nature human behaviour

Review article

https://doi.org/10.1038/s41562-024-01939-z

Causal inference on human behaviour

Received: 17 March 2023

Accepted: 27 June 2024

Published online: 23 August 2024

Check for updates

Drew H. Bailey[1], Alexander J. Jung[2], Adriene M. Beltz[3], Markus I. Eronen[4], Christian Gische[5], Ellen L. Hamaker[6], Konrad P. Kording[7,8], Catherine Lebel[9,10], Martin A. Lindquist[11], Julia Moeller[12], Adeel Razi[13,14,15,16], Julia M. Rohrer[17], Baobao Zhang[18] & Kou Murayama[2,19]

Making causal inferences regarding human behaviour is difficult given the complex interplay between countless contributors to behaviour, including factors in the external world and our internal states. We provide a non-technical conceptual overview of challenges and opportunities for causal inference on human behaviour. The challenges include our ambiguous causal language and thinking, statistical under- or over-control, effect heterogeneity, interference, timescales of effects and complex treatments. We explain how methods optimized for addressing one of these challenges frequently exacerbate other problems. We thus argue that clearly specified research questions are key to improving causal inference from data. We suggest a triangulation approach that compares causal estimates from (quasi-)experimental research with causal estimates generated from observational data and theoretical assumptions. This approach allows a systematic investigation of theoretical and methodological factors that might lead estimates to converge or diverge across studies.

Many human behaviours and experiences are difficult or impossible to manipulate in controlled settings, and yet their underlying causal mechanisms are at the heart of research questions in several fields, such as economics[1], epidemiology[2], political science[3,4], psychology and neuroscience[5–8], and sociology[9,10]. Advances in the study of causal inference have increased the potential for causally informative analyses from observational data of human behaviour[1,2,11,12]. Combining these with technological advances that have made longitudinal data more available[13] holds great potential for behavioural research.

In this Review, we bring together perspectives from a variety of fields. We describe issues of causal interpretation that are common to these disciplines, as well as available approaches to address them, with the aim of providing a non-technical, integrative conceptual overview (a glossary of key terms is provided in Box 1). Although many issues we discuss (such as lack of specificity in language) apply to causal inference in general, we focus on longitudinal data because scholars have made many exciting methodological advances in their analysis that may provide partial solutions. Even with longitudinal data, inferring causality is difficult—perhaps even more so than researchers imagine. In this Review, we hope to provide a potential future road map for researchers seeking to answer causal questions pertaining to human behaviour. We identify key challenges to causal inference on human behaviour

[1]School of Education, University of California, Irvine, Irvine, CA, USA. [2]Hector Research Institute of Education Sciences and Psychology, University of Tübingen, Tübingen, Germany. [3]Department of Psychology, University of Michigan, Ann Arbor, MI, USA. [4]Department of Theoretical Philosophy, University of Groningen, Groningen, the Netherlands. [5]Department of Psychology, Humboldt-Universität zu Berlin, Berlin, Germany. [6]Methodology and Statistics, Faculty of Social and Behavioural Sciences, Utrecht University, Utrecht, the Netherlands. [7]Department of Bioengineering, University of Pennsylvania, Philadelphia, PA, USA. [8]Department of Neuroscience, University of Pennsylvania, Philadelphia, PA, USA. [9]Alberta Children's Hospital Research Institute, Calgary, Alberta, Canada. [10]Department of Radiology, University of Calgary, Calgary, Alberta, Canada. [11]Department of Biostatistics, Johns Hopkins University, Baltimore, MD, USA. [12]Leipzig University, Leipzig, Germany. [13]Turner Institute for Brain and Mental Health, School of Psychological Sciences, Monash University, Clayton, Victoria, Australia. [14]Monash Biomedical Imaging, Monash University, Clayton, Victoria, Australia. [15]Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK. [16]CIFAR Azrieli Global Scholars Program, CIFAR, Toronto, Ontario, Canada. [17]Wilhelm Wundt Institute for Psychology, Faculty of Life Sciences, Leipzig University, Leipzig, Germany. [18]Maxwell School of Citizenship and Public Affairs, Syracuse University, Syracuse, NY, USA. [19]Research Institute, Kochi University of Technology, Kochi, Japan. e-mail: dhbailey@uci.edu

BOX 1

Glossary of key terms

Treatment or intervention (synonyms): Any deliberate change made to one or more variables in a model (often with the goal of uncovering causal effects).

(Randomized) experiment: A controlled study in which the treatment is allocated to the participants at random.

Quasi-experiment: A study in which naturally occurring events determine who receives a treatment (for example, country of birth or school of participants' choice).

Observational study: A non-experimental study (that is, one in which no random perturbations are given) that records naturally occurring phenomena.

Causal effect: The hypothetical difference in a variable *Y* given different conditions of an intervention *X* for a population (average causal effect), a part of a population (conditional causal effect) or a person (individual causal effect).

Estimand: The quantity (for example, a parameter) representing the (causal) effect of interest in a statistical analysis.

Estimate: The statistically estimated or calculated value derived from data to approximate the true value of an estimand.

Directed acyclic graph (DAG): A visual representation of the causal assumptions of a theory or model. Nodes in a DAG correspond to variables and are either connected by directed edges (indicating a causal influence) or not (indicating the absence of a causal influence). DAGs help to clarify whether possible covariates serve as mediators, confounders or colliders.

Mediator: If a treatment *X* causally influences *M* and *M* causally influences the outcome *Y*, then *M* is the mediator of the treatment. Whereas conditioning on *M* biases the estimate of the total effect, it might be necessary to condition on mediators when one is interested in estimating which path or mechanism is the reason for a causal effect.

Confounder: If a variable *C* causally impacts both the treatment *X* and the outcome *Y*, then *C* is a confounder for the treatment effect. Not conditioning the treatment effect on *C* biases the estimate.

Collider: If both the treatment *X* and the outcome *Y* causally influence a variable *C*, then *C* is a collider. Conditioning the treatment effect on *C* biases the estimate.

and provide tentative solutions for single challenges. Boxes 2–8 summarize each identified challenge, point to challenge-specific solutions and suggest literature for further reading. We further suggest how to deal with imperfect causal effect estimates in a more general manner.

Major challenges

Failing to ask appropriately specific research questions

It may sound obvious that causal inference can succeed only if we specify the causal effect of interest. However, an underappreciated reason why a large number of research questions have not yet been addressed using recent advances in the tools and language of causal inference is that researchers are often unclear (even to themselves) about what question their analysis is attempting to answer—specifically, whether the researchers are interested in estimating a causal effect at all. Although in some fields the default answer may be a clear 'yes' (for example, in economics[14]), researchers in other fields (for example, psychology) may show less consensus on this question.

BOX 2

Asking appropriately specific research questions

Unspecific research questions impede selecting the optimal design, models and techniques to answer them.

Tentative solutions

- Clarify the research question: do you attempt to estimate causal effects, and, if so, which kinds of effects?
- Can the question be answered by the data at hand?
- Can the estimate be bounded within a range of plausible values under realistic assumptions?

Literature suggestions

- Gelman and Imbens[15]
- Hamaker et al.[64]
- Hernán[18]
- Hernán and Robins[2]
- Lundberg et al.[65]

Let us start with the kinds of questions we hear every day in scientific settings and in our daily lives:

(1) 'Why are some people happier than others?'
(2) 'Does happiness predict a person's later health status?'
(3) 'What are the economic returns to completing an additional year of schooling?'

The ease with which these questions can be mapped onto well-defined causal effects varies. Question 1 is a natural and perhaps useful starting place. However, although it is unambiguously causal, it specifies no cause at all, resulting in a 'reverse' causal question[15] that cannot be directly addressed with methods intended to address forward causal inferences (what is the effect of *X* on *Y*?). Such a 'why' question is also sometimes called a 'backward causal question', because it starts with the effect and asks for a cause. Question 2 is framed as a statistical, not a causal, problem (does *X* predict *Y*?). The answer could be used for actuarial purposes when trying to forecast future health status or to select people for training or intervention. Only question 3 is both specific and (unambiguously) causal: the cause and effect of interest are reasonably clear.

