Using and distinguishing evidence from non-randomised studies of interventions

Jamie Hartmann-Boyce*, Nicola Lindson*, Lisa Bero, Jo Dumville, Ella Flemyng, Barney Reeves, Peter Tugwell, David B Wilson, Hugh Sharma Waddington

*joint first authors

Non-randomized studies of interventions (NRSIs) are defined in the Cochrane Handbook for Systematic Reviews of Interventions as any quantitative study estimating the effects
(benefit or harm) of an intervention that does not use randomization to allocate individuals (or other units such as groups of individuals) to intervention groups.[1] NRSIs use a variety of designs and methods to estimate and quantify the causal effect on the outcome(s) of interest of a health care intervention, programme or policy (Box 1). It is increasingly recognised that NRSIs have an important role to play in systematic reviews of health care interventions (e.g.,[2]), including on topics where interventions are typically directed to a population, such as in the areas of public health and planetary health policy, or where the effect of interest sought is of an intervention delivered in real-world settings[3], as well as in assessing harms and long-term effects of clinical practice interventions.

NRSI study designs can be described in many ways. For example, they may use cohort, cross-sectional or case-control data, estimate intervention effects with respect to contemporaneous or historical comparators, or both, and measure or analyse outcomes at individual patient level or in the aggregate. Unfortunately, study design labels are often used inconsistently [1], [4], [5], which can make it difficult to evaluate a study's risk of bias. For example, regarding confounding – that is, common causes of intervention group allocation and outcome other than the intervention of interest – it is important to know specific aspects of a study's design, such as whether baseline (pre-intervention) data were incorporated in analysis, or whether the same units are follow-up over time versus a repeated cross-section. This helps in establishing the extent to which the study can identify a causal relationship from the association measured between intervention and outcome, which is often of most relevance for policy and practice decision-making.

This paper aims to help readers consider when to use NRSI evidence, to ensure reviewers incorporate all evidence relevant for the review question, and how to differentiate between different NRSIs based on some key features, which assists reviewers in evaluating the risk of bias. Cochrane guidance states that NRSI eligibility criteria should be defined on design features, not labels, but it can be difficult to understand what this means in practice so that it can be applied to systematic reviews consistently and transparently. This work was initiated to fill this gap in the guidance, based on the experience of Cochrane authors in struggling with how to apply it. The paper accompanies tutorials published in Cochrane (forthcoming tutorials currently in development). We also provide a glossary of study design labels and their commonly associated features.

When are NRSIs most useful in informing decisions about intervention effects?

When trying to determine the effects of a healthcare intervention, randomized controlled trials (RCTs) which are designed and conducted rigorously are widely considered the least biased form of primary evidence, due to their ability to address measured and unmeasured confounding. However, there are instances where clinicians, policy makers, or systematic reviewers may need to consider NRSIs. Some examples can be seen in Table 1.

Some NRSIs may also be known and used in policy and practice, but be ineligible for a systematic review that only planned to include RCTs or certain types of NRSIs. In these cases, reviewing this NRSI evidence can improve the relevance of the review for decision making (e.g., by helping raise awareness about concerns regarding these studies through risk-of-bias assessment).

Key features that can help differentiate NRSIs

Study design features should be used to differentiate between NRSIs in systematic reviews, for example to clarify inclusion criteria and to perform adequate risk-of-bias assessment. Study design labels, such as “controlled before-after” or “interrupted time series”, are often used by authors as short-hand names to distinguish between different kinds of NRSI. Because the labels are applied inconsistently, and because studies are designed and conducted with varying degrees of complexity, it is important to specify NRSI with respect to the key design features, as highlighted in Cochrane guidance (Reeves et al. 2019). This section presents four types of design feature that can be used to characterise NRSI, drawing on Reeves et al. [1] and the Campbell Collaboration [6] (Box 2). This represents a spectrum of methodological rigor, in which RCTs can be considered as the benchmark (e.g. following a similar approach to ROBINS-I [7]) and other designs are classified based on their ability to control for confounding (Table 2).

How was the intervention effect measured?

