Received: 4 October 2020 | Revised: 6 March 2023 | Accepted: 21 March 2023

DOI: 10.1002/wcs.1650

OVERVIEW

WIREs COGNITIVE SCIENCE WILEY

Causal inference in cognitive neuroscience

David Danks[1] | **Isaac Davis**[2]

[1]Halicioglu Data Science Institute, Department of Philosophy, University of California San Diego, La Jolla, California, USA

[2]Department of Psychology, Yale University, New Haven, Connecticut, USA

Correspondence
David Danks, Halicioglu Data Science Institute, Department of Philosophy, University of California San Diego, 9500 Gilman Drive, MC 0555, La Jolla, CA 92093-0021, USA.
Email: ddanks@ucsd.edu

Edited by: Wayne Wu, Editor

Abstract

Causal inference is a key step in many research endeavors in cognitive science and neuroscience, and particularly cognitive neuroscience. Statistical knowledge is sufficient for prediction and diagnosis, but causal knowledge is required for action and intervention. Most statistics courses and textbooks emphasize the difficulty of causal inference, focusing on the maxim that "correlation does not mean causation": there can be multiple causal possibilities, often many of them, consistent with given observed statistics. This paper focuses instead on the conceptual issues and assumptions that confront causal and other kinds of inference, primarily focusing on cognitive neuroscience. We connect inference methods with goals and challenges, and provide concrete guidance about how to select appropriate tools for the scientific task.

This article is categorized under:

Psychology > Theory and Methods

Philosophy > Foundations of Cognitive Science

KEYWORDS

causal inference, cognitive neuroscience, methodology

1 | INTRODUCTION

One key goal of the cognitive and neural sciences is causal knowledge, whether for explanation, prediction, or control. While mere correlation can be sufficient for goals such as simple observational prediction, we often aim to understand the underlying causal structures that generate some behavior or phenomenon. As cognitive scientists and neuroscientists, we want to understand, for instance, cognitive learning and the representations that result, or the ways in which parts of the brain are causally connected to generate behavior, or the fine-grained details that lead to reliable functioning of a neural circuit, or one of many other causal structures at a variety of levels within the mind/brain. We do not simply seek to predict or capture the statistics of behavior, but to further understand the mechanisms—cognitive and neural—that generated the behavior.

Causation can be a slippery notion; researchers are often unclear about exactly what they have in mind. In this article, we adopt a broadly interventionist notion (Woodward, 2005): *C* causes *E* just when a possible (not necessarily feasible) intervention on *C* would probabilistically alter *E* when all else is held fixed. This notion of causation is deliberately agnostic about mechanistic details, underlying constituents, and so forth. Nonetheless, it captures much of what matters about causation in the cognitive and neural sciences, including for (many) explanations, predictions, and both laboratory and clinical interventions.

All causal knowledge, including this relatively agnostic type, can be notoriously difficult to acquire. As every statistics course emphasizes, correlations are insufficient for causal knowledge. Given a correlation or association between two factors A and B, there will typically be multiple possible causal explanations of that informational connection. Evidence about A's value could be informative about B's value (and vice versa) for many different reasons: perhaps A causes B, or B causes A, or there is some unobserved third factor that causes them both (a so-called confounder), or some combination of these possibilities. Or perhaps there is a noncausal explanation for the correlation or association; for example, the association might be accidental (e.g., due to small sample size), or a selection bias might induce an association, or an inappropriate statistical test was used to infer the association, or some other possibility. In light of these challenges, we must use methods specifically for causal inference to acquire causal knowledge.

At the same time, there is often confusion about exactly how we might acquire causal knowledge, particularly if our data come from noisy measurements in not-fully-randomized experiments on poorly understood systems within highly variable individuals. One might justifiably wonder if causal inference is simply hopeless in these cases; perhaps there is nothing to be done beyond calculation of the relevant descriptive statistics. While this response is understandable, we suggest that it is too quick. We ultimately cannot sidestep the maxim that "even the most sophisticated statistical analysis [including causal inference methods] cannot salvage a badly designed experiment" (Box et al., 1978, p. vii), but there are methods that can provide significant insight, as long as they are appropriately selected.

In this article, we examine what kinds of knowledge are required for various scientific goals, and the corresponding methods that can provide that kind of insight. For reasons of space, we largely sidestep issues of experimental design. Every statistical and inferential method requires assumptions, and the assumptions that we are justified in making often depend on the experimental design. Despite this close connection, we primarily consider what should be done given a particular dataset, not how the data were collected.

We focus throughout on cognitive neuroscience, as it is one of (if not *the*) most data-intensive of the cognitive and neural sciences. The development and spread of high-resolution neuroimaging modalities have enabled cognitive neuroscientists to collect terabytes of data about each individual participant, which require appropriate statistical and data science methods for analysis. There are many complex challenges that arise for causal learning in cognitive neuroscience (e.g., Ramsey et al., 2010), and so we need to be particularly thoughtful about our choices of methods in this domain. Having said that, many of our discussions are not unique to cognitive neuroscience, but rather apply whenever we try to answer scientific questions with large amounts of data. We thus hope that much of what we say here will be useful in other parts of the cognitive and neural sciences as they increasingly use richer and larger data sources (e.g., Peterson et al., 2021).