Which question we ask has implications for data collection and analysis. A researcher interested in question 1 might be best served by conducting a literature review or, if there is little prior work, pursuing an exploratory line of research. A researcher interested in question 2 would not need to worry about confounding because the question is concerned only with prediction without further qualification; additional variables would be of interest only to improve the predictive accuracy of the model. But researchers often still attempt to control for third variables as confounders, which highlights the fact that such analyses are often associated with an underlying causal question, even if it is not explicated—a source of substantial confusion[16–18].

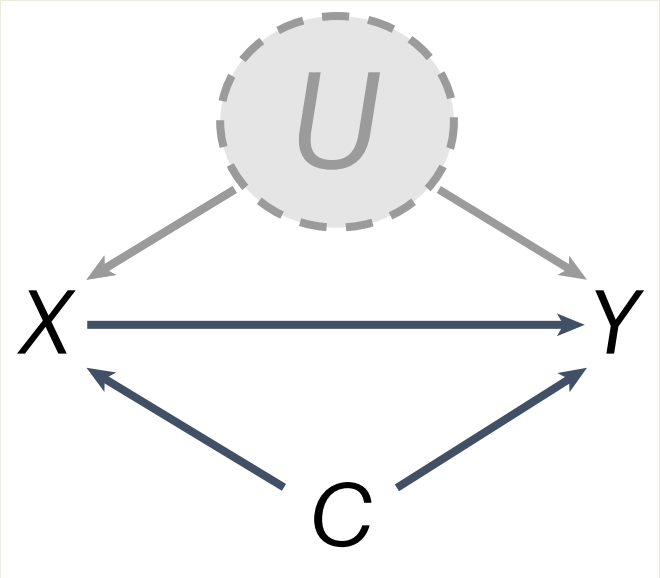
Under-control

Even clearly specified questions like question 3 do not guarantee clear causal answers. A researcher attempting to estimate the effect of happiness on health by statistically controlling for potential confounders (that influence both earlier happiness and later health) will have a difficult time convincing a reasonable sceptic. A positive estimate, a sceptic might argue, is plausibly still confounded by unmeasured common causes (for example, existing health conditions[19]). Omitted confounders are a problem for causal inference approaches that

https://doi.org/10.1038/s41562-024-01939-z

BOX 3

Under-control



Confounders bias the estimate of causal effects if not adequately controlled for. If a variable *C* (confounder) is known to cause both *X* and *Y*, then the estimate of the causal effect of *X* on *Y* will be biased if *C* is not adequately controlled for. Unknown or unobserved confounders (depicted as *U*) are a main threat to valid causal inference.

Example: The estimated effect of years of schooling (*X*) on earnings (*Y*) may be upwardly biased if childhood human capital (*C*) is not ruled out by the design or analysis.

Tentative solutions

- Clarify the causal model before data collection to assess potential confounders.
- Measure as many confounders as possible and adequately adjust for them when estimating the causal effect.
- Actively discuss assumptions under which your estimate approximates a causal effect.
- Consider a design-based solution or methods that control for unobserved confounders.

Literature suggestions

- Angrist and Pischke[48]
- Cinelli and Hazlett[22]
- Cinelli et al.[30]
- Pearl et al.[100]
- Rohrer[7]
- Usami et al.[84]

BOX 4

Over-control

Controlling for colliders biases the estimate of the causal effect. For example, if *X* and *Y* both cause another variable *C* (collider), conditioning on *C* biases the estimate of the causal effect of *X* on *Y* (even if *X* and *Y* are unconditionally independent).

If the causal estimand is a total effect, controlling for mediators biases the estimate. For example, if the effect of *X* on *Y* is mediated by *M*, conditioning on *M* biases the estimate of the total effect of *X* on *Y*.

Example: The estimated effect of years of schooling (*X*) on earnings (*Y*) may be downwardly biased if adult neighbourhood (*C*) or post-schooling occupation (*M*) is statistically controlled.

Tentative solutions

- Clarify the causal model to decide which variables (not) to control for.
- Do not merely control for all covariates in your dataset.
- Actively discuss assumptions under which your estimate approximates a causal effect.

Literature suggestions

- Achen[101]
- Cinelli et al.[30]
- Elwert and Winship[32]
- Greenland et al.[27]
- Hoyle et al.[29]
- Wysocki et al.[31]

rely on third-variable adjustment (as opposed to well-implemented design-based approaches[1]). In the study of human behaviour, individuals may anticipate what they expect to happen, and many behaviours or emotional reactions[20] precede the anticipated events. For longitudinal research on causal mechanisms affecting human behaviour, anticipated events may confound prospective associations between psychological variables such as happiness and subsequent outcomes.

One may argue that, as long as researchers control for 'major' sources of confounding, they can obtain relatively accurate causal estimates. In principle, for some types of data and some kinds of research questions, major confounders may be limited in number, but we suspect that this is rarely the case for observational studies of human behaviour, as it is often complex and influenced by numerous factors. At the least, there is no way to prove that all sources of confounding have been controlled for. There have been attempts to identify most relevant confounders using a robustness check approach[21]; however, measured confounders are often limited by design. As can be seen in various formulas that quantify the bias due to multiple omitted confounders[22], the cumulative impact of those omitted confounders (including their higher-order interactive effects) could indeed rarely be ignorable. This may explain some of the widespread contradictions between findings from experimental and observational studies across fields[5,23,24].

In addition, even when confounders are known, the quality of their measurement can limit causal inferences. For example, when health recommendations change, the associations between behaviours labelled as potentially harmful or helpful (for example, vitamin consumption) and other health-related behaviours (for example, smoking) and outcomes (for example, heart health) sometimes change as well, which suggests that more health-conscious individuals are more likely to respond to health recommendations[25]. But statistically controlling for measures of demographics, socio-economic status and other health-related behaviours is not sufficient to eliminate the time-varying association between labelled behaviours and health-related outcomes. This suggests that 'health-consciousness' cannot be sufficiently captured by these measures—and thus, we cannot fully adjust for them.

Finally, even if researchers include sufficient measures of the right confounders, causal estimates can still be biased if the model is misspecified—for example, if it fails to account for measurement error or inadequately captures the function through which changes in the confounder affect the causal variable of interest and the outcome[26].

BOX 5

Effect heterogeneity

The effect of a treatment *X* on the outcome *Y* may differ between individuals, time points, contexts and other conditions. Average treatment effects provide estimates for central trends that may fail to describe many of the individuals in a heterogeneous population. Systematic sampling and systematic analysis of heterogeneity can enrich our understanding of causal effects in heterogeneous populations.

Tentative solutions

- Systematically sample heterogeneous populations.
- Use statistical methods that detect heterogeneity.
- Whenever possible, test assumptions of homogeneity before applying them.
- Actively discuss assumptions under which your estimate approximates an average causal effect.

Literature suggestions

- Athey and Imbens[102]
- Bryan et al.[34]
- Geng et al.[103]
- Gische et al.[80]
- Moeller[104]
- Montoya et al.[59]
- Pearl and Bareinboim[105]
- Wager and Athey[106]

Over-control

A different sceptic might argue that a reported smaller-than-expected estimate of the effect of earlier happiness on later health has resulted from statistical over-control. Over-control happens when researchers (wrongly) control for the variables that are the causal consequence of the causal variable of interest, resulting in post-treatment bias[27–29]. Post-treatment bias can arise from conditioning on mediators (variables caused by the causal variable of interest that affect the outcome of interest) or colliders (variables influenced by the causal variable of interest and the outcome).

