing qualitative studies may also provide evidence to support hypothesised mechanisms. To contribute to future evaluations, archiving new primary data is good practice where possible in appropriate repositories such as those offered by many university libraries or, if appropriate, the UK Data Service (https://ukdataservice.ac.uk). These datasets can be used for comparisons in future follow-ups, or for comparative studies.

Reporting, critical appraisal and evidence synthesis

Summary

- Clearly defining the purpose of a given synthesis is essential 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 study designs.
- While diversity in designs and outcomes 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.

This section signposts readers to guidance for reporting different stages and uses of evaluations and discusses issues specific to identifying and appraising evidence from natural experimental evaluations and synthesising such evidence across studies. Using and combining evidence, within a single evaluation, are discussed in *Design and planning: concepts*.

Reporting guidelines

Accurate, clear and comprehensive reporting of natural experimental evaluations will facilitate critical appraisal and evidence synthesis and encourage best use of evidence for decision-making. *Table 7* provides details of guidance likely to be useful to researchers conducting natural experimental evaluations.

Bringing together evidence from a set of natural experimental evaluations requires consideration of issues throughout the critical appraisal and evidence synthesis process. Defining the overarching research question(s) for a given evidence synthesis, that is why a given set of studies is being synthesised, will inform decisions on how best to do it.[26] Each element of the scope of a conventional effectiveness synthesis, that is the populations, interventions, comparators and outcomes, may be difficult to define, for example because of a lack of standardisation of intervention or exposure and control conditions, multiplicity of outcomes in various formats, or varying styles and standards of reporting across the studies. Assessing effectiveness in a narrow sense – for example by producing a pooled effect size estimate using meta-analysis[165] – may not be the most useful question, or even an answerable one for a given type of intervention or a given set of evidence. For some review topics it may be more valuable to examine effectiveness in a broader sense,[166] to investigate equity impacts rather than overall effectiveness,[165] or to establish causality, describe people's lived experiences, or explore mechanisms of action in respect of particular interventions.[165] 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.[167,168] As with a primary natural experimental study, any evidence synthesis should have a protocol to demonstrate that the results are not selectively chosen. With a synthesis of natural experimental evaluations, 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 (see *Design and planning: practicalities*).

TABLE 7 Reporting guidance

Reporting guidance	Focus
SPIRIT 2013: Standard Protocol Items: Recommendations for Interventional Trials[60]	This guidance for reporting clinical trial protocols may be useful for identifying key features of the natural experimental evaluation to include in a protocol. A reporting guideline is in development for protocols for observational studies (SPIROS)[59]
STROBE: Strengthening the Reporting of Observational Studies in Epidemiology[156]	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[157]	Checklist for reporting non-randomised behavioural and public health intervention evaluations
RECORD: REporting of studies Conducted using Observational Routinely-collected health Data[158]	An extension to STROBE, 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[62]	An adaptation of the TIDieR guidance,[159] 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)[160]	Guidance and 28 item checklist for reporting economic evaluations of health interventions
SRQR: Standards for Reporting Qualitative Research[161]	Guidance for reporting qualitative research that is intended to be relevant across differing paradigms and methods
Triple C (Case study, Context, Complex interventions) reporting principles[126]	Reporting principles for case study evaluations of the role of context in complex interventions
PRISMA-P: Preferred Reporting Items for Systematic reviews and Meta-Analyses for Protocols 2015[162]	Guidance for reporting systematic review protocols
PRISMA 2020[163]	Guideline and checklist to promote complete reporting of systematic reviews
SWiM: Synthesis without Meta-analysis[164]	Guidance for reporting synthesis without meta-analysis in systematic reviews

Identifying and managing natural experimental evaluation 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. Formalised bibliographic search using Medical Subject Heading terms or familiar text strings to filter a literature search by study design works less well for natural experimental studies than for trials. 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.[169]

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, for example epidemiology and economics. In some disciplines, for example 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 natural experimental evaluations may be further complicated by the multifaceted nature of many of these studies, for example the reporting of multiple analyses or mixed methods from a single study. Defining relevant time periods may be difficult as evaluations are often examining a natural experiment that occurred at a time before, possibly decades prior to, the evaluation. These issues have implications for the literature search strategy, achieving a balance between the sensitivity and the specificity of the search. It may be helpful to involve an information scientist and a range of topic experts from different disciplines in formulating a suitable search strategy.

A particular challenge with data extraction relates to the diversity of studies included in many syntheses. It may be difficult to determine which effect estimate to extract from the multiplicity reported in a given study.[170] Details

of the natural experiments and important information about the context need to be captured uniformly to enable synthesis.[62,147]

Critical appraisal of natural experimental evaluations

Critical appraisal may be required to understand the rigour of an individual study or undertaken as part of a synthesis of evidence. When critically appraising a natural experimental evaluation, reviewers may wish to use tools for assessing risk of bias. It should be noted that these tools typically assume a focus on a minimally biased primary effect size estimate as the main objective of each study, whereas many natural experimental evaluations entail more complex study designs often including qualitative and mixed-method analyses. With that caveat, the Risk Of Bias In Non-Randomised Studies - of Interventions (ROBINS-I)[17] is the most thorough of the quantitative critical appraisal tools currently available. Applying it requires expertise and time to learn the process,[171,172] and some features of natural experimental studies can be problematic to fit into this type of appraisal process. When using ROBINS-I, few evaluations of natural experiments 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. When a broad definition of natural experiment is used (as we advocate in this guidance, see *Concepts and definitions*), a wide variety of study designs may be included in the review resulting in a wide range of internal validity across the included studies.

Where studies are multifaceted, tools such as ROBINS-I may be difficult to apply.[15,173] While ROBINS-I may be the most obvious off-the-shelf tool to use, 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 natural experimental evaluations, such as regression discontinuity or instrumental variables designs.[174] In some cases the alternative ROBINS-E, which is designed to assess the risk of bias in observational epidemiological studies, may be more appropriate.[175] Other tools that may be useful include the Effective Public Health Practice Project quality assessment tool,[176] and the Cochrane Risk of Bias 2 with Cochrane Effective Practice and Organisation of Care guidance notes.[177] For some types of natural experimental evaluations, and some systematic reviews, an appraisal framework more like those used for qualitative research may be more appropriate.[178,179]

Synthesising results from natural experimental evaluations

Synthesising evaluations of natural experiments contributes to making best use of available evidence. The likely diversity in these studies provides opportunities for further 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 designs, diversity of the content and context of the interventions, and the range of outcomes studied. Some studies may report multiple effect estimates for the same outcome, multiple outcomes, or both. In some cases it may be possible to derive standardised outcome measures suitable for inclusion in a meta-analysis, but standardisation is not always possible or meaningful if, for example, there is a potential trade-off between competing good outcomes (e.g. more cycling at the expense of walking in transport studies).[180] The included studies may report 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.*[155]). Investigating health inequity requires study data to be amenable to grouping to make appropriate comparisons, which should be clearly reported.[181]

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 (such as those drawn from large-scale administrative datasets) 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, for example by comparing meta-analysis findings with syntheses not using meta-analysis of other studies of different designs. Furthermore, the requirement for methodological homogeneity among studies included in a meta-analysis may raise

questions about the external validity of the pooled result, if an evaluative bias may have been introduced by excluding certain types of intervention or context from analysis solely because dissimilar study designs or outcomes were used.[10]

Approaches to synthesis that do not use meta-analysis, such as guidance provided by Cochrane[182,183] and the Realist and Meta-narrative Evidence Syntheses: Evolving Standards group,[184] may often be useful when synthesising natural experimental evaluations. The specifics of the evidence synthesis questions and studies involved will determine the appropriate methods to use, with a mixed-methods design often useful.[167,168] On some occasions, the focus of the review may be to combine effect estimates with simulation models[19] or to conduct a synthesis of economic evidence.[185]