Throughout this article, we will use two running examples of hypothetical studies to better illustrate the goals, challenges, and methods that commonly arise in cognitive neuroscience. For the first example, consider an experiment in which participants are shown a series of gambles—winning W with probability p, or losing L with probability $1 - p$—and choose whether to accept or reject each gamble (e.g., Botvinik-Nezer et al., 2019). Participants are scanned using one's preferred imaging modality prior to, during, and after each decision. Within this paradigm, there are several research questions we might pose. We might ask whether the neural data enable: classification of participants in terms of risk-aversion (Canessa et al., 2013), or inferring which type of gamble is being considered in a particular trial (Vilares et al., 2017), or prediction of whether a participant will accept or reject a particular gamble. Alternately, we might seek to act in the world by finding interventions that affect participants' judgments. Or we might ask questions about how risky choices are represented in the brain, and the mechanisms through which they give rise to decisions (Tom et al., 2007). Of course, these questions are neither exhaustive nor exclusive: for example, questions of implementation have obvious relevance to designing interventions, and classifying a participant's risk aversion could inform the choice of intervention.

As a second example, consider efforts to better understand the neural and cognitive mechanisms of individuals on the psychotic spectrum, including schizophrenia and bipolar diagnoses. There are significant patterns and heterogeneity in the symptoms of individuals on this spectrum (Weinberger & Harrison, 2011), perhaps suggesting that it is not unidimensional (Borsboom, 2017). Multiple studies have shown certain brain similarities for individuals on this spectrum (Argyelan et al., 2014; Hibar et al., 2018; Long et al., 2020; van Erp et al., 2018), though with some differences (Rashid et al., 2014; Rashid et al., 2016). Given these findings, a natural study would be to (i) identify tasks with potential performance variation across this spectrum; (ii) use fMRI (or other neuroimaging modality) while participants perform those tasks; and (iii) measure a number of other markers and correlates of psychosis. We could then ask whether we can determine people's diagnoses from the imaging data, or attempt to predict either present or future symptoms. We could even (potentially) use the neuroimaging data to try to understand the neural mechanisms that underlie

performance differences. As with the first running example, these questions are neither exclusive nor exhaustive, and many studies attempt to answer multiple of them. Nonetheless, we can already begin to see how these different questions will require different types of knowledge to answer.

We thus begin in Section 2 by outlining three different common research goals—classification, prediction, and intervention—as well as the types of background knowledge about the world, including our measurement methods, that we must have to make progress towards each goal. In particular, only the third goal requires *causal* knowledge. We then turn in Section 3 to selection of methods, and accompanying complexities, when addressing a particular goal. The division between causal and noncausal knowledge will again prove relevant in our choices of methods. We conclude that section with practical guidance about the steps and considerations that should factor into our methodological choices. We then turn in Section 4 to some broader theoretical and methodological challenges that arise for causal inference in cognitive neuroscience.

2 | GOALS AND KNOWLEDGE IN COGNITIVE NEUROSCIENCE

2.1 | Noncausal goal: Classification

At a high level, the aim of any classification study is to link some experimental or participant-specific condition to patterns of neural structure or activity, and thereby enable us to use those neural signatures to classify future states or conditions. This can include detecting persistent conditions such as neurodegenerative disorders (e.g., Alzheimer's disease) based on changes in the structure or functionality of a subject's brain (Rathore et al., 2017). This can also include so-called "mind-reading" studies (Norman et al., 2006) that classify transient cognitive states (e.g., accepting/rejecting a gamble) at a particular time (Vilares et al., 2017). As this last example suggests, classification challenges are sometimes described using the language of prediction, for example, "can we use neuroimaging data to predict whether someone will accept gamble *G*?" From a computational perspective, though, this question is fundamentally a classificatory one: we aim to use the neural data to classify participants as *G*-accepters versus *G*-rejectors.[1] While background knowledge about relevant brain features can be helpful in developing a classifier given neural data, many high-profile examples have been purely data-driven.

Regardless of the particular focus, training a classifier from neural data neither requires nor provides causal knowledge about pathways from experimental conditions to brain states. Instead, classification requires two pieces of noncausal knowledge. First, we require a set of "ground-truth" categories and data from which we can learn: we need to know, for at least some individuals, their state or condition so that we can find neural patterns or signatures that correlate with it. In diagnosis studies, the ground truth comes from patient diagnoses by a medical professional; in "mind-reading" studies, the experimental condition, supplemented with some cognitive psychology, typically provides ground truth (e.g., the gamble at that moment, coupled with the assumption that the participant read and understood the gamble).

Second, we require a set of brain features and other data to use as input for the classifier. These could comprise activation levels for a single neuron (Gold & Shadlen, 2001), average response levels across individual voxels in a specific brain region (Poline et al., 1997), patterns of responses across many voxels in multiple brain regions (O'Toole et al., 2007), or aggregate measures of different brain structure across multiple regions (Bassett & Sporns, 2017). Much of the challenge in developing a classifier is choosing the right set of brain features (see below), and so we typically use both statistical tools and prior knowledge about brain regions and cognitive functions. Regardless of the exact approach, the aim of classification is to learn a function that maps neural measurements to category labels (e.g., from fMRI data to predicted gamble acceptance) as accurately as possible. In general, this goal does not require causal knowledge, as any informative "signal" of the category label is useful, regardless of whether it is a cause.