Using the previous example of the causal effect of happiness on health, imagine that a researcher controlled for sleep quality (assessed at the same time as happiness). Statistical control for sleep quality may reduce the estimated effect of happiness on health. But if happiness truly causally influences sleep quality, which in turn influences health (rather than sleep quality influencing both happiness and health), the researcher has controlled for a mediator—a pathway through which happiness might affect health. Thus, they underestimate the true effect of happiness on health. The issue is further complicated because sleep quality may also be a confounder that affects both happiness and health; thus, if we do not adjust for it, we may overestimate the true effect of happiness on health. It is thus unclear whether it is a 'good' or 'bad' control[5,30,31] because of uncertainty about the causal order of variables: plausibly, sleep quality is both a confounder and a mediator and thus cannot be used to recover the effect of happiness on health without additional information.

Furthermore, if the researcher controls for a factor plausibly influenced by both health and happiness, this may result in collider bias[32]. Such 'control' can even occur without explicit statistical control if the sample is selective. If both happiness and health make it more likely that an individual participates in the study, study participation becomes a collider. In such a scenario, even if happiness causally affects health, one may fail to detect such an effect or underestimate its magnitude.

A well-known real-world example of collider bias comes from research using administrative data from police stops to estimate the magnitude of racial discrimination in police officers' use of force. These studies are based on the following logic. If officers are more likely to use force in otherwise similar situations (with regard to suspect behaviour) on members of one group than on members of another, this would be a sign of discrimination. Discriminatory use of force implies a causal effect of the suspect's perceived race (*X*) on the use of force (*Y*) that is not mediated via suspect behaviour (*M*). However, in reality, perceived race directly influences officers' decisions to stop suspects in the first place. Being stopped is thus a collider, influenced by both perceived race (*X*) and criminal activity (*M*). Members of the racial group discriminated against may thus be on average less likely to be engaged in criminal activity when they are stopped. Paradoxically, in the presence of discriminatory stopping but not discriminatory use of force, we might expect a lower rate of use of force against the group stopped more, conditional on being stopped. Thus, relying on data from stops alone to attempt to adjust for suspect behaviour may lead us to underestimate racial discrimination in police officers' use of force[33].

Effect heterogeneity

Many popular causal inference techniques from observational data are useful for estimating the average causal effect. A focus on average causal effects is convenient because they provide a simple summary of causal effects, averaged across time and individuals within a sample; and under certain circumstances, which need to be tested, they allow for generalizations to the population. Furthermore, averaging can be useful to estimate causal effects, because counterfactuals for a specific individual cannot be directly observed. However, people are unique and change over time[34,35], and such heterogeneity is easily overlooked if only a single average treatment effect is examined. A failure to appropriately model the heterogeneity of causal effects across covariates, across individuals or across time can even lead to biased estimates of an average causal effect under some reasonable conditions[36–38], particularly in the absence of a randomized experiment.

A practical implication of effect heterogeneity is that it makes the interpretation of a new finding that conflicts with the previous literature challenging. In the absence of a strong causal design, researchers cannot tell whether the unexpected finding originates from bias (a lack of internal validity), heterogeneity of the causal effect of interest (a lack of external validity) or random variation. Consider a hypothetical finding where, among individuals with a severe medical diagnosis, optimism is more strongly predictive of subsequent mortality than in other populations. A possible explanation for this finding is that the causal effect of optimism is moderated by one's condition, such that it is particularly important to stay positive when one's health is compromised. Another is that there is no such differential effect of optimism on health in this population, but rather unmeasured prognostic details in the patient population affect both current levels of optimism and later mortality. The theoretical underdetermination of why conflicting findings arise complicates the process of attempting to make cumulative progress in the study of human behaviour.

Interference

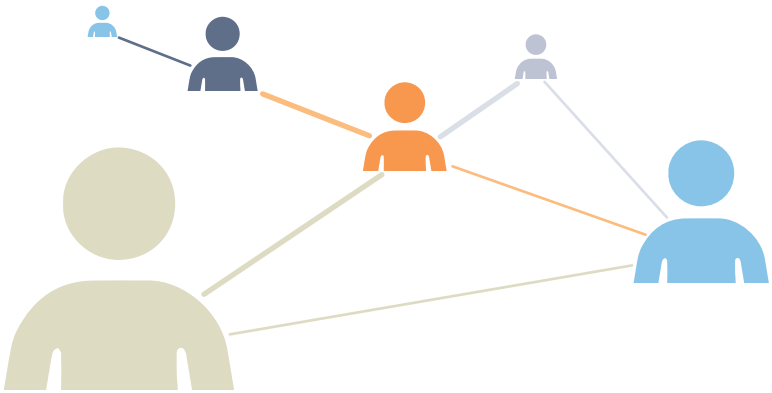
Many causal inference techniques assume the absence of interference (also referred to as spillover effect or contamination), meaning that the potential outcome of one individual is not affected by the causal variable of interest for other individuals[39]. However, when observational data are collected in a sample of people who could potentially interact with each other (for example, data from school or other social settings), there is a good possibility that the assumption is violated, in that the treatment of interest affects not only treated individuals but also untreated individuals used to approximate the counterfactual.

For example, if we attempt to estimate the effect of years of schooling on later earnings by comparing children with more years of schooling to their siblings with fewer years of schooling (whose outcomes

https://doi.org/10.1038/s41562-024-01939-z

BOX 6

Interference/spillover



A standard assumption of most approaches to causal inference is that there is no interference or spillover. That means that the potential outcome of an individual is not affected by the treatment status of other individuals. However, in some practical applications, individuals participating in a study do interact or communicate with each other, which increases the risk of potential spillovers.

Tentative solutions

- Consider analysis- or design-based solutions.
- Conduct robustness checks with subgroups or settings under which spillovers are unlikely.
- Consider alternative estimands that explicitly take interference into account.

Literature suggestions

- Benjamin-Chung et al.[107]
- Hudgens and Halloran[108]
- Imai et al.[109]
- Tchetgen and VanderWeele[110]
- Zhang et al.[111]

may approximate the counterfactual had their siblings not received additional schooling), an important assumption is that one's earnings are not affected by one's sibling's schooling. However, research on the effect of people's education on their siblings' subsequent educational choices or performance suggests that spillovers are common—siblings' educational experiences appear to affect each other[40–42]. If individuals influence each other, an individual influenced by the causal variable of interest, X, subsequently influences others with different values of X, complicating the accurate estimation of causal effects on Y; associations between X and Y are likely to be biased even when they are adjusted for confounders.

Timescales of effects

We have already argued that effects can vary across individuals, and average and individual effects can also vary substantially across timescales. An estimate of an intervention effect may be specific to the time at which the intervention occurs and when the outcome is measured[35]. Having an alcoholic drink is likely to increase one's self-confidence for a short period; not finding an effect on self-confidence measured a week (or even years) after having a drink does not indicate that alcohol does not influence self-confidence. The effects of a different intervention (for example, the effects of parenting on subsequent illegal behaviour), however, may only show up after long periods; conversely, daily diary or ecological momentary assessments would not be useful to assess such effects directly, although they may offer insights into some of the intervening processes. How to connect the various timescales, and how to make sense of this, poses a major challenge[43], particularly when behaviours affect each other bidirectionally[44].

Fat-handed and other complex interventions

Returning to the question of the effects of happiness on health, let us assume that a researcher was able to address these issues and obtained an unbiased causal estimate of the effect of a year-long sustained increase in happiness on various health outcomes years later. Using this estimate to gauge the benefits of various interventions to increase health still would not be straightforward. The problem is that there are different ways to change happiness (for example, a self-help workshop, exercise or a magic happiness pill), and the hypothesized causal consequences of happiness on health could be different depending on the intervention implemented. One school of thinking about causality refers to this as a lack of consistency; in this line of reasoning, asking for the effects of happiness on health would not be a well-defined question to begin with, as happiness is not (directly) manipulable[45,46]. Another school of thinking about causality maintains that such effects can be well defined[47], but that still leaves it open that the intervention used to induce happiness interacts with the effects of happiness on health. In any case, moving from causal effects of variables that cannot be directly manipulated to interventions would not be straightforward.