Intervention effects measure the magnitude of the change in outcomes relative to a benchmark called the “comparator”. The intervention effect is the difference in outcomes measured between units (e.g. individuals or groups) receiving the intervention, programme or policy (such as anti-smoking health messaging) and units receiving the comparator (for example, no intervention or an alternative, those receiving differing levels of the intervention, or measured before versus after receiving it). When considering NRSI evidence, it should be clear whether the intervention effect is measured as:

- The difference between two groups, for example those receiving an anti-smoking health messaging intervention and those not receiving it, collected from both groups at the same point in time (i.e., contemporaneously);
- The difference before versus after receiving the intervention (within groups), for example comparing outcomes in people before versus after when they received the intervention (i.e., historically); or
- Contemporaneously *and* historically (also called the “double-difference”) (i.e., the intervention effect was measured as the difference between groups in the within-group change over time).

In addition, some NRSIs provide estimates that are generalizable to sub-sample of the study population, such as only those within the bandwidth around the allocation threshold of the forcing variable in a regression discontinuity design (RDD), or only those who adhere to the intervention in an instrumental variable study. Even when they are designed and conducted appropriately, these studies produce local intervention effects that might differ from the population intervention effect produced in representative trial, due to the different sub-samples over which the intervention effect estimate is calculated.

Single-group (uncontrolled) pre-test post-test designs (historical comparisons) are often considered unreliable, owing to the inability of the study to address historical intervention confounds (other interventions occurring at the same time) or maturation effects (e.g., regression to the mean). Similarly, NRSIs where data are only collected at end-line (e.g., cross-sectional and case-control designs) may be considered unreliable if they are not adequately able to demonstrate that the intervention clearly preceded the outcome (temporal precedence).

What determined who received the intervention?

One of the strengths of RCTs is that participation in the intervention of interest is determined randomly. This helps reduce the risk of bias, since people that receive health interventions have certain characteristics which make them systematically different (i.e., incomparable) to people that don’t.

Knowledge about what determines intervention receipt, as in the case of RCTs and some NRSIs, helps in addressing the risk of bias.

In NRSIs, whether a participant receives an intervention (the intervention allocation method) may be determined by an independent third party, a researcher, a policymaker, the patient's clinician or the patient themselves. For example, a clinician may choose to prescribe an intervention for a patient based on their characteristics, such as delivery in a hospital maternity-ward for women at risk of birth complications. Or a patient may specifically request an intervention because a friend or relative has received it, such as a health behavior change intervention to stop smoking, and therefore may be more motivated to see an improvement. Both scenarios could introduce bias, if the group to which the treated units are being compared does not have the same pre-existing patient characteristics or motivation.

The most rigorous NRSIs are usually those where someone (or something) independent determines who receives the intervention. The intervention allocation method might take different forms, such as a threshold on a scale variable (called a "forcing variable"). For example, antiretrovirals are prescribed to those with a CD4 count below a particular threshold[8]; vaccinations are given to those at a particular age [9], [10]. Modelling the relationship between the forcing variable and health outcomes like survival can produce an unbiased intervention effect estimate, using a method called the discontinuity design (or regression discontinuity design, RDD). Systematic reviews and meta-analyses that compared effect size estimates in RCTs and RDDs conducted at the same time in the same populations (so-called "internal replication studies") concluded they provided results that were statistically indistinguishable from one another and of very small absolute difference.[11], [12]

More generally, we are usually more confident about a NRSI when the allocation method is known and can be modelled.[13] So, for example, if we want to estimate the effect of ward delivery for at-risk pregnancies, we would want to ensure that the comparison observations are also of at-risk pregnancies (but who did not receive ward delivery). In contrast, studies where the allocation method is unknown – for example, where groups are formed purely by self-selection of patients (e.g. patients indicating in a survey whether they adhere to a particular intervention) – are more likely to produce biased intervention effect estimates.

What outcome data were collected, when, and from whom?