2.2 | Noncausal goal: Prediction

Unlike classification studies, which aim to identify patterns or regions of brain activity that correlate with observed conditions, "functional connectivity" studies aim to predict the future activity of some brain feature Y, given the current activity of some brain feature X. The implicit (and sometimes, explicit) hope is that the predictors of a future brain state can provide a guide toward—perhaps even be equal to—the actual causes. These types of predictive tasks thus require

a richer understanding of the neural circuits underlying cognitive processes, though not a fully causal one. These circuits are frequently identified using the statistical technique of Granger causality (or G-causality): if the value of X at time t provides information about the value of Y at time $t + 1$ (with all other variables in the system held fixed), then X is a "G-cause" of Y (Friston et al., 2014; Granger, 1980). Despite the name, it is important to note that "X is a G-cause of Y" does *not* imply that "X is a cause of Y" in the usual sense: intervening on a G-cause need not lead to a change in the G-effect. Having said that, there are specific conditions (e.g., if all relevant variables are measured at an appropriate rate so there are no confounders) when G-causation implies causation, though those conditions rarely hold in cognitive neuroscience. At the same time, G-causality is typically a *necessary* condition for causality, and so, at the least, these analyses can help to identify candidates for genuine causal connections.

Granger-causality analyses require time series data over some set of neural variables at approximately the same sampling rate as the underlying connections. One complicating factor specifically for fMRI measurements is the potential for undersampling, or mismatch between the true rate at which neural activity spreads across these variables, and the rate at which measurements are taken from those variables. Undersampling can lead to erroneous conclusions about the existence or direction of potential causal relationships if we use Granger causality (Huettel et al., 2004; Seth et al., 2013). Like most analysis methods for temporal data, Granger causality analysis requires a stationarity assumption about the data; certain properties of the data must be stable over time (though there are variations that permit weaker stationarity assumptions, as well as methods for transforming the data to force a type of stationarity). These approaches require fewer assumptions than causal discovery methods, but only provide information about stable predictors of future states, not stable *causes* of those future states.

2.3 | Causal goal: Intervention

Classification and prediction are both fundamentally observational goals: we use information from the world to classify an individual or infer what is likely to happen next. As is well-known, though, observational goals and knowledge do not provide reliable paths to action. From the observation that people are wearing sweaters outside, we can infer that it is likely cold; however, donning a sweater in the summer will not lower the outdoor temperature. In cognitive neuroscience, novel treatments for individuals on the psychosis spectrum, or interventions that alter (in some way) people's propensities to accept various gambles, require causal, not merely predictive, knowledge, as we must ensure that we act on causes of whatever we want to influence, rather than its effects. In contrast, both causes and effects can be observationally useful when we want to infer a state or feature.

If we are trying to learn whether an intervention on C will make a difference in E, the gold-standard approach is to randomly manipulate C across a population of individuals, and then see whether there are systematic differences in the resulting E values. The primary benefit of manipulating (rather than observing) is that we can distinguish between causes and effects, as a manipulation of an effect will not make a difference in contrast with manipulation of a cause. The primary benefit of randomization is that we can have high confidence that any resulting changes in E are due to our manipulation, rather than unobserved factors that happened to covary with our actions. In cognitive neuroscience, however, we can rarely conduct this type of experiment, whether because we cannot directly manipulate the factors of interest (e.g., cortical networks) or because we cannot randomize our manipulations (e.g., if our intervention depends on the individual's symptoms). Moreover, we are often interested in multiple potential causes, and the strategy of random manipulation does not readily scale.

A variety of methods have been developed for causal inference outside of randomized manipulations (see Section 3.3). While details differ between these methods, they all require stronger background knowledge or assumptions than classification or prediction methods. For example, many methods require some form of a Markov assumption: a variable is (statistically) independent of its noneffects, conditional on its direct causes. This assumption does not require prior knowledge of any specific causal relations, but it does require general knowledge about what might happen if C did cause E. More generally, all causal inference methods require some knowledge or assumptions about "how causation manifests in data" in a particular context, though they do not require that we know which causal relations actually occur in that context.

In terms of the data requirements, causal inference methods are not significantly different from classification or prediction methods. We need to have measurements of the factors that (potentially) matter, such as brain states, neural dynamics, behavioral responses, or symptoms. If we have time series data, then we can use timing information to improve causal inference, though as with Granger causality, "time order plus correlation" is usually not sufficient to

infer a causal relationship. This overlap in terms of data requirements is one reason that it is easy to slide from classificatory or predictive goals to interventional ones, as the same data can be used to ask all three kinds of questions. Each requires distinct types of background knowledge and assumptions, though, and so we must ensure that we are justified in shifting between goals in these ways.

3 | METHODS AND COMPLEXITIES

We turn now to some methods that are relevant for these different scientific goals in cognitive neuroscience. We discuss classification and prediction methods since many of them will be familiar to readers, and we aim to emphasize that these methods are *not* specifically causal (at least, not without significant additional assumptions). Throughout, we draw a distinction between determining which factors are relevant at all, versus determining how (or with what strength) particular factors matter. We call these "discovery" and "estimation," respectively, as the former type of method aims to discover a qualitative model (i.e., what matters) while the latter aims to estimate relations within such a model (perhaps implicitly specified).[2] Although methods could potentially perform both discovery and estimation, most address only one or the other (or perform them sequentially as discovery followed by estimation).

3.1 | Classification

Contemporary classification studies comprise two different high-level approaches: univariate methods that test for statistical correlations between individual variables (e.g., single voxel or region) and experimental conditions (Poldrack et al., 2011), and multivariate methods that identify patterns across multiple variables that correlate with experimental conditions (Haynes & Rees, 2006; Norman et al., 2006). Univariate methods make the unrealistic simplifying assumption that variables are (statistically) independent of one another. They thereby provide a straightforward way to identify aggregate brain regions that "light up" in a particular state or condition, and so are potentially useful for classification (though with potential concerns; see Kriegeskorte et al., 2006). Thus, univariate methods are generally useful for "coarse" classification, but can have limited resolution for more fine-grained classification problems. For example, univariate methods have been useful for identifying regions of the brain that are more active during creative problem solving (Fink et al., 2009), but not for understanding finer-grained representations.