Exploring heterogeneity between the natural experimental evaluation results can result in valuable insights from the synthesis, developing potential hypotheses about similarities and differences in mechanisms and study characteristics between the evaluations. The exploration of heterogeneity may be through qualitative methods, triangulation methods, or quantitative methods when these are statistically feasible and substantively meaningful.[186]

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.[26] When assessing certainty of the evidence is appropriate, the framework most often used is Grading of Recommendations, Assessment, Development and Evaluations (GRADE).[187] Applying this tool with natural experimental evaluations raises some issues similar to those noted above when assessing risk of bias. These include ensuring that evidence from evaluations of natural experiments is appropriately acknowledged in GRADE by correctly recognising the susceptibility of different study designs to bias, incorporating a variety of differing study designs, and selecting outcomes for synthesis.[16,18] The lack of scale available within GRADE can make it difficult to differentiate between certainty assessments.[16]

Infrastructure and information governance

Summary

- Natural experimental evaluations often use data that were 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 environments (TRE) model of curation and provision 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,[188] population surveys,[39] administrative datasets,[189] sales and purchasing data[45] and, increasingly, information gathered via mobile phones, fitness apps, and other forms of 'crowd-sourced' data.[190,191] 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
(*Boxes 5–7*), but outside these settings, acquisition of data can be complicated, uncertain, and time-consuming. The risks involved may deter researchers from attempting studies that require access data that has not previously been used for research, especially where novel linkages between datasets are involved. Funding bodies can therefore play an important role in facilitating access to routinely collected datasets, investing in infrastructure and in funding the curation of strategically important datasets.

BOX 5 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 had the training within the past 2 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 6 New Zealand's Integrated Data Infrastructure (IDI)

The IDI (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.

BOX 7 Brazil's Centre for Data and Knowledge Integration for Health (CIDACS)

CIDACS, the Centre for Data and Knowledge Integration for Health (https://cidacs.bahia.fiocruz.br/en/), was established to examine health outcomes of social policies such as cash transfer, sanitation, and housing programmes. CIDACS has an infrastructure to store, process, and link identified data and a facility for producing de-identified datasets.

The Unified Registry for Social Programmes (CadUnico) provides a population spine, to which has been added data on deaths, births, and infectious diseases for the Brazilian population. CadUnico contains socioeconomic and demographic information on more than 114 million people who have applied for, or are eligible for support from, over 20 government social benefit programmes.

Access to deidentified or anonymised CIDACS data is provided for research projects conducted in collaboration with CIDACS or to authorised staff from government agencies. To obtain access to CIDACS datasets, the researcher must obtain approval from a relevant Brazilian institutional research ethical committee, provide a clear analysis plan, comply with rules of access and access relevant data via a high-level security system, using an authorised virtual private network (VPN) with two-factor authentication.

CadUnico, Consultar dados do Cadastro Único (Unified Registry for Social Programmes, Brazil).

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 markedly increase the time and effort needed to obtain a research-ready dataset, if research culture, information governance procedures, accreditation requirements 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. Funding organisations can contribute to addressing these issues by funding data infrastructure and capacity building, and by negotiating with data owners to make routinely collected data available for natural experimental evaluations of policies and programmes.

Trusted research environments

Secure Anonymised Information Linkage, the Integrated Data Infrastructure, and CIDACs (see *Boxes 5–7*) are examples of 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 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.[192]

The TRE model has several 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 and creation of any derived variables.

Good practice considerations

The framework was developed in consultation with international experts and users of natural experimental evaluations and aims to raise awareness of issues to consider when planning an evaluation or using the results of an evaluation to influence population health decision-making. The researchers hope the framework will promote the conduct and use of evidence from methodologically robust evaluations of policies and other interventions, as they are implemented. With the large number of topics covered by the framework, the researchers were unable to cover each in comprehensive detail. Instead, they aimed to convey key information with signposting to more detailed information provided throughout.

The researchers conclude by summarising their main messages for planning, commissioning, conducting, reporting and using evidence from natural experimental evaluations. The researchers 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. The researchers have concentrated on messages that are practical and implementable. There are already many instances of good practice, from studies that take a systems approach to the evaluation of natural experiments to funders' mandates for open science practices and investment in capacity building and data infrastructures that facilitates natural experimental evaluations. Building on these will support the continued growth of high-quality natural experimental evaluations that produce useful evidence for population health decision-making.

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 most appropriate evaluation approach and assess the feasibility of the evaluation.

- Consider the variety and importance of stakeholders: for natural experimental evaluations, there may be stakeholders in the natural experiment (policy or event) who differ from those with a stake in its evaluation; 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 analytical approaches, incorporating perspectives from diverse disciplines, such as economics, epidemiology and the social and political sciences, 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, and no simple hierarchy of stronger and weaker methods.
- Consider the appropriateness of adopting a systems science approach to evaluating natural experiments as events within complex systems, as interactions between interventions and the contexts in which they occur may lead to a range of intended and unintended consequences.
- Natural experimental evaluations will usually be stronger if they use a combination of methods, including alternative methods of effect estimation, robustness checking and a mixture 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 its planning, protocol, analyses, and results, using established reporting standards where available (see *Table 7*), ensuring key details are in plain language appropriate for the evidence users. Report the results of all the planned analyses or explain why any of those indicated in the protocol were not progressed.
- Consider evaluation from a health equity perspective. Natural experiments may have differential effects across social groups and therefore impact on health inequalities across multiple and intersecting dimensions (such as socioeconomic position, gender, age, disability, ethnicity, etc.), either through the differing reach of interventions or differences in effectiveness.
- Evaluation of the strength of evidence from natural experimental evaluations should be based on detailed appraisal of the strengths and limitations of the studies and the applicability of their findings, including assessment of risk of bias where appropriate, 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 or methods-appropriate study plan is available prior to analysis commencing, findings are published in open access journals, and the relevant reporting guidelines are followed.
- Establish processes within funding bodies to facilitate flexible and timely responses to prospective natural experimental evaluation opportunities.
- As well as funding individual studies, support capacity building for natural experiments through investment in infrastructure (e.g. TREs) and the workforce (e.g. training in evaluation methods, data science and research software engineering).
- Negotiate with data owners to make routinely collected data both available and linkable to other datasets for natural experimental evaluations of policies and programmes.
- When commissioning natural experimental evaluations, be prepared to be flexible and pragmatic and accept that both the evaluability of the natural experiment and the feasibility of the evaluation require assessment, which may result in the full evaluation not being viable.

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

Provide guidance to authors and reviewers on requirements for reports of natural experimental evaluations (e.g. require a study protocol, use of appropriate reporting guidelines).

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.

Additional information

CRediT contribution statement

Peter Craig (https://orcid.org/0000-0002-7653-5832): Conceptualisation, Methodology, Investigation, Writing – original draft, Project administration, Funding acquisition.

Mhairi Campbell (https://orcid.org/0000-0002-4416-7270): Methodology, Investigation, Writing – original draft, Project administration.

Manuela Deidda (https://orcid.org/0000-0002-0921-6970): Methodology, Investigation, Writing – original draft.

Ruth Dundas (https://orcid.org/0000-0002-3836-4286): Methodology, Investigation, Writing – original draft.

Judith Green (https://orcid.org/0000-0002-2315-5326): Methodology, Investigation, Writing – original draft.

Srinivasa Vittal Katikireddi (https://orcid.org/0000-0001-6593-9092): Methodology, Investigation, Writing – original draft.

Jim Lewsey (https://orcid.org/0000-0002-3811-8165): Methodology, Investigation, Writing – original draft.

David Ogilvie (https://orcid.org/0000-0002-0270-4672): Methodology, Investigation, Writing – original draft.

Frank de Vocht (https://orcid.org/0000-0003-3631-627X): Methodology (co-lead), Investigation, Writing – original draft.