Outcome data may be collected before the intervention has commenced, at baseline (also called pre-test measurement), after the intervention at endline (post-test measurement), in intervention and comparison groups. Sometimes there are multiple periods of observation before and after the intervention (i.e. multiple pre-tests and/or multiple post-tests). In NRSI that require one or

more pre-tests, it is important that pre-test measurement is done before the intervention has been implemented (e.g., before the patient starts intervention). For example, some studies described as collecting controlled-before versus after data may in fact collect data only after the intervention has started, which would cause bias in the intervention effect estimate, since we cannot be sure that treated and comparison pre-tests accurately represent the pre-intervention characteristics of both groups.

Data may also be collected at individual patient or group levels (e.g., health facility, neighbourhood or other).

There is also discussion about the minimum number of observations before and after intervention for valid causal inference in an interrupted time-series (ITS) design, since a minimum number of observations at discrete time points is required to establish reliable trends in outcomes for the condition without intervention (the pre-tests) and the condition with intervention (the post-tests). For example, one internal replication study that measured outcomes from randomized clusters of individuals over an interrupted time-series suggested that valid ITS design, producing the most similar magnitude of effect as the RCT, needed six pre-test and six post-test observations.[14]

However, it is likely that far more time-series observations are needed in most studies. For example, it is important that the measurement of trends can account for time-varying sources of confounding like seasonality. In the classic interrupted time series, where there is only a single aggregate measurement at each time point (like a mortality rate, or monthly crime count), at least 50 time points are thought needed, with at least 20 before or during the intervention.[6]

In some studies, the same participants are followed up in successive rounds of data collection (also called “panel data”), whereas in others, data from repeated cross-sections or health episodes are collected from different (or some different and some same) participants in each round (“pseudo-panel data”). Panel data are thought to be at lower risk of bias than repeated cross-section data.[15] This is because attrition (i.e., losses to follow-up), which is a key potential source of bias in panels, can be measured. In contrast, differential selection of participants into study groups (e.g., ‘joiners’ who enroll after assignment of the intervention), a key potential source of bias in repeated cross-sections, is unmeasurable.

How well can the study address confounding?

Confounding is the distortion of the observed relationship between independent (i.e., intervention) and dependent (i.e., outcome) variables due to another variable that independently causes both (see, for example, A Beginners Guide to Confounding by Eveliina Ilola: https://s4be.cochrane.org/blog/2018/10/01/a-beginners-guide-to-

confounding/). Confounding makes it challenging to ascertain and measure the true causal relationship between the intervention, programme or policy and the outcome, because the confounding variable generates a spurious correlation between the intervention and outcome (i.e., an association that is non-causal). This is a particularly important issue in NRSIs, which can affect different types of designs in different ways, depending on the three other study design features listed above. Understanding and accounting for confounders by design, or controlling for them in analysis, is important for inferring causality in NRSIs, and hence determining whether the study is at risk of bias.

Observable confounders are factors that are measured and considered in analysis in primary studies, such as prognostic factors like sociodemographic characteristics and geographical location. Unobservable confounders are factors that are unmeasurable (or not typically measured) in primary studies, such as individual participant or group motivation and aptitudes. For example, it is usually possible to observe factors like patient age, sex, location and group membership, but difficult or impossible to measure aptitudes like ability or motivation for change.

NRSIs that can potentially address unobservable sources of confounding include discontinuity designs, and studies designed or analysed using instrumental variables (e.g., Mendelian randomization). However, it is important that these types of studies are well-designed and conducted; for example, in instrumental variables using Mendelian randomization the genetic marker must be closely related to intervention allocation and fixed over the course of the study (time-invariant). Some studies analysing controlled-before versus after data can address sources of unobservable confounding that are fixed over the course of a study, at the unit of analysis (e.g. the patient or health episode); an example might be innate ability in a study of psychosocial development. However, these types of studies cannot address sources of unobservable confounding that might be expected to vary over the course of a study, like participant motivation.