Moving the other way around—from interventions to conclusions about the effects of variables that cannot be manipulated directly—is equally challenging. For example, to estimate the effect of earlier happiness on later health, perhaps a researcher conducts a randomized

BOX 7

Fat-handed and other complex interventions

Sometimes interventions alter not only the treatment *X* but also other variables. If we intervene on *X* to see whether it is a cause of *Y*, an intervention that not only changes *X* but also changes some other causes of *Y* (that are not on the path from *X* to *Y*) is called a fat-handed intervention. For example, an intervention on a person's daily diet will change the daily calorie intake (treatment) but might also change the daily intake of vitamins (other variable). Both calorie intake and intake of vitamins affect the body mass index (outcome). The intervention on daily diet is thus fat handed.

Tentative solutions

- Consider triangulating across various interventions with different likely side effects.
- Measure and report potential side effects of your interventions.
- When possible, intervene on variables that are well defined and isolatable from other variables.

Literature suggestions

- Eberhardt and Scheines[112]
- Eronen[49]
- Mooij et al.[113]
- Peters et al.[114]
- Scheines[50]
- VanderWeele[26]

BOX 8

Timescales of effects

The latency of an intervention on an effect may be unclear. If the effect of *X* on *Y* has a latency of several years, it will not show when the time interval between measurements is just a few months. Moreover, effects may be non-stationary, meaning they can change over time, whereas most estimates of time effects assume stationarity and overlook the fact that over time, an effect can emerge, speed up, slow down or cease to exist.

Tentative solutions

- Consider the potential timescale of effects before data collection to measure causes and effects at appropriate intervals.
- Consider continuous time modelling.

Literature suggestions

- Aalen et al.[115]
- Driver and Voelkle[116]
- Røysland[117]
- Ryan and Hamaker[118]
- Voelkle et al.[56]

controlled trial of an intervention that targets participants' happiness directly, but not their health. In this case, the impact of the intervention on later health could provide a strong test of the causal effects of happiness on health (in such a design, assignment to the intervention would be an 'instrumental variable' for estimating the effect of happiness on health[48]). However, this would depend critically on the assumption that the intervention did not also directly influence other factors (for example, health-related motivation) that increase later health (in the instrumental variables framework, this is called the exclusion restriction assumption). Interventions that violate this assumption and thus change variables beyond the causal variable of interest are labelled 'fat-handed' interventions[49,50]. Interventions that target psychological constructs such as happiness are extremely likely to be fat-handed: how could one intervene on just happiness without also changing related psychological constructs, such as positive affect or elevated mood? One source of this problem is that constructs in the behavioural sciences are often unclearly or ambiguously defined, or conceptually overlap with other constructs[51], making them challenging for causal inference[52]. This makes it difficult to test the causal effects of such variables even when randomized controlled trials are feasible (for example, did the intervention affect health via happiness or via frequently elevated mood?). Such effects of constructs (rather than well-defined interventions) also have important practical implications: if happiness is hypothesized to mediate the effects of an intervention on health over a long period, it is far more efficient to optimize iterations of the intervention to generate larger impacts on happiness than on health. However, if the wrong mediator is identified, this strategy may not work.

Summary and implications of challenges

The major challenges above all come down to human behaviour being embedded in complex causal systems of which we can observe only a limited part. This leaves researchers in a difficult position. Sceptics with strong priors on either side of a scientific disagreement can offer plausible reasons why any study or set of similar studies presenting evidence against their position should not convince them of the opposite view. Observational studies under-control (or over-control) for various factors influencing the outcome, experiments directly influence more (or less) than the causal variable of interest and causal effects could apply to a small subgroup or no individual at all, limiting the usefulness of such estimates.

Tentative solutions

Given the challenges to drawing causal conclusions from longitudinal data outlined above, what can be done? Recent years have brought many potential (if partial) solutions. For under-control, there are methods that attempt to control for unobserved time-invariant confounders with the help of longitudinal data[53–55]. We may be able to get a more informative causal estimate for different timescales using continuous time modelling[56]. Effect heterogeneity in longitudinal data could be partially tackled using methods that stratify on post-treatment variables[57], through the estimation of nomothetic effects in bottom-up procedures by empirically identifying those patterns that generalize across person-specific effects (for example, group iterative multiple model estimation[58]) or through dynamic treatment regimes[59]. Methods for estimating precise causal effects for various data structures such as panel designs, intensive longitudinal data and neuroimaging have also been proposed[60–63].

However, in the absence of a strong causal design, it is often difficult to establish whether such solutions have generated unbiased estimates. The capacity of causal inference from observational data is substantially determined by how we design the research, not solely by advanced statistical techniques. We offer several potential approaches below.

Clarifying the research question

Perhaps the lowest-hanging fruit for improving inference is to improve the clarity of the research question, which may often involve narrowing its scope. Such clarity helps readers, but most importantly the researchers themselves, to determine (1) whether the purpose of the study is to estimate causal effects at all, (2) which kinds of effects the researcher is interested in[64–66] and (3) whether the question asked can be answered by the data at hand.

An important example that demonstrates the need for well-defined research questions comes from the Many Analysts Project[67]. The authors asked researchers to reanalyse a dataset to answer the question of whether soccer referees are more likely to give red cards to players with a dark skin tone. Worryingly, research teams obtained a wide range of effect sizes using the same data. However, Auspurg and Brüderl[68] observed that research teams appeared to interpret the research question differently, with some teams attempting to answer the descriptive question of whether players with a darker skin tone were more likely than lighter-skinned players to receive a red card, but other research teams inferring a more complex research question about the causal mechanisms through which such a difference might occur: "the *direct causal effect* of skin tone that remains after netting out confounders and 'productivity-relevant' mediators"[68]. When analyses were restricted to the latter category, the variability of estimates was much narrower.

In the previous case, the study would have benefited from more specificity in the research question. A clearly specified cause, effect, timescale and population of interest are necessary to provide credible estimates of causal effects. At the same time, it is important to keep in mind that the research question of interest may be broader, with individual estimands only providing partial answers. Likewise, there is still a space in science for backward causal questions (such as 'Why are some people healthier than others?') that often motivate the quest for answers to narrower forward causal questions. Such questions have an important role in human thinking and decision-making and can be useful for checking the assumptions of causal models and revising theories (Gelman and Imbens[15] discuss them and give examples). For example, noticing patterns across space and time in which people become ill may be useful for forming theories about why some people are healthier than others, which can be further tested with stronger causal designs. Furthermore, a researcher who has considered the backward question, 'Why do students in the same school frequently have such different levels of academic achievement?' might be less likely to mistakenly attribute test score differences to the causal effects of specific school or teacher characteristics in the absence of a strong set of baseline statistical controls. Considering more general questions is useful for thinking about generalizability, mechanisms, alternative hypotheses, and matches between data and model assumptions and thus should be considered in service of asking better forward causal questions and for better assumption-checking when trying to answer them.

Choosing the right design and variables

Once a causal question is set, we need to think carefully about the right design to permit appropriate causal inference[69]. This includes the decision about potential controlling variables to be measured, the time interval of measurement, the number of time points, sample characteristics, sample size and so on. Tools that help with such decision-making processes are increasingly available[7]. Researchers must pay careful attention to the measurement of constructs in the study of human behaviour (for example, depression and poverty), which pose unique challenges related to reliability and validity, along with conceptual challenges (for example, whether measurements are better understood as effects or causes of the construct of interest), all of which can affect causal inferences[26,70,71].