Multivariate methods like multi-voxel pattern analysis (MVPA) do not treat individual variables as independent of one another, but rather seek to identify patterns of activity that correlate with conditions. The multivariate space is significantly more computationally complex, so most methods first learn an "embedding" of the high-dimensional brain data into a low-dimensional space (e.g., using PCA or ICA). Given a low-dimensional embedding, these methods learn a "decision boundary" that maps onto the experimental categories (e.g., using linear methods like Support Vector Machines and Naive Bayes Classification, or nonlinear methods like deep neural networks). The low-dimensional embedding both makes classification more tractable, and also provides a heuristic for interpreting patterns of neural activity in the context of classification. For example, these methods can reveal whether different neural patterns track probabilities versus payoffs in a gamble.

The main challenge in MVPA (and other multivariate methods) is feature selection: fMRI can record roughly 100,000 voxels, and it is simply not feasible to learn all relevant subsets and patterns (Samuel Schwarzkopf & Rees, 2011), though methods such as recursive feature elimination and other regularization techniques can enable researchers to identify high-quality feature sets (Hanson et al., 2004; Hanson & Halchenko, 2008). Alternatively, one can use prior knowledge or an initial univariate analysis (perhaps generalized in various ways, such as including nearby voxels to the statistically correlated ones; Haxby et al., 2001; Mitchell et al., 2004), though with the risk of excluding voxels that are individually uncorrelated with the experimental condition, but are nonetheless part of a meaningful pattern of activity. Other methods for feature selection apply univariate analysis methods directly to whole patterns of variables, though heuristics must be used to address the inevitable combinatorial explosion of potential patterns to test (Norman et al., 2006). More generally, it is important to emphasize that classification does not, in general, provide a neural *explanation* of cognitive phenomena. That is, we can use these methods to identify brain regions or patterns of activity that reliably correlate with target conditions, but they do not (without further analysis) reveal the function that these regions or patterns serve in the cognitive process of interest, nor the causal role that they play in performing a

cognitive function. That being said, these features could provide the basis for robust, between-participant classification models (Poldrack et al., 2009).

In recent years, deep learning (DL) methods have become increasingly common in classification studies (Wen et al., 2018). DL methods can help automate the process of reducing the dimensionality of the input data, thereby selecting a small set of relevant features or dimensions on which to learn the decision boundary. However, while DL methods are useful for automating this dimensionality reduction, they also generally require a substantial quantity of data, which limits their application in studies with sparser data streams. Furthermore, DL methods have been criticized as being opaque: although DL methods can, given enough data, learn nearly any function or decision boundary, there is not always a straightforward way to interpret what the method has "learned" after the fact (Smucny et al., 2022).

3.2 | Prediction

Recall that the aim of functional connectivity studies is to identify functional circuits in the brain underpinning cognitive activity. These circuits are identified by finding the predictively relevant factors using time series data. The most common method for conducting this search is Granger causality analysis (GCA): if X both precedes Y temporally and provides information about Y (all else held fixed), then X is a "G-cause" of Y and can be used to predict Y. Thus, GCA involves autoregressing a set of time series variables to identify which variables most directly predict the values of which other variables (Friston et al., 2014; Granger, 1980). Importantly, GCA is generally used as a "model-free" method, in that it does not require strong assumptions about the structural connectivity underlying a small set of predetermined ROIs. Rather, GCA typically starts with a complete graph over a large set of ROIs, and gradually eliminates connections between variables that do not reliably predict each other. The final remaining connections will (ideally) correspond to a functional circuit of neural activation that co-occurs with certain cognitive processes.

GCA is a powerful method for time series analysis if we aim to predict future brain states given the current state. The method efficiently identifies informationally relevant predictors, and outputs a model that can be used for rapid inference. Its popularity is justified, but GCA faces many challenges when people try to use it for more than prediction. First, the informational value of a factor depends on what else was measured, so G-causality depends on the variable set. Second, GCA is very dependent on the temporal structure of the data, so easily fails if there is a mismatch between the timescale at which neural activity actually occurs, and the timescale at which measurements are taken (as is common with fMRI; Huettel et al., 2004). In such cases, the true connectivity structure may be quite different than what GCA reveals (Seth et al., 2013). Time series data are helpful for prediction, but not a panacea for the challenges of causal inference.

3.3 | Intervention

As noted above, methods that yield causal knowledge are distinguished partly by their assumptions. Many standard methods for observational goals can be used to guide interventions, if we have relevant background knowledge. For example, simple linear regression is a reliable method for causal estimation, if we know that we have measured all relevant factors, we have the correct temporal order, and so forth. Of course, we rarely have such strong background knowledge, but it is important to note that causal inference methods are not necessarily distinguished by their mathematics. In particular, Granger-causal relations actually *can* correspond to actual causal relations, but Granger-causality is only an effective, reliable causal learning method if we have substantial background knowledge, much of it causal in nature.[3] Standard applications of Granger-causality without those assumptions, though, do not yield causal knowledge.