Other types of study (and otherwise more rigorous designs with problems in conduct) are likely to address observable sources of confounding only. For example, patient-reported outcomes are often thought more likely to be biased in unblinded trials, especially when participants are observed repeatedly, than in studies using clinician-verified outcomes or administrative records of whatever design. For studies addressing observable confounding only, it is recommended that evaluators consider which confounders are particularly important for that study to take into account, based on the clinical evidence sought (see Riskofbias.info). Ideally, these should be pre-specified as per ROBINS-I.[7] These may depend on the clinical/contextual details of each review and could include sociodemographic factors (e.g., age, sex, location) that are common

predictors of the outcome, and where possible information about the method used to allocate intervention (e.g., known intervention allocation rule used by practitioner).[16]

The funder didn’t influence the results/outcomes of the study despite author affiliations with the funder.

Contributorship statement
All authors contributed to the conception and design of the work. JHB, NL and HSW drafted the work. All authors critically revised the work. All authors gave final approval of the version to be published. JHB, NL and HSW are guarantors for this article. JHB and NL are joint lead authors.

Conflicts of interest
All authors declare no conflicts of interest in regards to this manuscript.

Funding
NL, JHB and HSW’s work on this project was financially supported by Cochrane. No other authors received funding to support this work.

Data sharing
Data sharing is not applicable as no datasets were generated for this study

Patient and public involvement
Patients and the public were not involved in this manuscript.

Tables

Table 1. Examples of instances where NRSIs may need to be considered when evaluating intervention effects

Instance	Example(s)
Evidence about unintended consequences or adverse events (harms), which RCTs may not measure, or only measure imprecisely, and NRSIs may measure with limited risk of bias [17]	The effects of infant sleeping position on mortality [18]
Evidence about the effects of interventions conducted in real-world settings rather than under highly controlled (explanatory or proof-of-concept) settings, especially in settings or fields where the conduct of an RCT would be challenging, uncommon, or unethical	The impacts of e-cigarette promotion on non-smoking youth [19] or of juvenile curfews on drug-taking[20]
Evidence about long-term effects, where RCTs may be impractical,[1] due to the need to isolate control groups from the health care intervention over long periods	The long-term effects of mass deworming interventions on stunting, education and work [21]
Outcomes measured among groups of participants who are not well represented in other studies, such as those included under PROGRESS-Plus [22][2]	The effects of farmer field schools [23]
Evidence about rare primary outcomes, for which RCTs may be underpowered. In such cases, the RCT evidence base may comprise outcomes which are more accessible for researchers, but may be less clinically relevant for policy and practice.	Infectious disease mortality comprises most of the global burden of disease in childhood, but RCTs of environmental health interventions like drinking water treatment and hygiene promotion measure effects on morbidity [24]
Emergent health conditions, if RCTs are likely to take longer to design and conduct than the decision-making cycle requires	A review on the effects of personal protective equipment in reducing infection conducted in response to the COVID-19 pandemic [25]

[1] Even where receipt of an intervention is time-limited, so the risk of contamination of controls is minimised, it can be difficult practically to obtain funding for long-term follow-ups. However, long-term data collected from routine sources (e.g. health service administration) are increasingly being incorporated in randomized trials

[2] PROGRESS stands for place of residence, race/ethnicity/culture/language, occupation, gender/sex, religion, education, socioeconomic status and social capital; Plus refers to personal characteristics associated with discrimination (e.g., disability), features of relationships (e.g., parents who smoked) and instances when a person is temporarily disadvantaged (e.g., respite care).

Table 2: Glossary of study designs and their main features (see Box 2 for definitional concepts)[3]