Actively discussing the causal informativeness of a model

We argue that observational work should be discussed in terms of its implications for causal theories, where the causal implications are judged in a continuous, not a binary manner. A current norm in many fields studying human behaviour is to deliberately avoid making causal statements if the study lacks a randomized experimental design[18,72]. Although researchers should attempt to appropriately calibrate the strength of their claims to the strength of the evidence, failing to consider the causal implications of one's model (even a tentative one) may make it easier to overlook indications that assumptions are violated while probably conveying a causal interpretation to the reader anyway[16,17]. For example, consider a hypothetical study that regresses earnings on measures of previously measured cognitive test scores, social skills and demographic controls and interprets the implications of such findings for the (presumably causal) 'importance' of cognitive and social skills. Although the analysis is at risk for bias from all the potential factors listed in the previous section, the analyst might dismiss a critique about potential bias from the exclusion of other potential confounders from the model under the grounds that the analysis uses non-experimental data and is thus not 'causal'. However, this argument conflates the goal of the study (causal estimation) with the degree of certainty the method affords in its findings, paradoxically missing an opportunity to improve the analysis. Researchers should be able to communicate the goal of the study (for example, causal estimation) and the certainty with which estimates are causally informative (for example, low in an observational, cross-sectional setting; higher in a setting that rests on fewer assumptions) separately, rather than lumping them together under the single term 'causal' (causal estimation with high certainty). Presenting the study as a rough attempt at bounding causal estimates of the returns to cognitive and social skills would allow researchers to convey a great deal of uncertainty while still directly addressing the causal questions at hand. An approach to rule out different threats to causal inference of estimates could be to use a range of estimation strategies. Deming[73], Pion and Lipsey[74], and Ritchie and Tucker-Drob[75], for example, compare results across multiple identification strategies.

Triangulating on causal estimates

Clarifying research questions is a necessary but far-from-sufficient solution to identifying causal effects of constructs, particularly when only fat-handed interventions or observational data are available. How might a researcher interested in the causal effects of such a construct proceed, knowing that any finding can be dismissed for statistically over- or under-controlling or including an overly broad or overly narrow experimental manipulation, depending on the observed effect? Like others before[76–78], we argue that successful examples from the study of human behaviour triangulate a causal effect (or set of causal effects) of interest by considering multiple findings using various identification strategies and reasoning through them in the context of tentative theories and assumptions (Fig. 1). Triangulation may happen on different levels, across studies in the field as a whole, within a particular research programme (and thus potentially across studies) and within individual studies.

Researchers may consider two types of sources of evidence: (1) strong causal estimates of well-defined (but probably fat-handed) interventions (broadly defined, including both applied interventions and exogenous manipulations of behaviour designed solely for the purpose of improving causal understanding) and (2) non-experimental estimates from observational data. Estimates should be compared with each other whenever possible and interpreted considering theory and assumptions. Combinations of theories and assumptions that accommodate existing estimates are re-examined, with a focus on research questions and designs that can plausibly distinguish among them[78,79]. We describe an example below.

A researcher interested in the effects of a complex or fat-handed intervention on a set of outcomes might approach the problem from several different perspectives. For example, a researcher interested in the effects of cognitive skills on later educational and occupational success might be interested in this problem so that they can (1) better understand the causal dynamics between cognitive skills and educational inputs and/or (2) design interventions to better prepare students for educational or occupational success.

A researcher following the first approach might consider a large longitudinal dataset (Fig. 1, left box) and use recent advances in longitudinal data analysis techniques to address confounding via matching

and latent variable modelling[56,80–82], justifying the use of statistical controls from a causal inference perspective. They might test whether their conclusions are robust over variations in model specification[83,84] and assumptions about how cognitive skills produce variation in test scores[70,85]. They might attend to whether estimates are similar across a range of subgroups and time periods as well.

In many cases, estimates will vary, sometimes predictably across specifications (Fig. 1, bottom arrow). Even if estimates fall within a narrow range of values, they might differ across individuals, subgroups, measures and outcomes. The researcher might attempt coherent pattern matching[69], where in the absence of an intentional intervention, the researcher tests a list of theory- and assumption-driven predictions and calibrates the strength of the match between theory and results. The researcher might use robustness checks or falsification tests to attempt to quantify the amount of remaining bias in the estimate of interest. For example, concerned about bias in the estimated effects of corporal punishment on children's later antisocial behaviour, Larzelere and colleagues[86] compared estimated effects of corporal punishment to the estimated effects of other interventions not hypothesized to increase antisocial behaviour (for example, grounding or psychotherapy). Finally, the researcher might include some bounding exercise that allows for a lower bound of plausible estimates[87], such as a sensitivity analysis testing the magnitude of unobserved confounders necessary to produce the estimated effect[22].

In contrast, a researcher who is interested in developing an effective intervention on later educational or occupational success might approach the problem by reviewing the literature on field experiments or strong quasi-experimental evaluations of exogenous factors found to improve students' cognitive skills (Fig. 1, top box). Interventions are likely to vary in their fat-handedness (the extent to which they influence education via pathways other than students' cognitive skills alone). For example, a brief cognitive training intervention might plausibly influence later outcomes primarily via changes to cognitive skills. However, a broader intervention, such as an additional year of schooling, includes exposure to different mentors, peers and other social contexts over an extended period and is much more fat-handed (Fig. 1, right arrow). Thus, the researcher might consider whether within and across categories of interventions, subgroups and outcomes, improvements to cognitive skills are reliably related to improvement in educational or occupational success[88].

Finally, both researchers might attempt to obtain estimates of the effects of cognitive skills on educational or occupational success on the same scale (for example, what is the estimated effect of a hypothetical minimally invasive intervention that influences cognitive skills at age 18 by half a standard deviation on wages at age 35?) so they can be compared (Fig. 1, left arrow). In the presence of an exogenous intervention, the researcher might be able to systematically compare estimates of the causal effect of interest using an experimental design and an observational analysis in the same sample. This method, sometimes called a within-study comparison, sometimes finds converging[89,90] and other times diverging results[91]. Key to attempting to reconcile these estimates is a judgement about the likely net bias of both of them: if statistical controls are poor and interventions are all substantially fat-handed, for example, perhaps the net bias of both kinds of studies will go in the same direction[19].

If estimates differ substantially across methods, several possible explanations should be considered: Are the meanings of outcome scores measured in observational studies qualitatively different from the outcome scores in stronger causal designs, despite sharing the same label and even the same measure[92]? Are the statistical controls in observational studies likely to miss plausible confounders[93]? Are effects heterogeneous across samples and settings, which differ systematically across timescales and designs? If so, are estimates different primarily because they are unbiased estimates of different effects, or because they are estimates of the same underlying effect but are differently biased? Is there a structural model that can reconcile these findings, and if so, what predictions would it make about the results of future observational or experimental studies (Fig. 1, right box)? A large number of potential explanations can account for non-overlapping estimates[76,94]. Importantly, though, triangulating between experimental and observational analyses makes it possible to probe more of them, relative to an approach that views observational and experimental research as irreconcilable approaches to studying fundamentally different kinds of questions[79].

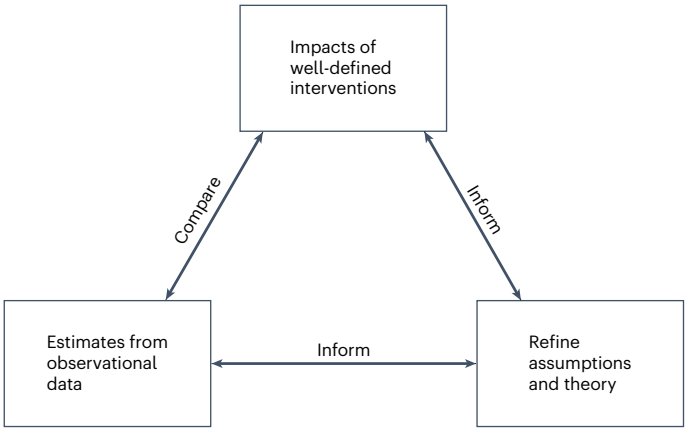


Fig. 1 | A framework for triangulating on causal effects. Designs of intervention and observational studies should be theory informed, and the results of interventional and observational studies need to be compared as well as to be interpreted in light of and to inform theory.

Several scientific practices enacted by researchers, but perhaps incentivized by systems, could enhance the efficiency of this process. In some cases, it may be possible a priori to design a research programme where a research question is well defined and constructs are operationalized sufficiently similarly across studies that estimates are comparable and reconcilable after they are conducted. For example, systematically exploring whether heterogeneity can account for apparently different findings would require larger, more systematic samples across both experimental and non-experimental designs (many such corresponding research programmes have been undertaken in recent years[34,95–97]). Meta-analytic models can incorporate estimates obtained from multiple identification strategies, samples and outcomes to attempt to identify sources of heterogeneity of estimates.