Over the past 40 years, a range of methods have been developed for causal learning when we cannot perform randomized manipulations. Causal discovery methods—many of which learn causal graphical models—can learn multivariate causal structures from observational, experimental, and mixed data (see, e.g., Spirtes et al., 1993; Eberhardt, 2009; Pearl, 2009; Malinsky & Danks, 2018; or many references therein). At a high level, these methods determine the set of causal structures—perhaps one, perhaps many—that could have produced these data (without assuming any privileged predictive target). Moreover, different algorithms have been developed for many complex scientific settings (e.g., CD-NOD for heterogeneous, nonstationary time series data; Huang et al., 2020), including many developed specifically for cognitive neuroscience data, particularly fMRI data (e.g., IMaGES, Ramsey et al., 2010; FASK, Sanchez-Romero et al., 2019). These methods have been shown to significantly outperform other methods for causal

discovery on synthetic fMRI data (Sanchez-Romero et al., 2019). Typically, causal discovery methods are reliable for all cases in which traditional observational methods work, as well as others for which observational methods are not reliable for causal discovery. That is, causal learning without substantial prior knowledge should almost always be done with causal discovery methods, rather than observational methods such as regression or Granger-causality.

Causal estimation methods—many formulated in the potential outcomes framework (see, e.g., Rubin, 2005 for an overview)—require some specification of the likely causal structure, and then estimate the strengths or causal effect sizes within that structure. The initial specification need not be exactly right; at the very least, we might allow that *C* causes *E*, but then estimate its effect size as zero. Many causal estimation methods focus on identifying the causal impact of a few key factors, but they need not be restricted in this way. Most relevant for cognitive neuroscience, dynamic causal modeling (DCM; Friston, 2009) is a causal estimation method that aims to learn causal influences given a specification of the relations and couplings between different neural regions. Mediation methods (VanderWeele, 2015) are another common example of causal estimation, as they aim to determine whether the causal influence of *C* on *E* flows entirely through (i.e., is mediated by) some third factor *M* (or set of factors).

Most causal discovery and estimation methods are intended for either static or dynamic data, but not both. Causal inference from time series data can be both easier (e.g., time order can disambiguate causal direction) and harder (e.g., datapoints and "shocks" are typically correlated) than working with static or independent and identically distributed (i.i.d.) data. Almost all causal inference algorithms have been designed and improved by exploiting features of the specific types of data (including measurement methods) and contexts. As a result, we typically have distinct causal inference methods for static and dynamic data, regardless of whether we are engaged in discovery or estimation.

3.4 | Practical guidance

We conclude this section with concrete guidance for identifying the best analysis method(s) for a given study. To choose the right method, it is crucial to clearly understand both the requirements for achieving the aim of a study, as well as the constraints imposed by the level of background knowledge and structure of the data. To this end, we structure our advice around three key questions (Table 1):

1. *What kind of knowledge do we need to achieve the goal of our study*? (rows in Table 1): If we are using brain data to diagnose a disorder, "read" a subject's mental states, or predict a subject's behavior, then we can usually achieve these goals with classification. If our goal is to understand the structure and temporal dynamics of neural circuits underlying cognitive processes, then we need predictive knowledge. If our goal is to identify interventions that will yield some target effect (e.g., a treatment to reduce symptom severity, or a stimulus change to shape behavior in novel ways), then our analysis must go beyond prediction to yield causal knowledge.
2. *What is the scope of our background knowledge*? (columns in Table 1): If our background knowledge is minimal, then our data might include many variables, only a few of which are actually relevant to the cognitive process or

TABLE 1 Examples of methods for different goals (rows), level of background knowledge (columns), and data types (within-cell).

	Level of background knowledge	
Goal of analysis	**Discovery**	**Estimation**
Classification	*Static*: MVPA, Deep Learning	*Static*: Decision trees
	Temporal: Feature extraction	*Temporal*: ARIMA
Prediction	*Static*: Lasso & other variable selection methods	*Static*: Regression
	Temporal: GCM	*Temporal*: Structural vector autoregression (SVAR)
Intervention	*Static*: FASK, FCI	*Static*: Potential outcomes
	Temporal: IMaGES, CD-NOD	*Temporal*: DCM

Abbreviations: ARIMA, autoregressive integrated moving average; CD-NOD, causal discovery-nonstationary/heterogeneous data; DCM, dynamic causal model; FASK, fast adjacency skewness; FCI, fast causal inference; GCM, Granger causal model; IMaGES, independent multiple sample greedy equivalence search; MVPA, multi-voxel pattern analysis.

condition under study. In such cases, our analysis will typically require some measure of discovery or feature selection. This process may also involve detecting causal or temporal relations between relevant features in a nonparametric fashion (i.e., detecting the presence of a relation without quantifying the "strength" of that relation). In other cases, our background knowledge may be sufficient to identify a small set of relevant variables or brain features. For example, prior classification studies may reveal which brain regions are consistently active during a particular cognitive task, and temporal analysis like GCA might further narrow down the set of potential causal factors. If we have a suitable "model," then our analysis can focus on fine-tuned estimations of the parameters in that model (e.g., the strength of causal relations between brain features).

3. *What is the structure of the data*? (within-cell in Table 1): While there are many potential factors to consider, a particularly important one is the temporal structure of the data. In many classification studies, for example, the brain data are generated by averaging activation levels of a particular region across a fixed interval of time, and then used to classify a participant. In a case like this, averaging across time will eliminate the temporal structure of the data. On the other hand, many studies involve highly dynamic data in which the temporal order of activations is critical, such as in functional connectivity studies. In most cases, time-series data will require very different analysis methods than static data, and it is therefore important to understand the temporal structure of the data when choosing an analysis method.

Of course, many of the distinctions we raise here are not all-or-nothing. For example, "discovery" and "estimation" should be interpreted as two endpoints of a spectrum, rather than two mutually exclusive and exhaustive possibilities. Moreover, as we have repeatedly emphasized, we also need to think about legitimate assumptions and other factors when choosing an algorithm within a cell. For reasons of space, we do not delve into those further steps in the decision process, but we note that other papers do provide such overviews for particular cells in the table (e.g., Nogueira et al., 2022 for causal discovery methods).