Martin White (https://orcid.org/0000-0001-7700-2352): Conceptualisation, Methodology, Investigation, Writing – original draft.

Acknowledgements

The researchers thank the participants of the workshops and online consultation for their contributions (see *Appendix 1*). The researchers acknowledge the support and insight of the advisory group and the oversight group.

Advisory group: Adrian Bauman (University of Sydney), Kate Tilling (University of Bristol), Marc Suhrcke (Luxembourg Institute of Social and Economic Research), Niamh Fitzgerald (University of Stirling), Sara Shaw (University of Oxford), Audrey Ceschia (Editor Lancet Public Health), Scott Lloyd (Middlesbrough Council and Associate Lead for Public Health Research, NIHR Clinical Research Network North East & North Cumbria), Sarah Sharples (CSA Dept for Transport).

Oversight group: Graham Hart (oversight group Chair, MRC PHIND panel), Peymané Adab (MRC-NIHR Methodology Advisory Group), Claire Kidgell (Assistant Director NIHR), Catherine Moody (Head of Population Health UKRI MRC), Tamsyn Derrick (Programme Manager Population Health Sciences and PHIND UKRI MRC).

The researchers thank colleagues at the MRC/CSO Social and Public Health Sciences Unit, including Lorna Dick for administrative support throughout the project, Enni Miller for communications support including website development, and Adam Gilinsky for developing the online consultation platform.

Data-sharing statement

De-identified content from the workshops and consultation are archived with the UK Data Service's ReShare repository (https://dx.doi.org/10.5255/UKDA-SN-857343) and are available on application via https://reshare.ukdataservice.ac.uk/. All other queries should be addressed to the corresponding author.

Ethics statement

Ethical approval for the workshops and online consultation was obtained in May 2022 from the University of Glasgow College of Social Sciences Ethics Committee (Reference number 400210210).

Information governance statement

The University of Glasgow is committed to handling all personal information according to the Data Protection Act (2018) and the General Data Protection Regulation (EU GDPR) 2016/679. Under the Data Protection legislation, the University of Glasgow is the Data Controller, and you can find out more about how we handle personal data, including how to exercise your individual rights and the contact details for our Data Protection Officer here www.gla.ac.uk/myglasgow/dpfoioffice/.

Disclosure of interests

Full disclosure of interests: Completed ICMJE forms for all authors, including all related interests, are available in the toolkit on the NIHR Journals Library report publication page at https://doi.org/10.3310/JTYW6582.

Primary conflicts of interest: Judith Green was a member of the NIHR Public Health Research (PHR) Funding Board (2017–23) the O'Brien Institute for Public Health, Calgary, Canada, International SAB, is a member of the NIHR PHR Behavioural Science Scientific Advisory Board and is supported by Wellcome Trust (Centre Grant 203109/Z/16/Z). Srinivasa Vittal Katikireddi and Jim Lewsey are members of the NIHR PHR Funding Board (2020–5). Srinivasa Vittal Katikireddi was a member of the NIHR Long COVID funding board (2021–2) and the NIHR Policy Research Units funding board (2023). Peter Craig, Mhairi Campbell, Ruth Dundas and Srinivasa Vittal Katikireddi are supported by the Medical Research Council (MC_UU_00022/2) and the Scottish Government Chief Scientist Office (SPHSU17). Srinivasa Vittal Katikireddi is also supported by the European Research Council (949582). Frank de Vocht is a member of the NIHR PHR Funding Board (2017–28), the Swedish Research Council Public Health Funding Committee, the UNSCEAR Expert Group on the evaluation of diseases of the circulatory system from radiation exposure (CircuDis), and a committee member of the Health Council of the Netherlands Future Hazards. Ruth Dundas is a member of the Wellcome Trust Population Health Advisory Board (2022–5), the NIHR Population Health Career Scientist Committee, is supported by UK Prevention Research Partnership (MR/ S037608/1) and the NIHR Advisory Board for Evaluation and co-creation to optimise use and benefits of the Healthy Start Scheme. David Ogilvie and Martin White are supported by the Medical Research Council (Unit programme MC_UU_00006/7). Martin White was a member of the NIHR PHR research funding panel (2009–20), Director of the PHR programme and chaired the Prioritisation Board (2014–20), member of the NIHR Strategy Board (2014–20), member of the MRC Population Health Sciences Strategy Group (PHSG) (2014–20), member of the MRC Public Health Intervention Development (PHIND) panel (2014–8).

Publications

Craig P, Campbell M, Bauman A, Deidda M, Dundas R, Fitzgerald N, *et al.* Making better use of natural experimental evaluation in population health. *BMJ*. 2022;**379**:e070872.

Craig P, Campbell M, Deidda M, Dundas R, Green G, Katikireddi SV, *et al.* Using natural experiments to evaluate population health and health system interventions: new framework for producers and users of evidence. *BMJ* 2025;**388**:e080505. https://doi.org/10.1136/bmj-2024-080505

Conference papers, seminars, workshops

Craig P. Natural experimental evaluations – new guidance. Interactive seminar, School of Public Health Research (SPHR) Network for the use of Natural Experiments in Public Health, online, September 2022.

Craig P. Natural experimental evaluations – new guidance. Methods masterclass, Maternal and Child Health Network (MatCHNet) online, November 2022.

Craig P, Ogilvie D, Campbell M, Dundas R, Katikireddi SV. Natural experimental evaluations for public policy and health systems: recent advances and updated guidance. Pre-conference workshop 15th European Public Health Conference 2022, Berlin, Germany, November 2022.

Green J, Shaw S, Paparini S. Medical imperialism, sociological imperialism? The politics of reporting guidelines BSA Medical Sociology Conference, Lancaster, September 2022.

Campbell M, Craig P. Updated guidance on using natural experiments to evaluate population health Interventions. UK PRP Prevention Research Conference, Edinburgh, UK, November 2023.

Craig P, Campbell M, Dundas R, Katikireddi SV. Using natural experiments to evaluate population health interventions. Workshop, Society for Social Medicine and Population Health 67th Annual Scientific Meeting, Newcastle, UK, September 2023.

Craig P, Green J, Dundas R. Guidance on using natural experiments to evaluate population health interventions. Skills building seminar, 16th European Public Health Conference 2023, Dublin, Ireland, November 2023.

de Vocht F. Methods to Evaluate Public Health Interventions. Erasmus University Summer Programme NIHES 2023, Rotterdam, Netherlands, August 2023.

Ogilvie D. Using natural experiments to evaluate population health interventions: updated and extended guidance for producers and users of evidence. Poster presentation, 22nd International Society of Behavioral Nutrition and Physical Activity (ISBNPA) Annual Meeting, Uppsala, Sweden, June 2023.

Patient and public involvement

The project was methodological. Throughout the course of the project, there was engagement with appropriate audiences such as relevant policy-makers, practitioners and local and national government representatives. The workshops and consultation exercise included input from evidence users as well as researchers.

Equality, diversity and inclusion

As a methodological project the researchers' review did not include any direct research participants. They were inclusive when identifying potential respondents for the expert workshops and online consultation. The researchers selected the advisory group to be as inclusive as possible. Workshops were held at times convenient for international participants and the writing group, and recruitment of participants for the advisory group, workshops and consultation aimed to be inclusive of gender, ethnicity and country.

Impact and learning

Plans for dissemination of the framework have been designed with the aim of reaching various audiences through a variety of formats. This includes conferences for both academics and practitioners, journal papers, online events and a dedicated website which collates these resources. Interest in the forthcoming framework has been demonstrated by invitations to present at Public Health Wales and facilitate a virtual methods workshop for early career researchers, public health trainees, and analysts for the Maternal and Child Health Network (MatCHNeT – funded by the UK Prevention Research Partnership).