Approach	Key definitional concepts	Effect measurement	Allocation to intervention	Data collection	Confounding type addressed
Randomized controlled trial (RCT)	Have the essential features of random assignment of study participants to intervention conditions, such as intervention and control.	Contemporaneous control[1]	Randomization	At least one post-test[5] and often one pre-test[6]	Unobservable[7] and observable[8]
Quasi-RCT	Uses a known, systematic method of assignment to intervention, such as alternation of participants ordered alphabetically to intervention groups.	Contemporaneous control	Forcing variable[4] (e.g., alternation)	At least one post-test and often one pre-test	Unobservable and observable
(Regression) discontinuity design (RDD))	Intervention assignment is explicitly known to the researcher and based on a threshold on a scaled baseline measure.	Contemporaneous comparison[2]	Forcing variable[4] (scale threshold)	At least one post-test; sometimes a pre-test	Unobservable and observable
Instrumental variables (IV) design	A pre-intervention variable (an instrument) is predictive of who gets intervention but has no direct causal effect on the outcome	Contemporaneous comparison	Forcing variable[4] (instrument)	At least one post-test	Unobservable and observable
Interrupted time series (ITS) design	Usually relies on aggregate data, examined for some number of time periods prior to intervention start and some number of time periods after the start of the intervention	Historical comparison[3]	Forcing variable[4] (time)	Multiple pre-tests and multiple post-tests	Unobservable and observable
Controlled-ITS design	An augmentation of ITS with one or more comparison series	Contemporaneous and historical comparison	Forcing variable (time), with non-randomized intervention group selection	Multiple pre-tests and multiple post-tests	Unobservable and observable
Synthetic control design	Usually applied to instances where there is no natural comparison group, where a comparison group time-series is created statistically from multiple external groups	Contemporaneous and historical comparison	Non-randomized selection by researchers	At least one post-test and multiple pre-tests	Time-invariant unobservable and observable
Non-equivalent groups with pre-test and post-test	Includes an intervention group and a non-randomly created comparison group, and a baseline measure of the outcome (a pretest), which can be used to adjust for observable	Contemporaneous and historical comparison	Non-randomized selection by planners, practitioners,	At least one pre-test and one post-test	Time-invariant unobservable and observable

[3] The selection of included design was based on expert opinion of the author group *Relies entirely on statistical adjustment for measured confounders; remains susceptible to unmeasured confounding

design[4]	confounding, as well as unobservable confounders that are typically fixed over the course of a study. May also be combined with matching of groups at pre-test using statistical methods		researchers or patients		
Non-equivalent groups, post-test only design (cross-section design)	Includes a program or intervention group and a non-randomly created comparison group. In its simplest form, only has a single assessment of the outcome following program participation, although variations might include follow-up assessments.	Contemporaneous comparison	Non-randomized selection by planners, practitioners, researchers or patients	One post-test	Observable only*
Case-control design	Selects participants based on the outcome. Cases are individuals who exhibit the outcome and comparison individuals are those who do not exhibit the outcome. Exposure to a prior variable of interest is assessed.	Contemporaneous or historical comparison	Non-randomized selection by planners, practitioners, researchers or patients	One post-test	Observable only*
Cohort design (non-equivalent group cohort design)	A treated and an untreated group are followed up over time. Data collection points by group may or may not be contemporaneous.	Contemporaneous or historical comparison	Non-randomized selection by planners, practitioners, researchers or patients	At least one post-test and sometimes a pre-test	Observable only*
Single group pre-test post-test design (uncontrolled-before versus after (BA) design or reflexive control design)	Selects study participants that receive treatment. Their status on one or more outcome variables is assessed before and after treatment participation. Due to the absence of a contemporaneous comparison, the design is limited in its ability to address sources of confounding.	Historical comparison	Non-randomized selection by planners, practitioners or patients	One pre-test and at least one post-test	Limited ability to control for specific, measured (observable) confounders

[4] including controlled-before versus after (CBA) and non-randomized controlled trial (NRCT) designs

Box 1. Other terms for NRSI

In the social sciences, the term “quasi-experimental design” (QED) is used to describe non-randomized studies of intervention effects.[26] NRSI are also called “natural experiments” where researchers have no role in determining intervention allocation (assignment and implementation), and where existing data, or data collected for other purposes, is used in analysis.[27]

Box 2. Definitional notes on study designs covered in Table 2

1 Contemporaneous control: measurement of a group that does not receive the intervention of interest (or receives something else) at the same time as the intervention group in randomized controlled trials.
2 Contemporaneous comparison: measurement of a group that does not receive the intervention of interest (or receives something else) at the same time as the intervention group in non-randomized studies of interventions (quasi-experiments).
3 Historical comparison: measurement at a time before the intervention has occurred in non-randomized studies of interventions.
4 Forcing variable: a variable that determines intervention allocation at a particular value or threshold in non-randomized studies of interventions.
5 Post-test: measurement after the intervention has been implemented.
6 Pre-test: measurement before the intervention has been implemented.
7 Unobservable confounding: factors affecting both intervention allocation and the outcome which cannot be (or are not) measured.
8 Observable confounding: measured factors affecting both intervention allocation and the outcome.