Many of the methods above can be integrated with a structural model—a formal model of behaviour that makes predictions about the results of future experiments[98]. For example, Todd and Wolpin[99] proposed and estimated a formal model of how children choose to attend school each year through high school, using data from the control group of a randomized experiment in which children and families in the treatment group received nutritional supplements and a subsidy conditional on children attending school. They showed that the subsidy increased school attendance by the amount predicted by their model, demonstrating its potential usefulness for both improving our understanding of why children (do not) attend school and forecasting the impacts of potential policies on school attendance in the future. We view such formalizations of causal models that make accurate predictions about the effects of future interventions as a worthy goal for the social sciences to aspire to. However, we note that a useful early step for many social scientists will be to understand that estimates yielded by different analyses can be mutually informative, either because they approximate the same estimand or because they approximate estimands that have a theoretically important relation to each other that can be used to triangulate on the estimand of interest. For example, a researcher studying the effects of different implementations of the same psychotherapy programme might catalogue the features of the

evaluation design, participant and programme features (for example, the baseline symptomology of different samples or the dosage of different implementations), and outcomes (for example, whether they are self-reported or tracked by administrative data) associated with larger or smaller estimated effects. Understanding what set of plausible explanations best account for variation in estimates may be useful for both understanding what features to include in a structural model and guiding future experimental efforts.

Triangulation will not necessarily provide a simple satisfactory answer to the broad question, 'What are the effects of cognitive skills on later educational and occupational success?' or the more specific question, 'What would be the effect of a hypothetical minimally invasive intervention that influences cognitive skills at age 18 by half a population standard deviation on wages at age 35?' But it might yield potentially useful answers to more specific causal questions, along with a broader model for predicting the impacts of future similar interventions. These predictions might be useful for improving the effectiveness of future interventions[80] and individualizing treatments[34,59,100] and for making testable predictions about underlying data-generating processes. Under such an approach, evidence from a wide variety of sources is admissible, and no single study is definitive. Estimates—even for a reasonably well-defined research question—can vary within and across studies for a large range of reasons reviewed above, and more estimates allow for a more comprehensive model of the underlying causal process(es) of interest.

Conclusion

Causal inference is hard, complicated by many logical problems and methodological challenges. Appreciating them is crucial, for we can overcome only the obstacles we are aware of. Yet, causal inference is at the heart of the study of human behaviour. Giving up on the goal of ascertaining causal mechanisms is therefore not an option. Recently, there have been substantial methodological advancements and debates about methods for causal inference. Studies attempting causal inference can now integrate the available tools from different disciplines and triangulate across the theories, methods and findings they have generated; this interdisciplinary dialogue seems vital to us.

References

1. Angrist, J. D. & Pischke, J.-S. The credibility revolution in empirical economics: how better research design is taking the con out of econometrics. *J. Econ. Perspect.* **24**, 3–30 (2010).
2. Hernán, M. A. & Robins, J. M. *Causal Inference: What If* (Chapman & Hall/CRC, 2020).
3. Aronow, P. M. & Miller, B. T. *Foundations of Agnostic Statistics* (Cambridge Univ. Press, 2019).
4. Keele, L. The statistics of causal inference: a view from political methodology. *Polit. Anal.* **23**, 313–335 (2015).
5. Foster, E. M. Causal inference and developmental psychology. *Dev. Psychol.* **46**, 1454–1480 (2010).
6. Marinescu, I. E., Lawlor, P. N. & Kording, K. P. Quasi-experimental causality in neuroscience and behavioural research. *Nat. Hum. Behav.* **2**, 891–898 (2018).
7. Rohrer, J. M. Thinking clearly about correlations and causation: graphical causal models for observational data. *Adv. Methods Pract. Psychol. Sci.* **1**, 27–42 (2018).
8. Rigoux, L. & Daunizeau, J. Dynamic causal modelling of brain–behaviour relationships. *NeuroImage* **117**, 202–221 (2015).
9. Gangl, M. Causal inference in sociological research. *Annu. Rev. Sociol.* **36**, 21–47 (2010).
10. Winship, C. & Morgan, S. L. The estimation of causal effects from observational data. *Annu. Rev. Sociol.* **25**, 659–706 (1999).
11. Imbens, G. W. & Rubin, D. B. *Causal Inference in Statistics, Social, and Biomedical Sciences* (Cambridge Univ. Press, 2015).
12. Pearl, J. *Causality: Models, Reasoning, and Inference* 2nd edn (Cambridge Univ. Press, 2009).
13. Hamaker, E. L. & Wichers, M. No time like the present. *Curr. Dir. Psychol. Sci.* **26**, 10–15 (2017).
14. Angrist, J. D. & Pischke, J.-S. *Mostly Harmless Econometrics: An Empiricist's Companion* (Princeton Univ. Press, 2009).
15. Gelman, A. & Imbens, G. *Why Ask Why? Forward Causal Inference and Reverse Causal Questions* Working Paper No. 19614 (NBER, 2013).
16. Alvarez-Vargas, D. et al. Hedges, mottes, and baileys: causally ambiguous statistical language can increase perceived study quality and policy relevance. *PLoS ONE* **18**, e0286403 (2023).
17. Haber, N. A. et al. Causal and associational language in observational health research: a systematic evaluation. *Am. J. Epidemiol.* **191**, 2084–2097 (2022).
18. Hernán, M. A. The C-word: scientific euphemisms do not improve causal inference from observational data. *Am. J. Public Health* **108**, 616–619 (2018).
19. Rohrer, J. M. & Lucas, R. E. Causal effects of well-being on health: it's complicated. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/wgbe4 (2020).
20. Hoemann, K., Devlin, M. & Barrett, L. F. Comment: emotions are abstract, conceptual categories that are learned by a predicting brain. *Emot. Rev.* **12**, 253–255 (2020).
21. Young, C. & Holsteen, K. Model uncertainty and robustness: a computational framework for multimodel analysis. *Sociol. Methods Res.* **46**, 3–40 (2017).
22. Cinelli, C. & Hazlett, C. Making sense of sensitivity: extending omitted variable bias. *J. R. Stat. Soc. B* **82**, 39–67 (2020).
23. Branwen, G. How often does correlation = causality? *Gwern.net* https://www.gwern.net/Correlation (2022).
24. Runge, J. Causal network reconstruction from time series: from theoretical assumptions to practical estimation. *Chaos* **28**, 075310 (2018).
25. Oster, E. Health recommendations and selection in health behaviors. *Am. Econ. Rev. Insights* **2**, 143–160 (2020).
26. VanderWeele, T. J. Constructed measures and causal inference: towards a new model of measurement for psychosocial constructs. *Epidemiology* **33**, 141–151 (2022).
27. Greenland, S., Judea, P. & Robins, J. M. Causal diagrams for epidemiologic research. *Epidemiology* **10**, 37–48 (1999).
28. Rosenbaum, P. R. From association to causation in observational studies: the role of tests of strongly ignorable treatment assignment. *J. Am. Stat. Assoc.* **79**, 41–48 (1984).
29. Hoyle, R. H., Lynam, D. R., Miller, J. D. & Pek, J. The questionable practice of partialing to refine scores on and inferences about measures of psychological constructs. *Annu. Rev. Clin. Psychol.* **19**, 155–176 (2023).
30. Cinelli, C., Forney, A. & Pearl, J. A crash course in good and bad controls. *Sociol. Methods Res.* https://doi.org/10.1177/00491241221099552 (2022).
31. Wysocki, A. C., Lawson, K. M. & Rhemtulla, M. Statistical control requires causal justification. *Adv. Methods Pract. Psychol. Sci.* **5**, 251524592210958 (2022).
32. Elwert, F. & Winship, C. Endogenous selection bias: the problem of conditioning on a collider variable. *Annu. Rev. Sociol.* **40**, 31–53 (2014).
33. Knox, D., Lowe, W. & Mummolo, J. Administrative records mask racially biased policing. *Am. Polit. Sci. Rev.* **114**, 619–637 (2020).
34. Bryan, C. J., Tipton, E. & Yeager, D. S. Behavioural science is unlikely to change the world without a heterogeneity revolution. *Nat. Hum. Behav.* **5**, 980–989 (2021).
35. Haslbeck, J. M. B. & Ryan, O. Recovering within-person dynamics from psychological time series. *Multivar. Behav. Res.* **57**, 735–766 (2022).