4 | BROADER CHALLENGES

The use of any particular method invariably raises a number of challenges, particularly for some of the more computationally complex methods that have been developed for use in cognitive neuroscience. For reasons of space, we cannot attempt a survey of all such issues for all such methods. We instead focus here on two complexities that are particularly challenging for causal inference in cognitive neuroscience; both problems have been previously described in much more detail by other authors (most notably, Ramsey et al., 2010 and Ramsey et al., 2011), but the importance of these challenges has not necessarily been recognized by the broader cognitive neuroscience community.

4.1 | Mixture problems

One challenge that often arises across many of the domains we have discussed is the problem of mixtures (Ramsey et al., 2011). In statistical inference, a *mixture distribution* is a probability distribution $P_M(X)$ that is the combination of different component distributions, such as might result from combining data from individual participants who differ in important ways.[4] In many experiments, we compute means or medians, or otherwise focus on the group-level performance instead of examining each individual participant. This approach enables us to average out some of the participant-specific "noise," and the resulting averages can be tested against our theories about "normal" cognitive or neural performance. If our experimental participants truly are the same except for some independent noise or error, then this strategy will work well. However, if there are relevant cognitive or neural differences between individual participants, then this aggregation and averaging can qualitatively change the structure of the data.

To put this more concretely, the best model of a group might not be the best model of any individual (Ashby et al., 1994; Maddox, 1999), and so we must be careful not to infer that any particular individual has the same predictive or causal structures as the best-fitting model for the group. We must take care when generalizing results from individual to population, or from population to individual: a causal model that fits the aggregate data may not tell us anything useful about individual cognition, and vice versa. The exception is if our participants actually are all relevantly similar, but if we knew the underlying causal structures (cognitive or neural) to test this assumption, then causal inference would no longer be needed.

Moreover, mixture distributions can result even within a single individual's data, and so a shift to individual difference analyses will not necessarily help matters. In particular, we may be studying a cognitive or neural process that is itself composed of other processes, and where we do not know the nature and number of these other processes. To draw on one of our running examples, decision-making under risk can involve several different cognitive processes, including quantitative reasoning (e.g., computing the expected reward of several different gambles), emotion ("winning" and "losing" often incur strong emotional reactions, which can influence decision making), working memory (especially for decisions that involve many different options), and so forth. If we know the exact component processes, then we can adjust our statistical analyses so that they can support reliable causal inference. But again, that knowledge of component processes is often precisely what we are trying to learn through our research.

Mixture concerns for individual participants are even more prevalent in brain-mapping studies, as the proper level of analysis for identifying functional brain regions or activity patterns is itself a subject of ongoing study and debate in neuroscience (Bassett & Sporns, 2017). Since many brain structures may be recruited for many different cognitive functions, the activity we record in a particular brain structure or region may reflect multiple ongoing cognitive or neurological processes, either simultaneously, or sequentially throughout a trial. Analyzing the resulting measurements as if they reflect the activity of a single, coherent brain structure performing a single function may result in sampling artifacts: spurious connections or correlations which do not necessarily reflect any meaningful connection in the brain, and which would not be inferred under a different measurement process (McCaffrey & Danks, 2018). Furthermore, depending on our knowledge of the underlying mixture components, it may be difficult or impossible to determine which connections are spurious, which poses a serious challenge for genuine causal inference from neuroimaging data.

This last concern points to a more general kind of "mixture" problem: the phenomena we wish to study (and the data or outputs relevant to the phenomena) often occur at multiple, mutually interacting levels of analysis simultaneously. For example, the relevant systems may correspond to localized activity within particular brain regions, to connections between individual brain regions, to networks of brain regions and hierarchies of component networks, or aggregate measures of network structure applied across these network hierarchies (Bassett & Bullmore, 2006; Bassett & Sporns, 2017). In such cases, our ability to infer meaningful causal relations between measurements depends critically on our assumptions about which parts of the system we ought to model, how those parts interact, and how reliably we can isolate those parts from the rest of the system. That is, we often face a deep-seated "chicken-and-egg" problem for reliable causal inference in cognitive neuroscience.

4.2 | "Chicken-and-egg" problems in cognitive neuroscience

We understand a chicken-and-egg problem to be one in which we have two different challenges, and where solving one requires solving (or making assumptions about) the other. For causal inference, the problem is that the "right" causal factors are those that are part of the actual causal relations, but we can only determine the actual causal relations if we have the right causal factors. When we face a chicken-and-egg problem, we must find a way to break the justificatory feedback loop. One route is to try to infer both components simultaneously (e.g., Chalupka et al., 2016), but such methods are computationally complex and require significant assumptions. Alternatively, we can iteratively and sequentially address each challenge with the aim (or hope) of reaching equilibrium between the two components. This strategy is typically tractable—computationally, theoretically, and experimentally—but also typically not guaranteed to reach the best pair of answers.