References

1. 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. *J Epidemiol Community Health* 2012;**66**:1182–6. https://doi.org/10.1136/jech-2011-200375
2. The Committee for the Prize in Economic Sciences in Memory of Alfred Nobel. *Answering Causal Questions Using Observational Data*. Stockholm: Royal Swedish Academy of Sciences; 2021.
3. Snell A, Reeves A, Rieger M, Galea G, Mauer-Stender K, Mikkelsen B, Stuckler D. WHO Regional Office for Europe's natural Experiment Studies Project: an introduction to the series. *Eur J Public Health* 2018;**28**:1–3. https://doi.org/10.1093/eurpub/cky195
4. National Institutes of Health Office of Disease Prevention. *A Report from the Federal Partners Meeting of the National Institutes of Health Pathways to Prevention Workshop: Methods for Evaluating Natural Experiments in Obesity*. National Institutes of Health; 2018. URL: https://prevention.nih.gov/sites/default/files/2019-01/ObesityMethodsP2PFederalPartnersMeetingReport.pdf (accessed 11 December 2021).
5. European Federation of Academies of Sciences and Humanities (ALLEA), Federation of European Academies of Medicine (FEAM). *Health Inequalities Research. New Methods, Better Insights?* Royal Netherlands Academy of Arts and Sciences (KNAW). URL: https://allea.org/wp-content/uploads/2021/11/Health_Inequalities.pdf (accessed 11 December 2021).
6. Bärnighausen T, Tugwell P, Røttingen J-A, Shemilt I, Rockers P, Geldsetzer P, *et al.* Quasi-experimental study designs series – paper 4: uses and value. *J Clin Epidemiol* 2017;**89**:21–9. https://doi.org/10.1016/j.jclinepi.2017.03.012
7. Basu S, Meghani A, Siddiqi A. Evaluating the health impact of large-scale public policy changes: classical and novel approaches. *Annu Rev Public Health* 2017;**38**:351–70. https://doi.org/10.1146/annurev-publhealth-031816-044208
8. Craig P, Katikireddi SV, Leyland A, Popham F. Natural experiments: an overview of methods, approaches, and contributions to public health intervention research. *Annu Rev Public Health* 2017;**38**:39–56. https://doi.org/10.1146/annurev-publhealth-031816-044327
9. Tugwell P, Knottnerus JA, McGowan J, Tricco A. Big-5 quasi-experimental designs. *J Clin Epidemiol* 2017;**89**:1–3.
10. 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. https://doi.org/10.1136/bmj.n2061
11. Craig P, Campbell M, Deidda M, Dundas R, Green G, Katikireddi SV, *et al.* Using natural experiments to evaluate population health and health system interventions: new framework for producers and users of evidence. *BMJ* 2025;**388**:e080505. https://doi.org/10.1136/bmj-2024-080505
12. Craig P, Campbell M, Bauman A, Deidda M, Dundas R, Fitzgerald N, *et al.* Making better use of natural experimental evaluation in population health. *BMJ* 2022;**379**:e070872.
13. de Vocht F, Katikireddi SV, McQuire C, Tilling K, Hickman M, Craig P. Conceptualising natural and quasi experiments in public health. *BMC Med Res Methodol* 2021;**21**:32.
14. Dunning T. *Natural Experiments in the Social Sciences: A Design-Based Approach*. Cambridge, MA: Cambridge University Press; 2012.
15. 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. *Am J Public Health* 2022;**112**:548–52. https://doi.org/10.2105/AJPH.2021.306658

16. Hilton Boon M, Thomson H, Shaw B, Akl EA, Lhachimi SK, López-Alcalde J, *et al.*; GRADE Working Group. Challenges in applying the GRADE approach in public health guidelines and systematic reviews: a concept article from the GRADE Public Health Group. *J Clin Epidemiol* 2021;**135**:42–53. https://doi.org/10.1016/j.jclinepi.2021.01.001

17. 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. https://doi.org/10.1136/bmj.i4919

18. 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. *J Clin Epidemiol* 2022;**142**:200–8. https://doi.org/10.1016/j.jclinepi.2021.11.026

19. Brozek JL, Canelo-Aybar C, Akl EA, Bowen JM, Bucher J, Chiu WA, *et al.*; GRADE Working Group. GRADE Guidelines 30: the GRADE approach to assessing the certainty of modeled evidence – an overview in the context of health decision-making. *J Clin Epidemiol* 2021;**129**:138–50. https://doi.org/10.1016/j.jclinepi.2020.09.018

20. Petticrew M, McKee M, Lock K, Green J, Phillips G. In search of social equipoise. *BMJ* 2013;**347**:f4016.

21. Hamad R, Elser H, Tran DC, Rehkopf DH, Goodman SN. How and why studies disagree about the effects of education on health: a systematic review and meta-analysis of studies of compulsory schooling laws. *Soc Sci Med* 2018;**212**:168–78. https://doi.org/10.1016/j.socscimed.2018.07.016

22. Avendano M, de Coulon A, Nafilyan V. Does longer compulsory schooling affect mental health? Evidence from a British reform. *J Public Econ* 2020;**183**:104137. https://doi.org/10.1016/j.jpubeco.2020.104137

23. Grytten J, Skau I, Sørensen R. Who dies early? Education, mortality and causes of death in Norway. *Soc Sci Med* 2020;**245**:112601. https://doi.org/10.1016/j.socscimed.2019.112601

24. Meghir C, Palme M, Simeonova E. Education and mortality: evidence from a social experiment. *Am Econ J: Appl Econ* 2018;**10**:234–56. https://doi.org/10.1257/app.20150365

25. Davies NM, Dickson M, Davey Smith G, Windmeijer F, van den Berg GJ. *The Causal Effects of Education on Adult Health, Mortality and Income: Evidence from Mendelian Randomization and the Raising of the School Leaving Age*. IZA discussion paper no. 12192. Bonn: IZA – Institute of Labor Economics; 2019. https://doi.org/10.2139/ssrn.3390179

26. 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. *J Epidemiol Community Health* 2020;**74**:203–8. https://doi.org/10.1136/jech-2019-213085

27. 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 Intern Med* 2017;**177**:44–50. https://doi.org/10.1001/jamainternmed.2016.6811

28. Baxter AJ, Dundas R, Popham F, Craig P. How effective was England's teenage pregnancy strategy? A comparative analysis of high-income countries. *Soc Sci Med* 2021;**270**:113685.

29. 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**:994–1003. https://doi.org/10.1111/add.13310

30. 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. *Lancet Publ Health* 2018;**3**:e333–40.

31. Codreanu MA, Waters T. *Do Work Search Requirements Work? Evidence from a UK Reform Targeting Single Parents*. London: Institute for Fiscal Studies; 2023.

32. 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. *Lancet Publ Health* 2020;**5**:e157–64.