References

[1] B. C. Reeves, G. A. Wells, and H. Waddington, “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.*, vol. 89, pp. 30–42, Sept. 2017, doi: 10.1016/j.jclinepi.2017.02.016.

[2] I. J. Saldanha *et al.*, “Inclusion of nonrandomized studies of interventions in systematic reviews of intervention effectiveness: an update,” 2022.

[3] H. S. Waddington, P. F. Villar, and J. C. Valentine, “Can non-randomised studies of interventions provide unbiased effect estimates? A systematic review of internal replication studies,” *Eval. Rev.*, vol. 47, no. 3, pp. 563–593, 2023.

[4] S. Polus *et al.*, “Heterogeneity in application, design, and analysis characteristics was found for controlled before-after and interrupted time series studies included in Cochrane reviews,” *J. Clin. Epidemiol.*, vol. 91, pp. 56–69, Nov. 2017, doi: 10.1016/j.jclinepi.2017.07.008.

[5] M. Vigneri, M. Clarke, J. Exley, P. Tugwell, V. Welch, and H. White, “Epidemiology and development economics two sides of the same coin in impact evaluation,” *J. Clin. Epidemiol.*, vol. 144, pp. 16–21, Apr. 2022, doi: 10.1016/j.jclinepi.2021.11.029.

[6] D. B. Wilson *et al.*, “A Study Design Nomenclature for Systematic Reviews Assessing the Effectiveness of Interventions,” *Campbell Collab. Rev.*, 2025.

[7] J. A. Sterne *et al.*, “ROBINS-I: a tool for assessing risk of bias in non-randomised studies of interventions,” *BMJ*, p. i4919, Oct. 2016, doi: 10.1136/bmj.i4919.

[8] J. Bor, E. Moscoe, P. Mutevedzi, M.-L. Newell, and T. Bärnighausen, “Regression Discontinuity Designs in Epidemiology: Causal Inference Without Randomized Trials,” *Epidemiology*, vol. 25, no. 5, pp. 729–737, Sept. 2014, doi: 10.1097/EDE.0000000000000138.

[9] N. E. Basta and M. E. Halloran, “Evaluating the Effectiveness of Vaccines Using a Regression Discontinuity Design,” *Am. J. Epidemiol.*, vol. 188, no. 6, pp. 987–990, June 2019, doi: 10.1093/aje/kwz043.

[10] M. Eyting, M. Xie, F. Michalik, S. Heß, S. Chung, and P. Geldsetzer, “A natural experiment on the effect of herpes zoster vaccination on dementia,” *Nature*, vol. 641, no. 8062, pp. 438–446, May 2025, doi: 10.1038/s41586-025-08800-x.

[11] D. D. Chaplin *et al.*, “THE INTERNAL AND EXTERNAL VALIDITY OF THE REGRESSION DISCONTINUITY DESIGN: A META-ANALYSIS OF 15 WITHIN-STUDY COMPARISONS,” *J. Policy Anal. Manage.*, vol. 37, no. 2, pp. 403–429, Mar. 2018, doi: 10.1002/pam.22051.