In fact, cognitive neuroscience faces *two* distinct chicken-and-egg problems. One such problem was just noted: we are often interested in causal relations between phenomena that occur on multiple, mutually interacting levels of analysis simultaneously. In these cases, the proper system or level of analysis at which to define variables of interest is rarely obvious, and may be highly context-specific. For example, if we are interested in determining the neural basis of risky decision-making, then we need to know whether to use localized activity within a particular region, or connections between active regions, or networks of regions, or network types distributed throughout the brain, or some other variables (i.e., we face the problem of identifying the correct ontology). In response, we might first conjecture some plausible causal factors involved in risky decision-making (based on our prior knowledge), and attempt to infer a causal structure using those factors. On the basis of various "failures of fit" for the causal structure, we can then examine specific causal factors to try to improve our measure or understanding of them. This step may require us to change our variables, perhaps by developing novel measurement methods. These new variables can be used for further causal inference. And so on until we reach a mutual equilibrium between causal factors and causal structure. This iterative

process rarely occurs in any single paper, but arguably plays out over multiple papers (e.g., in the gradual identification and clarification of the resting state network).

The second chicken-and-egg problem is more general and applies across multiple sciences, though it is particularly problematic for cognitive neuroscience. On the one hand, our causal inference methods can only be shown to be reliable if we know something about the actual "ground truth" so that we can check the outputs of our methods. On the other hand, successfully learning those true causal structures in the world requires reliable methods. In other words, demonstrating reliability of causal inference methods requires knowledge of the truth, but reliable methods are required to obtain the truth. Since the "ground truth" is rarely directly observable in cognitive neuroscience, this chicken-and-egg problem is even more challenging. For this reason, reliable inference methods in cognitive neuroscience may be context-specific: a method that works well for one particular problem, with one particular set of background assumptions, may not be reliable at all for other problems under other assumptions. The challenge is not that we have no reliable methods at all, but rather that our methods will typically be reliable only for particular contexts or purposes, and so we must intentionally and intelligently select an appropriate method.

This issue is particularly salient in neuroimaging, as the causal inference methods are usually quite complex but we have limited ground truths about causal relations in the brain. For example, both GCA and DCM are used to infer directed connections between regions of interest in the brain, but differ in their assumptions and applications. DCM is a model-based method (Friston, 2009) that often requires a simple model with only a handful of regions of interest, and so the problem of unobserved common causes looms large. GCA is a model-free method that uses Granger-causality over all regions of interest to identify potential G-causal connections (Roebroeck et al., 2005; Roebroeck et al., 2011), and so is not reliable in certain conditions, including measurement noise (Friston et al., 2014) and undersampling (Seth et al., 2013). However, we cannot necessarily combine the two methods into one that is reliable in a wider set of cases, precisely because we do not necessarily know what constitutes improvement for these methods, as we do not know what is really happening (causally) in the brain. People frequently turn to synthetic data in such situations (e.g., Sanchez-Romero et al., 2019), but that approach requires substantial background knowledge.

If we are careless in examining our assumptions, or overgeneralize a method from one goal to another, we may end up with inaccurate models of the world, improperly verified by unreliable methods. We ought not use a method just because "everyone else in my field uses it" or "it worked for the last paper." Rather, we should consider the assumptions or conditions for reliability of some method, and then endeavor to use it only when it is applicable.

5 | CONCLUSIONS

Causal knowledge is critical for many, but certainly not all, scientific investigations. We must think carefully about whether we aim for classification and prediction, or instead intervention (and some kinds of explanation). For the former goals, causal knowledge is typically not required, though it can be useful in various ways. The latter goals always require causal knowledge, and so we must ensure that we use appropriate methods to acquire it. This choice of appropriate method depends more on the assumptions that we are justified in making, rather than the exact mathematics; many computational algorithms can be used for either prediction or causal estimation, for instance, depending on our background knowledge. Most scientists understand the importance of experimental design and measurement methods when choosing how to analyze their data. The importance of our scientific goals and other assumptions is less well understood, but no less relevant. In particular, we must be careful about drawing causal conclusions from methods (and assumptions) that cannot support them.

We are obviously unable to consider all of the challenges and complexities underlying the different methods used in cognitive neuroscience. Most notably, we have not carefully articulated the constraints imposed by different measurement methods and data collection frameworks. For example, is the temporal resolution of our imaging modality appropriate for the phenomenon of interest? Or are there intermediate data processing steps that can introduce problematic artifacts into our data (e.g., Power et al., 2012)? These types of issues can significantly influence the suitability of a particular method. More generally, we close by emphasizing that widely used methods are not necessarily those that *should* be used in a particular analysis. A common method may be inappropriate for a particular study because many others are interested in a different question or goal, or because particular assumptions are violated in this case, or a host of other reasons. By carefully considering our scientific goals, background knowledge, and assumptions, we can select appropriate methods (and avoid inappropriate ones) that provide causal information even for noisy observational data.

AUTHOR CONTRIBUTIONS

David Danks: Conceptualization (lead); supervision (lead); writing – original draft (equal); writing – review and editing (lead). **Isaac Davis:** Conceptualization (supporting); writing – original draft (equal); writing – review and editing (supporting).

CONFLICT OF INTEREST STATEMENT

The authors have declared no conflicts of interest for this article.

DATA AVAILABILITY STATEMENT

Data sharing is not applicable to this article as no new data were created or analyzed in this study.

ORCID

David Danks https://orcid.org/0000-0003-4541-5966

RELATED WIREs ARTICLES

Bayesian data analysis

ENDNOTES

[1] More generally, there is rarely a "bright line" between classification and prediction. In practice, people often seem to use different terms based solely on whether we are trying to infer one value from a small set of possibilities (classification) versus many possibilities (prediction).

[2] There is not necessarily a clean division between discovery and estimation, as "factor F does not matter" (discovery) is arguably the same as "zero relevance for F" (estimation). Nonetheless, estimation methods typically presuppose substantially more background knowledge than discovery methods.