33. Brewer M, Dang T, Tominey E. *Universal Credit: Welfare Reform and Mental Health*. Bonn: IZA Institute of Labor Economics; 2022.
34. Angrist JD. Lifetime earnings and the Vietnam era draft lottery: evidence from social security administrative records. *Am Econ Rev* 1990;**80**:313–36.
35. Nguyen QC, Rehkopf DH, Schmidt NM, Osypuk TL. Heterogeneous effects of housing vouchers on the mental health of US adolescents. *Am J Public Health* 2016;**106**:755–62.
36. Fetzer T, Graeber T. Measuring the scientific effectiveness of contact tracing: evidence from a natural experiment. *Proc Natl Acad Sci USA* 2021;**118**:e2100814118. https://doi.org/10.1073/pnas.2100814118
37. Fetzer T. *Measuring the Epidemiological Impact of a False Negative: Evidence from a Natural Experiment*. Warwick: Department of Economics, University of Warwick; 2021.
38. Hainmueller J, Lawrence D, Martén L, Black B, Figueroa L, Hotard M, *et al.* Protecting unauthorized immigrant mothers improves their children's mental health. *Science* 2017;**357**:1041–4.
39. Goodman A, van Sluijs EMF, Ogilvie D. Impact of offering cycle training in schools upon cycling behaviour: a natural experimental study. *Int J Behav Nutr Phys Activ* 2016;**13**:34. https://doi.org/10.1186/s12966-016-0356-z
40. Kromydas T, Campbell M, Chambers S, Boon MH, Pearce A, Wells V, Craig P. The effect of school summer holidays on inequalities in children and young people's mental health and cognitive ability in the UK using data from the millennium cohort study. *BMC Public Health* 2022;**22**:154. https://doi.org/10.1186/s12889-022-12540-2
41. 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? *Prev Med* 2018;**108**:17–22. https://doi.org/10.1016/j.ypmed.2017.12.025
42. Rutter H, Savona N, Glonti K, Bibby J, Cummins S, Finegood DT, *et al.* The need for a complex systems model of evidence for public health. *Lancet* 2017;**390**:2602–4. https://doi.org/10.1016/S0140-6736(17)31267-9
43. Creswell JW, Clark VLP. *Designing and Conducting Mixed Methods Research*. London: SAGE Publications Ltd; 2011.
44. Fetters MD, Curry LA, Creswell JW. Achieving integration in mixed methods designs: principles and practices. *Health Serv Res* 2013;**48**:2134–56. https://doi.org/10.1111/1475-6773.12117
45. 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**:2697–707. https://doi.org/10.1111/add.15478
46. Green J, Roberts H, Petticrew M, Steinbach R, Goodman A, Jones A, Edwards P. Integrating quasi-experimental and inductive designs in evaluation: a case study of the impact of free bus travel on public health. *Evaluation* 2015;**21**:391–406. https://doi.org/10.1177/1356389015605205
47. 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**:E881–7.
48. 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 Pol* 2021;**102**:102104. https://doi.org/10.1016/j.foodpol.2021.102104
49. Bazeley P. *Integrating Analyses in Mixed Methods Research*. London. 2018. URL: https://methods.sagepub.com/book/integrating-analyses-in-mixed-methods-research (accessed 4 August 2023).
50. 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'. *Public Health* 2022;**211**:1–4.

51. Bennett A, Checkel JT, editors. *Process Tracing: From Metaphor to Analytic Tool*. Cambridge, MA: Cambridge University Press; 2015.

52. 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**:877. https://doi.org/10.1186/s12889-021-10926-2

53. Beach D, Pedersen RB. *Process-Tracing Methods: Foundations and Guidelines*. Ann Arbor, MI: University of Michigan Press; 2019.

54. Byrne D. Evaluating complex social interventions in a complex world. *Evaluation* 2013;**19**:217–28.

55. Leviton LC, Khan LK, Rog D, Dawkins N, Cotton D. Evaluability assessment to improve public health policies, programs, and practices. *Annu Rev Public Health* 2010;**31**:213–33.

56. 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. *Milbank Q* 2011;**89**:206–25.

57. Summerskill W, Collingridge D, Frankish H. Protocols, probity, and publication. *Lancet* 2009;**373**:992.

58. Marsden J, Cousijn J, Stapleton J. Pre-registration: not a daunting practice. *Addiction* 2022;**117**:845–6. https://doi.org/10.1111/add.15819

59. 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 Res Protocol* 2020;**9**:e17864.

60. 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 (Clin Res Ed.)* 2013;**346**:e7586. https://doi.org/10.1136/bmj.e7586

61. Baldwin JR, Pingault J-B, Schoeler T, Sallis HM, Munafò MR. Protecting against researcher bias in secondary data analysis: challenges and potential solutions. *Eur J Epidemiol* 2022;**37**:1–10.

62. 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**:k1079.

63. Cheetham M, Atkinson P, Gibson M, Katikireddi S, Moffatt S, Morris S, *et al.* Exploring the mental health effects of Universal Credit: a journey of co-production. *Persp Publ Health* 2022;**142**:209–12.

64. Cullerton K, White M. Understanding and Managing Corporate Conflicts of Interest. In: Maani N, Petticrew M, Galea S, editors. *The Commercial Determinants of Health*. New York: Oxford University Press; 2022. pp. 307–18. https://doi.org/10.1093/oso/9780197578742.003.0030

65. Hernán MA, Robins JM. Using big data to emulate a target trial when a randomized trial is not available. *Am J Epidemiol* 2016;**183**:758–64. https://doi.org/10.1093/aje/kwv254

66. Remiro-Azócar A, Heath A, Baio G. Conflating marginal and conditional treatment effects: comments on 'Assessing the performance of population adjustment methods for anchored indirect comparisons – a simulation study'. *Stat Med* 2021;**40**:2753–8.

67. Bazo-Alvarez JC, Morris TP, Pham TM, Carpenter JR, Petersen I. Handling missing values in interrupted time series analysis of longitudinal individual-level data. *Clin Epidemiol* 2020;**12**:1045–57. https://doi.org/10.2147/clep.S266428

68. Lopez Bernal J, Cummins S, Gasparrini A. The use of controls in interrupted time series studies of public health interventions. *Int J Epidemiol* 2018;**47**:2082–93. https://doi.org/10.1093/ije/dyy135

69. Lousdal ML. An introduction to instrumental variable assumptions, validation and estimation. *Emerg Themes Epidemiol* 2018;**15**:1. https://doi.org/10.1186/s12982-018-0069-7

70. Zhang ML, Adepeju M, Thomas R. Estimating the effects of crime maps on house prices using an (un) natural experiment: a study protocol. *PLOS ONE* 2022;**17**:e0278463.

71. Brodersen KH, Gallusser F, Koehler J, Remy N, Scott SL. Inferring causal impact using Bayesian structural time-series models. *Ann Appl Stat* 2015;**9**:247–74.
72. 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**:2021–31.
73. Zhou Z-H. *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: CRC Press; 2012.
74. Blakely T, Lynch J, Simons K, Bentley R, Rose S. Reflection on modern methods: when worlds collide: prediction, machine learning and causal inference. *Int J Epidemiol* 2019;**49**:2058–64. https://doi.org/10.1093/ije/dyz132
75. Bentley R, Baker E, Simons K, Simpson JA, Blakely T. The impact of social housing on mental health: longitudinal analyses using marginal structural models and machine learning-generated weights. *Int J Epidemiol* 2018;**47**:1414–22. https://doi.org/10.1093/ije/dyy116
76. Imai K, Keele L, Tingley D. A general approach to causal mediation analysis. *Psychol Methods* 2010;**15**:309–34. https://doi.org/10.1037/a0020761
77. Moore H, O'Malley C, Lloyd S, Eskandari F, Rose K, Butler M, *et al. 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*. Paper presented at: European and International Congress on Obesity, Dublin, August, 2020.
78. Kraska M. Repeated Measures Design. In: Salkind N, editor. *Encyclopedia of Research Design*. London: SAGE Publications Ltd; 2012. pp. 1244–7. https://doi.org/10.4135/9781412961288
79. Mozes M, van der Vegt I, Kleinberg B. A repeated-measures study on emotional responses after a year in the pandemic. *Sci Rep* 2021;**11**:23114.
80. Fergusson DM, Horwood LJ, Boden JM, Mulder RT. Impact of a major disaster on the mental health of a well-studied cohort. *JAMA Psychiat* 2014;**71**:1025–31. https://doi.org/10.1001/jamapsychiatry.2014.652
81. Sedgwick P. Before and after study designs. *BMJ* 2014;**349**:g5074. https://doi.org/10.1136/bmj.g5074
82. Oldenburg CE, Moscoe E, Bärnighausen T. Regression discontinuity for causal effect estimation in epidemiology. *Curr Epidemiol Rep* 2016;**3**:233–41. https://doi.org/10.1007/s40471-016-0080-x
83. Keele LJ, Titiunik R. Geographic boundaries as regression discontinuities. *Political Analysis* 2015;**23**:127–55. https://doi.org/10.1093/pan/mpu014
84. Behrman JA. The effect of increased primary schooling on adult women's HIV status in Malawi and Uganda: universal primary education as a natural experiment. *Soc Sci Med* 2015;**127**:108–15. https://doi.org/10.1016/j.socscimed.2014.06.034
85. Lee DS, Lemieux T. Regression discontinuity designs in economics. *J Econ Lit* 2010;**48**:281–355. https://doi.org/10.1257/jel.48.2.281
86. Hanandita W, Tampubolon G. Does poverty reduce mental health? An instrumental variable analysis. *Soc Sci Med* 2014;**113**:59–67. https://doi.org/10.1016/j.socscimed.2014.05.005
87. Mellon J. *Rain, Rain, Go Away: 192 Potential Exclusion-Restriction Violations for Studies Using Weather as an Instrumental Variable*. SSRN 3715610. Rochester, NY: SSRN; 2022.
88. Ikenwilo D. A difference-in-differences analysis of the effect of free dental check-ups in Scotland. *Soc Sci Med* 2013;**83**:10–8. https://doi.org/10.1016/j.socscimed.2013.01.027
89. Dimitrovová K, Perelman J, Serrano-Alarcón M. Effect of a national primary care reform on avoidable hospital admissions (2000–2015): a difference-in-difference analysis. *Soc Sci Med* 2020;**252**:112908. https://doi.org/10.1016/j.socscimed.2020.112908