36. Goldsmith-Pinkham, P., Hull, P. & Kolesár, M. *Contamination Bias in Linear Regressions* Working Paper No. 30108 (NBER, 2022).
37. Goodman-Bacon, A. Difference-in-differences with variation in treatment timing. *J. Econ.* **225**, 254–277 (2021).
38. Wu, W., Carroll, I. A. & Chen, P.-Y. A single-level random-effects cross-lagged panel model for longitudinal mediation analysis. *Behav. Res Methods* **50**, 2111–2124 (2018).
39. Rubin, D. B. Causal inference using potential outcomes. *J. Am. Stat. Assoc.* **100**, 322–331 (2005).
40. Altmejd, A. et al. O brother, where start thou? Sibling spillovers on college and major choice in four countries. *Q. J. Econ.* **136**, 1831–1886 (2021).
41. Heckman, J. & Karapakula, G. *Intergenerational and Intragenerational Externalities of the Perry Preschool Project* Working Paper No. 25889 (NBER, 2019).
42. Karbownik, K. & Özek, U. *Setting a Good Example? Examining Sibling Spillovers in Educational Achievement Using a Regression Discontinuity Design* Working Paper No. 26411 (NBER, 2019).
43. Bringmann, L. F. et al. Psychopathological networks: theory, methods and practice. *Behav. Res Ther.* **149**, 104011 (2022).
44. Dietrich, J., Schmiedek, F. & Moeller, J. Academic motivation and emotions are experienced in learning situations, so let's study them: introduction to the special issue. *Learn. Instr.* **81**, 101623 (2022).
45. Robins, J. M., Scheines, R., Spirtes, P. & Wasserman, L. Uniform consistency in causal inference. *Biometrika* **90**, 491–515 (2003).
46. VanderWeele, T. J. & Hernán, M. A. Causal inference under multiple versions of treatment. *J. Causal Inference* **1**, 1–20 (2013).
47. Pearl, J. Does obesity shorten life? Or is it the soda? On non-manipulable causes. *J. Causal Inference* **6**, 20182001 (2018).
48. Angrist, J. D. & Pischke, J.-S. *Mastering 'Metrics: The Path from Cause to Effect* (Princeton Univ. Press, 2014).
49. Eronen, M. I. Causal discovery and the problem of psychological interventions. *N. Ideas Psychol.* **59**, 100785 (2020).
50. Scheines, R. The similarity of causal inference in experimental and non-experimental studies. *Phil. Sci.* **72**, 927–940 (2005).
51. Bringmann, L. F., Elmer, T. & Eronen, M. I. Back to basics: the importance of conceptual clarification in psychological science. *Curr. Dir. Psychol. Sci.* **31**, 340–346 (2022).
52. Spirtes, P. & Scheines, R. Causal inference of ambiguous manipulations. *Phil. Sci.* **71**, 833–845 (2004).
53. Bollen, K. A. & Brand, J. E. A general panel model with random and fixed effects: a structural equations approach. *Soc. Forces* **89**, 1–34 (2010).
54. Hamaker, E. L., Kuiper, R. M. & Grasman, R. P. P. P. A critique of the cross-lagged panel model. *Psychol. Methods* **20**, 102–116 (2015).
55. Zyphur, M. J. et al. From data to causes I: building a general cross-lagged panel model (GCLM). *Organ. Res. Methods* **23**, 651–687 (2020).
56. Voelkle, M. C., Oud, J. H. L., Davidov, E. & Schmidt, P. An SEM approach to continuous time modeling of panel data: relating authoritarianism and anomia. *Psychol. Methods* **17**, 176–192 (2012).
57. Frangakis, C. E. & Rubin, D. B. Principal stratification in causal inference. *Biometrics* **58**, 21–29 (2002).
58. Beltz, A. M. & Gates, K. M. Network mapping with GIMME. *Multivar. Behav. Res.* **52**, 789–804 (2017).
59. Montoya, L. M. et al. The optimal dynamic treatment rule superlearner: considerations, performance, and application to criminal justice interventions. *International J. Biostat.* **19**, 217–238 (2023).
60. Gische, C. & Voelkle, M. C. Beyond the mean: a flexible framework for studying causal effects using linear models. *Psychometrika* **87**, 868–901 (2022).
61. Imai, K. & Kim, I. S. When should we use unit fixed effects regression models for causal inference with longitudinal data? *Am. J. Polit. Sci.* **63**, 467–490 (2019).
62. Sobel, M. E. & Lindquist, M. A. Causal inference for fMRI time series data with systematic errors of measurement in a balanced on/off study of social evaluative threat. *J. Am. Stat. Assoc.* **109**, 967–976 (2014).
63. Usami, S. Within-person variability score-based causal inference: a two-step estimation for joint effects of time-varying treatments. *Psychometrika* **88**, 1466–1494 (2022).
64. Hamaker, E. L., Mulder, J. D. & van IJzendoorn, M. H. Description, prediction and causation: methodological challenges of studying child and adolescent development. *Dev. Cogn. Neurosci.* **46**, 100867 (2020).
65. Lundberg, I., Johnson, R. & Stewart, B. M. What is your estimand? Defining the target quantity connects statistical evidence to theory. *Am. Sociol. Rev.* **86**, 532–565 (2021).
66. Rohrer, J. M. & Murayama, K. These are not the effects you are looking for: causality and the within-/between-persons distinction in longitudinal data analysis. *Adv. Methods Pract. Psychol. Sci.* **6**, 251524592211408 (2023).
67. Silberzahn, R. et al. Many analysts, one data set: making transparent how variations in analytic choices affect results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).
68. Auspurg, K. & Brüderl, J. Has the credibility of the social sciences been credibly destroyed? Reanalyzing the 'many analysts, one data set' project. *Socius* **7**, 237802312110244 (2021).
69. Shadish, W. R, Cook, T. D & Campbell, D. T. *Experimental and Quasi-Experimental Designs for Generalized Causal Inference* (Houghton, Mifflin, 2002).
70. Rhemtulla, M., van Bork, R. & Borsboom, D. Worse than measurement error: consequences of inappropriate latent variable measurement models. *Psychol. Methods* **25**, 30–45 (2020).
71. Westfall, J. & Yarkoni, T. Statistically controlling for confounding constructs is harder than you think. *PLoS ONE* **11**, e0152719 (2016).
72. Grosz, M. P., Rohrer, J. M. & Thoemmes, F. The taboo against explicit causal inference in nonexperimental psychology. *Perspect. Psychol. Sci.* **15**, 1243–1255 (2020).
73. Deming, D. Early childhood intervention and life-cycle skill development: evidence from Head Start. *Am. Econ. J. Appl. Econ.* **1**, 111–134 (2009).
74. Pion, G. M. & Lipsey, M. W. Impact of the Tennessee Voluntary Prekindergarten Program on children's literacy, language, and mathematics skills: results from a regression-discontinuity design. *AERA Open* **7**, 233285842110413 (2021).
75. Ritchie, S. J. & Tucker-Drob, E. M. How much does education improve intelligence? A meta-analysis. *Psychol. Sci.* **29**, 1358–1369 (2018).
76. Steiner, P. M., Wong, V. C. & Anglin, K. A causal replication framework for designing and assessing replication efforts. *Z. Psychol.* **227**, 280–292 (2019).
77. Munafò, M. R. & Davey Smith, G. Robust research needs many lines of evidence. *Nature* **553**, 399–401 (2018).
78. Colnet, B. et al. Causal inference methods for combining randomized trials and observational studies: a review. *Stat. Sci.* **39**, 165–191 (2024).
79. Wan, S., Brick, T. R., Alvarez-Vargas, D. & Bailey, D. H. Triangulating on developmental models with a combination of experimental and nonexperimental estimates. *Dev. Psychol.* **59**, 216–228 (2022).
80. Gische, C., West, S. G. & Voelkle, M. C. Forecasting causal effects of interventions versus predicting future outcomes. *Struct. Equ. Modeling* **28**, 475–492 (2021).
81. Imai, K., Kim, I. S. & Wang, E. H. Matching methods for causal inference with time-series cross-sectional data. *Am. J. Polit. Sci.* **67**, 587–605 (2021).