[12] H. Sharma Waddington *et al.*, “Quasi-experiments are a valuable source of evidence about effects of interventions, programs and policies: commentary from the Campbell Collaboration Study Design and Bias

Assessment Working Group," *J. Clin. Epidemiol.*, vol. 152, pp. 311–313, Dec. 2022, doi: 10.1016/j.jclinepi.2022.11.005.
[13]T. D. Cook, W. R. Shadish, and V. C. Wong, "Three conditions under which experiments and observational studies produce comparable causal estimates: New findings from within-study comparisons," *J. Policy Anal. Manage.*, vol. 27, no. 4, pp. 724–750, Sept. 2008, doi: 10.1002/pam.20375.
[14]A. Fretheim *et al.*, "A reanalysis of cluster randomized trials showed interrupted time-series studies were valuable in health system evaluation," *J. Clin. Epidemiol.*, vol. 68, no. 3, pp. 324–333, Mar. 2015, doi: 10.1016/j.jclinepi.2014.10.003.
[15]E. J. Caruana, M. Roman, J. Hernández-Sánchez, and P. Solli, "Longitudinal studies," *J. Thorac. Dis.*, vol. 7, no. 11, p. E537, 2015.
[16]H. Sharma Waddington *et al.*, "Risk of Bias in Experiments, Quasi-Experiments and Natural Experiments across Disciplines: Discussion Paper and Assessment Framework," *Campbell Syst. Rev.*, 2025.
[17]S. Golder, Y. K. Loke, and M. Bland, "Meta-analyses of Adverse Effects Data Derived from Randomised Controlled Trials as Compared to Observational Studies: Methodological Overview," *PLoS Med.*, vol. 8, no. 5, p. e1001026, May 2011, doi: 10.1371/journal.pmed.1001026.
[18]R. Gilbert, G. Salanti, M. Harden, and S. See, "Infant sleeping position and the sudden infant death syndrome: systematic review of observational studies and historical review of recommendations from 1940 to 2002," *Int. J. Epidemiol.*, vol. 34, no. 4, pp. 874–887, Aug. 2005, doi: 10.1093/ije/dyi088.
[19]R. Begh *et al.*, "Electronic cigarettes and subsequent cigarette smoking in young people: A systematic review," *Addiction*, vol. 120, no. 6, pp. 1090–1111, June 2025, doi: 10.1111/add.16773.
[20]D. B. Wilson, C. Gill, A. Olaghere, and D. McClure, "Juvenile Curfew Effects on Criminal Behavior and Victimization: A Systematic Review," *Campbell Syst. Rev.*, vol. 12, no. 1, pp. 1–97, Jan. 2016, doi: 10.4073/csr.2016.3.
[21]V. A. Welch *et al.*, "Deworming and adjuvant interventions for improving the developmental health and well-being of children in low- and middle-income countries: a systematic review and network meta-analysis," *Campbell Syst. Rev.*, vol. 12, no. 1, pp. 1–383, Jan. 2016, doi: 10.4073/csr.2016.7.
[22]J. O'Neill *et al.*, "Applying an equity lens to interventions: using PROGRESS ensures consideration of socially stratifying factors to illuminate inequities in health," *J. Clin. Epidemiol.*, vol. 67, no. 1, pp. 56–64, Jan. 2014, doi: 10.1016/j.jclinepi.2013.08.005.
[23]H. Waddington *et al.*, "Farmer field schools for improving farming practices and farmer outcomes: A systematic review," *Campbell Syst. Rev.*, vol. 10, no. 1, pp. i–335, 2014.
[24]H. Sharma Waddington, E. Masset, S. Bick, and S. Cairncross, "Impact on childhood mortality of interventions to improve drinking water, sanitation

and hygiene (WASH) to households: systematic review and meta-analysis," Mar. 14, 2023, *Infectious Diseases (except HIV/AIDS)*. doi: 10.1101/2023.03.13.23287185.
[25] J. H. Verbeek *et al.*, "Personal protective equipment for preventing highly infectious diseases due to exposure to contaminated body fluids in healthcare staff," *Cochrane Database Syst. Rev.*, vol. 2020, no. 5, May 2020, doi: 10.1002/14651858.CD011621.pub5.
[26] W. R. Shadish, T. D. Cook, and D. T. Campbell, *Experimental and Quasi-Experimental Designs for Generalized Causal Inference*. Boston, MA, USA: Houghton, Mifflin and Company, 2002.
[27] P. Craig *et al.*, "Using natural experiments to evaluate population health and health system interventions: new framework for producers and users of evidence," *BMJ*, vol. 388, p. e080505, Mar. 2025, doi: 10.1136/bmj-2024-080505.