90. Thayer WM, Hasan MZ, Sankhla P, Gupta S. An interrupted time series analysis of the lockdown policies in India: a national-level analysis of COVID-19 incidence. *Health Policy Plan* 2021;**36**:620–9. https://doi.org/10.1093/heapol/czab027

91. Linden A. Conducting interrupted time-series analysis for single-and multiple-group comparisons. *Stata J* 2015;**15**:480–500.

92. Ranapurwala SI, Ringwalt CL, Pence BW, Schiro S, Fulcher N, McCort A, *et al.* State medical board policy and opioid prescribing: a controlled interrupted time series. *Am J Prev Med* 2021;**60**:343–51. https://doi.org/10.1016/j.amepre.2020.09.015

93. Rieger M, Wagner N, Mebratie A, Alemu G, Bedi A. The impact of the Ethiopian health extension program and health development army on maternal mortality: a synthetic control approach. *Soc Sci Med* 2019;**232**:374–81. https://doi.org/10.1016/j.socscimed.2019.05.037

94. Drummond MF, Sculpher MJ, Claxton K, Stoddart GL, Torrance GW. *Methods for the Economic Evaluation of Health Care Programmes*. Oxford: Oxford University Press; 2015.

95. Edwards RT, McIntosh E. *Applied Health Economics for Public Health Practice and Research*. Oxford: Oxford University Press; 2019.

96. NICE. *Methods for the Development of NICE Public Health Guidance: Process and Methods Guides*. London: NICE; 2012.

97. McIntosh E, Luengo-Fernandez R. Economic evaluation. Part 1: introduction to the concepts of economic evaluation in health care. *J Fam Plann Reprod Health Care* 2006;**32**:107–12.

98. Walker S, Fox A, Altunkaya J, Colbourn T, Drummond M, Griffin S, *et al.* Program evaluation of population-and system-level policies: evidence for decision making. *Med Decis Making* 2022;**42**:17–27.

99. Kreif N, Mirelman AJ, Love-Koh J, Kim S, Moreno-Serra R, Revill P, *et al.* From impact evaluation to decision-analysis: assessing the extent and quality of evidence on 'value for money' in health impact evaluations in low-and middle-income countries. *Gates Open Res* 2021;**5**:1.

100. Alfonso YN, Bishai D, Bua J, Mutebi A, Mayora C, Ekirapa-Kiracho E. Cost-effectiveness analysis of a voucher scheme combined with obstetrical quality improvements: quasi experimental results from Uganda. *Health Policy Plan* 2015;**30**:88–99.

101. Leyland AH, Ouédraogo S, Nam J, Bond L, Briggs AH, Gray R, *et al.* Evaluation of health in pregnancy grants in Scotland: a natural experiment using routine data. *Publ Health Res* 2017;**5**:1–278.

102. Craig, Cooper C, Gunnell D, Haw S, Lawson K, Macintyre S, *et al. Using Natural Experiments to Evaluate Population Health Interventions*. Glasgow: Medical Research Council; 2011.

103. Angrist J, Bettinger E, Bloom E, King E, Kremer M. Vouchers for private schooling in Colombia: evidence from a randomized natural experiment. *Am Econ Rev* 2002;**92**:1535–58. https://doi.org/10.1257/000282802762024629

104. Manca A, Austin PC. *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 08/20. York: Department of Economics, University of York; 2008.

105. Kreif, G, Sadique. Statistical methods for cost-effectiveness analyses that use observational data: a critical appraisal tool and review of current practice. *Health Econ* 2013;**22**:486–500. https://doi.org/10.1002/hec.2806

106. Smith RD, Petticrew M. Public health evaluation in the twenty-first century: time to see the wood as well as the trees. *J Public Health (Oxf)* 2010;**32**:2–7.

107. Petticrew M, Cummins S, Ferrell C, Findlay A, Higgins C, Hoy C, *et al.* Natural experiments: an underused tool for public health? *Public Health* 2005;**119**:751–7.

108. Deidda M, Geue C, Kreif N, Dundas R, McIntosh E. A framework for conducting economic evaluations alongside natural experiments. *Soc Sci Med* 2019;**220**:353–61. https://doi.org/10.1016/j.socscimed.2018.11.032

109. Briggs A, Sculpher M, Claxton K. *Decision Modelling for Health Economic Evaluation*. Oxford: Oxford University Press; 2006.

110. Marsh K, IJzerman M, Thokala P, Baltussen R, Boysen M, Kaló Z, *et al.*; ISPOR Task Force. Multiple criteria decision analysis for health care decision making – emerging good practices: report 2 of the ISPOR MCDA Emerging Good Practices Task Force. *Value Health: J Int Soc Pharmacoecon Outc Res* 2016;**19**:125–37.

111. Walker S, Griffin S, Asaria M, Tsuchiya A, Sculpher M. Striving for a societal perspective: a framework for economic evaluations when costs and effects fall on multiple sectors and decision makers. *Appl Health Econ Health Policy* 2019;**17**:577–90. https://doi.org/10.1007/s40258-019-00481-8

112. Squires H, Chilcott J, Akehurst R, Burr J, Kelly MP. A framework for developing the structure of public health economic models. *Value Health: J Int Soc Pharmacoecon Outc Res* 2016;**19**:588–601.

113. Nixon RM, Thompson SG. Methods for incorporating covariate adjustment, subgroup analysis and between-centre differences into cost-effectiveness evaluations. *Health Econ* 2005;**14**:1217–29.

114. Jones AM, Lomas J, Moore P, Rice N. 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. *J Royal Stat Soc: Ser A (Stat Soc)* 2015;**179**:951–74.

115. Basu A. Estimating person-centered treatment (PeT) effects using instrumental variables: an application to evaluating prostate cancer treatments. *J Appl Econ (Chichester, England)* 2014;**29**:671–91. https://doi.org/10.1002/jae.2343

116. Teixeira-Pinto A, Normand SLT. Correlated bivariate continuous and binary outcomes: issues and applications. *Stat Med* 2009;**28**:1753–73.

117. Gomes M, Ng ES-W, Grieve R, Nixon R, Carpenter J, Thompson SG. Developing appropriate methods for cost-effectiveness analysis of cluster randomized trials. *Med Decis Mak* 2012;**32**:350–61.

118. Petrou S, Gray A. Economic evaluation using decision analytical modelling: design, conduct, analysis, and reporting. *BMJ* 2011;**342**:d1766.