82. Zyphur, M. J. et al. From data to causes II: comparing approaches to panel data analysis. *Organ. Res. Methods* **23**, 688–716 (2020).
83. Lüdtke, O. & Robitzsch, A. A comparison of different approaches for estimating cross-lagged effects from a causal inference perspective. *Struct. Equ. Modeling* **29**, 888–907 (2022).
84. Usami, S., Murayama, K. & Hamaker, E. L. A unified framework of longitudinal models to examine reciprocal relations. *Psychol. Methods* **24**, 637–657 (2019).
85. Bond, T. N. & Lang, K. The evolution of the black–white test score gap in grades K–3: the fragility of results. *Rev. Econ. Stat.* **95**, 1468–1479 (2013).
86. Larzelere, R. E., Cox, R. B. & Smith, G. L. Do nonphysical punishments reduce antisocial behavior more than spanking? A comparison using the strongest previous causal evidence against spanking. *BMC Pediatr.* **10**, 10 (2010).
87. Oster, E. Unobservable selection and coefficient stability: theory and evidence. *J. Bus. Econ. Stat.* **37**, 187–204 (2019).
88. Athey, S., Chetty, R., Imbens, G. W. & Kang, H. *The Surrogate Index: Combining Short-Term Proxies to Estimate Long-Term Treatment Effects More Rapidly and Precisely* Working Paper No. 26463 (NBER, 2019).
89. Weidmann, B. & Miratrix, L. Lurking inferential monsters? Quantifying selection bias in evaluations of school programs. *J. Policy Anal. Manage.* **40**, 964–986 (2021).
90. Dehejia, R. H. & Wahba, S. Causal effects in nonexperimental studies: reevaluating the evaluation of training programs. *J. Am. Stat. Assoc.* **94**, 1053–1062 (1999).
91. LaLonde, R. J. Evaluating the econometric evaluations of training programs with experimental data. *Am. Econ. Rev.* **76**, 604–620 (1986).
92. Protzko, J. Effects of cognitive training on the structure of intelligence. *Psychon. Bull. Rev.* **24**, 1022–1031 (2017).
93. Schmidt, F. L. Beyond questionable research methods: the role of omitted relevant research in the credibility of research. *Arch. Sci. Psychol.* **5**, 32–41 (2017).
94. Meehl, P. E. Why summaries of research on psychological theories are often uninterpretable. *Psychol. Rep.* **66**, 195–244 (1990).
95. Chaku, N., Kelly, D. P. & Beltz, A. M. Individualized learning potential in stressful times: how to leverage intensive longitudinal data to inform online learning. *Comput. Hum. Behav.* **121**, 106772 (2021).
96. Moeller, J. et al. Generalizability crisis meets heterogeneity revolution: determining under which boundary conditions findings replicate and generalize. Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/5wsna (2022).
97. Dunning, T. et al. (eds). *Information, Accountability, And Cumulative Learning: Lessons From Metaketa I* (Cambridge Univ. Press, 2019).
98. Low, H. & Meghir, C. The use of structural models in econometrics. *J. Econ. Perspect.* **31**, 33–58 (2017).
99. Todd, P. E. & Wolpin, K. I. Assessing the impact of a school subsidy program in Mexico: using a social experiment to validate a dynamic behavioral model of child schooling and fertility. *Am. Econ. Rev.* **96**, 1384–1417 (2006).
100. Pearl, J., Glymour, M. & Jewell, N. P. *Causal Inference in Statistics: A Primer* (John Wiley & Sons, 2016).
101. Achen, C. H. Let's put garbage-can regressions and garbage-can probits where they belong. *Confl. Manage. Peace Sci.* **22**, 327–339 (2005).
102. Athey, S. & Imbens, G. Recursive partitioning for heterogeneous causal effects. *Proc. Natl Acad. Sci. USA* **113**, 7353–7360 (2016).
103. Geng, E. H., Holmes, C. B., Moshabela, M., Sikazwe, I. & Petersen, M. L. Personalized public health: an implementation research agenda for the HIV response and beyond. *PLoS Med.* **16**, e1003020 (2019).
104. Moeller, J. Averting the next credibility crisis in psychological science: within-person methods for personalized diagnostics and intervention. *J. Pers. Oriented Res.* **7**, 53–77 (2021).
105. Pearl, J. & Bareinboim, E. Transportability of causal and statistical relations: a formal approach. *Proc. AAAI Conf. Artif. Intell.* **25**, 247–254 (2011).
106. Wager, S. & Athey, S. Estimation and inference of heterogeneous treatment effects using random forests. *J. Am. Stat. Assoc.* **113**, 1228–1242 (2018).
107. Benjamin-Chung, J. et al. Spillover effects in epidemiology: parameters, study designs and methodological considerations. *Int. J. Epidemiol.* **47**, 332–347 (2018).
108. Hudgens, M. G. & Halloran, M. E. Toward causal inference with interference. *J. Am. Stat. Assoc.* **103**, 832–842 (2008).
109. Imai, K., Jiang, Z. & Malani, A. Causal inference with interference and noncompliance in two-stage randomized experiments. *J. Am. Stat. Assoc.* **116**, 632–644 (2021).
110. Tchetgen, E. J. T. & VanderWeele, T. J. On causal inference in the presence of interference. *Stat. Methods Med. Res.* **21**, 55–75 (2012).
111. Zhang, C., Mohan, K. & Pearl, J. Causal inference with non-IID data using linear graphical models. *Adv. Neural Inf. Process. Syst.* **35**, 13214–13225 (2022).
112. Eberhardt, F. & Scheines, R. Interventions and causal inference. *Phil. Sci.* **74**, 981–995 (2007).
113. Mooij, J. M., Magliacane, S. & Claassen, T. Joint causal inference from multiple contexts. *J. Mach. Learn. Res.* **21**, 3919–4026 (2020).
114. Peters, J., Bühlmann, P. & Meinshausen, N. Causal inference by using invariant prediction: identification and confidence intervals. *J. R. Stat. Soc. B* **78**, 947–1012 (2016).
115. Aalen, O., Røysland, K., Gran, J., Kouyos, R. & Lange, T. Can we believe the DAGs? A comment on the relationship between causal DAGs and mechanisms. *Stat. Methods Med. Res.* **25**, 2294–2314 (2016).
116. Driver, C. C. & Voelkle, M. C. in *Continuous Time Modeling in the Behavioral and Related Sciences* (eds Van Montfort, K. et al.) 79–109 (Springer International, 2018).
117. Røysland, K. A martingale approach to continuous-time marginal structural models. *Bernoulli* **17**, 895–915 (2011).
118. Ryan, O. & Hamaker, E. L. Time to intervene: a continuous-time approach to network analysis and centrality. *Psychometrika* **87**, 214–252 (2022).

Acknowledgements

This Review resulted from a cross-disciplinary workshop discussing such approaches (https://www.longitudinaldataanalysis.com/). The workshop and collaboration were funded by the Jacobs Foundation and CIFAR. The funders had no role in the decision to publish or in the preparation of this manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Correspondence and requests for materials should be addressed to Drew H. Bailey.

Peer review information *Nature Human Behaviour* thanks Jörn-Steffen Pischke and Rebecca Johnson for their contribution to the peer review of this work.