119. Breeze PR, Squires H, Ennis K, Meier P, Hayes K, Lomax N, *et al.* Guidance on the use of complex systems models for economic evaluations of public health interventions. *Health Econ* 2023;**32**:1603–25. https://doi.org/10.1002/hec.4681

120. Zucchelli E, Jones A, Rice N. The evaluation of health policies through dynamic microsimulation methods. *Int J Microsim* 2012;**5**:2–20.

121. Espinoza MA, Manca A, Claxton K, Sculpher MJ. The value of heterogeneity for cost-effectiveness subgroup analysis: conceptual framework and application. *Med Decis Mak* 2014;**34**:951–64.

122. Cookson R, Robson M, Skarda I, Doran T. Equity-informative methods of health services research. *J Health Organ Manag* 2021;**35**:665–81.

123. Asaria M, Griffin S, Cookson R. Distributional cost-effectiveness analysis: a tutorial. *Med Decis Mak* 2016;**36**:8–19.

124. Verguet S, Laxminarayan R, Jamison DT. Universal public Finance of tuberculosis treatment in India: an extended cost-effectiveness analysis. *Health Econ* 2015;**24**:318–32.

125. Wong G, Westhorp G, Manzano A, Greenhalgh J, Jagosh J, Greenhalgh T. RAMESES II reporting standards for realist evaluations. *BMC Med* 2016;**14**:96. https://doi.org/10.1186/s12916-016-0643-1

126. Shaw S, Paparini S, Murdoch J, Green J, Greenhalgh T, Hanckel B, *et al.* TRIPLE C reporting principles for case study evaluations of the role of context in complex interventions. *BMC Med Res Methodol* 2023;**23**:115.

127. 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**:h1258.

128. 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**:1167.

129. Jones S, Hamilton S, Bell R, Araújo-Soares V, Glinianaia SV, Milne EMG, *et al.* What helped and hindered implementation of an intervention package to reduce smoking in pregnancy: process evaluation guided by normalization process theory. *BMC Health Serv Res* 2019;**19**:297. https://doi.org/10.1186/s12913-019-4122-1

130. 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*. London: National Institute for Health Research; 2018.

131. 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**:551.

132. 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 Res* 2016;**4**:1–154.

133. 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**:605–12. https://doi.org/10.1016/j.healthplace.2012.01.003

134. 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**:497.

135. Green J, Perkins C, Steinbach R, Edwards P. Reduced street lighting at night and health: a rapid appraisal of public views in England and Wales. *Health Place* 2015;**34**:171–80. https://doi.org/10.1016/j.healthplace.2015.05.011

136. Burges Watson D, Adams J, Azevedo LB, Haighton C. Promoting physical activity with a school-based dance mat exergaming intervention: qualitative findings from a natural experiment. *BMC Public Health* 2016;**16**:609. https://doi.org/10.1186/s12889-016-3308-2

137. 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 Soc* 2021;**2**:100029.

138. 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. *Int J Environ Res Public Health* 2019;**16**:1994.

139. Hewer RMF, Hill S, Amos A, consortium S-R. Student perceptions of smoke-free school policies in Europe – a critical discourse analysis. *Crit Publ Health* 2022;**32**:509–22. https://doi.org/10.1080/09581596.2020.1856332

140. Halkjær S, Lueg R. The effect of specialization on operational performance. *Int J Oper Prod Manag* 2017;**37**:822–39. https://doi.org/10.1108/IJOPM-03-2015-0152

141. Gibson K, Pollard TM, Moffatt S. Social prescribing and classed inequality: a journey of upward health mobility? *Soc Sci Med* 2021;**280**:114037.

142. 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. *Publ Health Res* 2018;**6**:1–248.

143. Guevara T, Sarmiento OL, Higuera D, Useche AF, Rubio MA, Wilches MA, *et al. Urban Transformations and Health: Results from the TransMiCable Evaluation [Urban Health in Latin America (SALURBAL) Project]*. 2020. URL: https://drexel.edu/~/media/Files/lac/Publications/TransMiCableENG.ashx?la=en (accessed 14 July 2023).

144. 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. *Publ Health Res* 2018;**6**:1–110.

145. Katikireddi SV, Bond L, Hilton S. Changing policy framing as a deliberate strategy for public health advocacy: a qualitative policy case study of minimum unit pricing of alcohol. *Milbank Q* 2014;**92**:250–83. https://doi.org/10.1111/1468-0009.12057

146. Hawe P, Shiell A, Riley T. Theorising interventions as events in systems. *Am J Community Psychol* 2009;**43**:267–76. https://doi.org/10.1007/s10464-009-9229-9

147. 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.

148. Ng SW, Colchero MA, White M. How should we evaluate sweetened beverage tax policies? A review of worldwide experience. *BMC Publ Health* 2021;**21**:1941. https://doi.org/10.1186/s12889-021-11984-2

149. Maini R, Mounier-Jack S, Borghi J. How to and how not to develop a theory of change to evaluate a complex intervention: reflections on an experience in the Democratic Republic of Congo. *BMJ Global Health* 2018;**3**:e000617. https://doi.org/10.1136/bmjgh-2017-000617

150. Moffatt S, White M, Mackintosh J, Howel D. Using quantitative and qualitative data in health services research – what happens when mixed method findings conflict? [ISRCTN61522618]. *BMC Health Serv Res* 2006;**6**:28. https://doi.org/10.1186/1472-6963-6-28

151. Green J, Hanckel B, Petticrew M, Paparini S, Shaw S. Case study research and causal inference. *BMC Med Res Methodol* 2022;**22**:307. https://doi.org/10.1186/s12874-022-01790-8

152. Mahoney J. The logic of process tracing tests in the social sciences. *Soc Method Res* 2012;**41**:570–97.

153. Robinson WS. The Logical Structure of Analytic Induction. In: Gomm R, Hammersley M, Foster P, editors. *Case Study Method: Key Issues, Key Texts*. London: SAGE Publications Ltd; 1951. p. 187.

154. 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. *Publ Health Res* 2019;**7**:1–46.

155. 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**:e004017. https://doi.org/10.1136/bmjgh-2020-004017

156. Elm E, 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**:806–8. https://doi.org/10.1136/bmj.39335.541782.AD

157. 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. *Am J Public Health* 2004;**94**:361–6.

158. Benchimol EI, Smeeth L, Guttmann A, Harron K, Moher D, Petersen I, *et al.*; RECORD Working Committee. The REporting of studies Conducted using Observational Routinely-collected health Data (RECORD) statement. *PLOS Med* 2015;**12**:e1001885.

159. 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**:g1687.

160. 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. *Int J Technol Assess Health Care* 2022;**38**:3–9.

161. O'Brien BC, Harris IB, Beckman TJ, Reed DA, Cook DA. Standards for reporting qualitative research: a synthesis of recommendations. *Acad Med: J Assoc Am Med Coll* 2014;**89**:1245–51.

162. Shamseer L, Moher D, Clarke M, Ghersi D, Liberati A, Petticrew M, *et al.*; the PRISMA-P Group. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015: elaboration and explanation. *BMJ* 2015;**349**:g7647.

163. 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. https://doi.org/10.1136/bmj.n160

164. 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**:l6890.

165. 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. *J Clin Epidemiol* 2014;**67**:56–64.

166. 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. *J Clin Epidemiol* 2013;**66**:1230–43.

167. Sandelowski M, Voils CI, Barroso J. Defining and designing mixed research synthesis studies. *Res School: Nation Refer J Mid-South Educ Res Assoc Univ Alabama* 2006;**13**:29.

168. Heyvaert M, Hannes K, Onghena P. *Using Mixed Methods Research Synthesis for Literature Reviews: The Mixed Methods Research Synthesis Approach*. London: SAGE Publications Ltd; 2016.

169. 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. *J Clin Epidemiol* 2017;**89**:30–42. https://doi.org/10.1016/ j.jclinepi.2017.02.016

170. López-López JA, Page MJ, Lipsey MW, Higgins JP. Dealing with effect size multiplicity in systematic reviews and meta-analyses. *Res Synth Method* 2018;**9**:336–51.

171. Thomson H, Craig P, Hilton-Boon M, Campbell M, Katikireddi SV. Applying the ROBINS-I tool to natural experiments: an example from public health. *Syst Rev* 2018;**7**:15. https://doi.org/10.1186/s13643-017-0659-4

172. 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. *J Clin Epidemiol* 2021;**140**:22–32. https://doi.org/10.1016/j.jclinepi.2021.08.022

173. 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. *J Clin Epidemiol* 2017;**89**:43–52. https://doi.org/10.1016/j.jclinepi.2017.02.015

174. Villar PF, Waddington H. Within study comparisons and risk of bias in international development: systematic review and critical appraisal. *Campbell Syst Rev* 2019;**15**:e1027.

175. ROBINS-E Development Group. *Risk of Bias in Non-randomized Studies: Of Exposure (ROBINS-E)*. 2023. URL: www.riskofbias.info/welcome/robins-e-tool (accessed 14 July 2023).

176. Effective Public Health Practice Project. *Quality Assessment Tool for Quantitative Studies*. Hamilton, Canada. 2022. URL: www.ephpp.ca/PDF/Quality%20Assessment%20Tool_2010_2.pdf (accessed 21 July 2022).

177. Cochrane Effective Practice and Organisation of Care (EPOC). *Suggested Risk of Bias Criteria for EPOC Reviews: EPOC Resources for Review Authors*. 2017. URL: https://epoc.cochrane.org/sites/epoc.cochrane.org/files/public/uploads/Resources-for-authors2017/suggested_risk_of_bias_criteria_for_epoc_reviews.pdf (accessed 14 July 2022).

178. 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. *J Clin Epidemiol* 2018;**97**:49–58.

179. Critical Appraisal Skills Programme (CASP). *Qualitative Checklist*. 2018. URL: https://casp-uk.net/images/checklist/documents/CASP-Qualitative-Studies-Checklist/CASP-Qualitative-Checklist-2018.pdf (accessed 16 June 2023).

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

181. Welch V, Petticrew M, Petkovic J, Moher D, Waters E, White H, Tugwell P; PRISMA-Equity Bellagio Group. Extending the PRISMA statement to equity-focused systematic reviews (PRISMA-E 2012): explanation and elaboration. *J Clin Epidemiol* 2016;**70**:68–89.

182. McKenzie JE, Brennan SE. Synthesizing and Presenting Findings Using Other Methods. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. New York: John Wiley & Sons; 2019. pp. 321–47.

183. Noyes J, Booth A, Cargo M, Flemming K, Harden A, Harris J, *et al.* Qualitative Evidence. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. New York: John Wiley & Sons; 2019. pp. 525–45.

184. Pawson R, Greenhalgh T, Harvey G, Walshe K. Realist review: a new method of systematic review designed for complex policy interventions. *J Health Serv Res Pol* 2005;**10**:21–34.

185. Shemilt I, Mugford M, Vale L, Marsh K, Donaldson C. *Evidence-Based Decisions and Economics: Health Care, Social Welfare, Education and Criminal Justice*. New York: John Wiley & Sons; 2011.

186. Deeks JJ, Higgins JP, Altman DG, Group CSM. Analysing data and undertaking meta-analyses. In: Higgins JPT, Thomas J, Chandler J, Cumpston M, Li T, Page MJ, Welch VA, editors. *Cochrane Handbook for Systematic Reviews of Interventions*. New York: John Wiley & Sons; 2019. pp. 241–84.

187. 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. *J Clin Epidemiol* 2011;**64**:383–94.

188. 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 Netw Open* 2022;**5**:e220077.

189. 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**:b4469.

190. Hong J, McArthur DP, Livingston M. The evaluation of large cycling infrastructure investments in Glasgow using crowdsourced cycle data. *Transportation* 2020;**47**:2859–72.

191. 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). *Transp Res Part A: Pol Pract* 2020;**133**:109–21.

192. Goldacre B, Morley J. *Better, Broader, Safer: Using Health Data for Research and Analysis*. London: Department of Health and Social Care; 2022.

Appendix 1

- List of advisory group members
- List of oversight group members
- List of workshop contributors
- List of consultation contributors
- *Table 8* Consultation questions
- *Table 9* Consultation participants' roles and organisations

Advisory group members:

Adrian Bauman, University of Sydney
Kate Tilling, University of Bristol
Marc Suhrcke, Luxembourg Institute of Social and Economic Research
Niamh Fitzgerald, University of Stirling
Sara Shaw, University of Oxford
Audrey Ceschia, Editor Lancet Public Health
Scott Lloyd, Middlesbrough Council and Associate Lead for Public Health Research, NIHR Clinical Research Network North East & North Cumbria
Sarah Sharples, CSA Dept for Transport

Oversight group members:

Graham Hart (Chair), MRC PHIND panel
Peymané Adab, MRC-NIHR Methodology Advisory Group
Claire Kidgell, Assistant Director NIHR
Catherine Moody, Head of Population Health UKRI MRC
Tamsyn Derrick, Programme Manager Population Health Sciences and PHIND UKRI MRC
Peter Craig, representing project team, MRC/CSO SPHSU, University of Glasgow

Workshop participants who provided their name for acknowledgement:

Agnes Erzse, Bridget C. Foley, Chris Bonell, Don Nutbeam, Elzo Junior, Igor Francetic, Jennifer Yip, Joanne McKenzie, Lindsay McLaren, Mark Robinson, Mauricio Barreto, Rodrigo Reis, Ruth Hunter, Simon Turner, Thad Dunning

Consultation contributors:

Aaron Reeves, Abraham George, Adrian Bauman, Benedict Armstrong, Bridget Foley, Christina Vogel, Cristina Fernandez-Garcia, Diane T Finegood, Hannah Forde, Jennifer Yip, Michele Hilton Boon, Emma Lawlor, Elzo Pereira Pinto Junior, Emily Widnall, Famke J.M. Molenberg, Frank Kee, Hugh Sharma Waddington, Jack Benton, James White, Janis Baird, Jenna Panter, Jessica Renzella, Joanne McKenzie, Lindsay McLaren, Magdalena Opazo Breton, Mark Robinson, Martin O'Flaherty, Matt Egan, Melanie Crane, Miriam Alvarado, Niamh Fitzgerald, Nick Cavill, Olga Lucia Sarmiento, Peter Tennant, Rachel Thomson, Rebecca Cannings-John, Ronan Lyons, Ruth Salway, Sam Harper, Sandro Galea, Sara Benjamin-Neelon, Siv Steffen Nygaard, Tamilore Sonubi, Tony Blakely

TABLE 8 Consultation questions

Each of the eight framework sections were accompanies by the same questions asking:	
Do you agree with the content of the proposed 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	
Additional questions:	
Consultation section: Concepts and definitions	If there are any key terms missing from the glossary, please describe in the comments section below.
Consultation section: Qualitative methods	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. 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.
Consultation section: Good practice considerations	Please suggest amendments and/or further recommendations that you think we should emphasise, based on the content of the earlier sections.

TABLE 9 Consultation participants' roles and organisations

Participant role (some provided multiple responses)	
Role	***No.***
Member of funding board	6
Funding body representative	1
Journal editor	8
Intervention researcher	18
Quantitative researcher	39
Qualitative researcher	17
Policy-maker	1
Practitioner	2
Clinician	4
Other	1
Details provided for 'Other'	Public Health Consultant
Participant organisation (some provided multiple responses)	
Organisation	***No.***
University	41
Public sector organisation	5
Non-profit organisation	2
For-profit	2
Other	0

EME
HSDR
HTA
PGfAR
PHR

Part of the NIHR Journals Library
www.journalslibrary.nihr.ac.uk

Published by the NIHR Journals Library