[3] And if we do have such background knowledge, then there are typically more effective causal inference methods that we should use instead.

[4] For example, a bimodal distribution resulting from the "mixing" of two Gaussians; $P_M(X) = \alpha\, P_1(X) + (1 - \alpha)\, P_2(X)$.

FURTHER READING

Calafato, M. S., Thygesen, J. H., Ranlund, S., Zartaloudi, E., Cahn, W., Crespo-Facorro, B., Diez-Revuelta, A., Di Forti, M., Risk, G., Hall, M. H., Iyegbe, C., Jablensky, A., Kahn, R., Kalaydjieva, L., Kravariti, E., Lin, K., McDonald, C., McIntosh, A. M., McQuillin, A., ... Bramon, E. (2018). Use of schizophrenia and bipolar disorder polygenic risk scores to identify psychotic disorders. *The British Journal of Psychiatry*, *213*(3), 535–541.

Plis, S., Danks, D., Freeman, C., & Calhoun, V. (2015). Rate-agnostic (causal) structure learning. In C. Cortes, N. D. Lawrence, D. D. Lee, M. Sugiyama, & R. Garnett (Eds.), *Advances in neural information processing systems 28* (pp. 3303–3311). The NIPS Foundation.

Ruderfer, D. M., Ripke, S., McQuillin, A., Boocock, J., Stahl, E. A., Pavlides, J. M. W., Mullins, N., Charney, A. W., Ori, A. P., Loohuis, L. M. O., Domenici, E., Di Florio, A., Papiol, S., Kalman, J. L., Trubetskoy, V., Adolfsson, R., Agartz, I., Agerbo, E., Akil, H., ... Psychosis Endophenotypes International Consortium, Wellcome Trust Case-Control Consortium. (2018). Genomic dissection of bipolar disorder and schizophrenia, including 28 subphenotypes. *Cell*, *173*(7), 1705–1715.

REFERENCES

Argyelan, M., Ikuta, T., DeRosse, P., Braga, R. J., Burdick, K. E., John, M., Kingsley, P. B., Malhotra, A. K., & Szeszko, P. R. (2014). Resting-state fMRI connectivity impairment in schizophrenia and bipolar disorder. *Schizophrenia Bulletin*, *40*(1), 100–110.

Ashby, F. G., Maddox, W. T., & Lee, W. W. (1994). On the dangers of averaging across subjects when using multidimensional scaling or the similarity-choice model. *Psychological Science*, *5*, 144–151.

Bassett, D. S., & Bullmore, E. D. (2006). Small-world brain networks. *The Neuroscientist*, *12*, 512–523.

Bassett, D. S., & Sporns, O. (2017). Network neuroscience. *Nature Neuroscience*, *20*, 353–364.

Borsboom, D. (2017). A network theory of mental disorders. *World Psychiatry*, *16*(1), 5–13.

Botvinik-Nezer, R., Iwanir, R., Holzmeister, F., Huber, J., Johannesson, M., Kirchler, M., Dreber, A., Camerer, C. F., Poldrack, R. A., & Schonberg, T. (2019). fMRI data of mixed gambles from the neuroimaging analysis replication and prediction study. *Scientific Data*, *6*(1), 1–9.

Box, G. E. P., Hunter, J. S., & Hunter, W. G. (1978). *Statistics for experimenters* (1st ed.). Wiley.

Canessa, N., Crespi, C., Motterlini, M., Baud-Bovy, G., Chierchia, G., Pantaleo, G., Tettamanti, M., & Cappa, S. F. (2013). The functional and structural neural basis of individual differences in loss aversion. *Journal of Neuroscience*, *33*(36), 14307–14317.

Chalupka, K., Eberhardt, F., & Perona, P. (2016). Multi-level cause-effect systems. In *Proceedings of the 19th International Conference on Artificial Intelligence and Statistics*, PMLR 51:361–369.

Eberhardt, F. (2009). Introduction to the epistemology of causation. *Philosophy Compass*, *4*, 913–925.

Fink, A., Grabner, R. H., Benedek, M., Reishofer, G., Hauswirth, V., Fally, M., Neuper, C., Ebner, F., & Neubauer, A. C. (2009). The creative brain: Investigation of brain activity during creative problem solving by means of EEG and fMRI. *Human Brain Mapping*, *30*(3), 734–748.

Friston, K. (2009). Causal modelling and brain connectivity in functional magnetic resonance imaging. *PLoS Biology*, *7*, e1000033.

Friston, K. J., Bastos, A. M., Oswal, A., van Wijk, B., Richter, C., & Litvak, V. (2014). Granger causality revisited. *NeuroImage*, *101*, 796–808.

Gold, J. I., & Shadlen, M. N. (2001). Neural computations that underlie decisions about sensory stimuli. *Trends in Cognitive Sciences*, *5*(1), 10–16.

Granger, C. W. J. (1980). Testing for causality: A personal viewpoint. *Journal of Economic Dynamics and Control*, *2*, 329–352.

Hanson, S. J., & Halchenko, Y. O. (2008). Brain reading using full brain support vector machines for object recognition: There is no "face" identification area. *Neural Computation*, *20*(2), 486–503.

Hanson, S. J., Matsuka, T., & Haxby, J. V. (2004). Combinatorial codes in ventral temporal lobe for object recognition: Haxby (2001) revisited: Is there a "face" area? *NeuroImage*, *23*(1), 156–166.

Haxby, J. V., Ida Gobbini, M., Furey, M. L., Ishai, A., Schouten, J. L., & Pietrini, P. (2001). Distributed and overlapping representations of faces and objects in ventral temporal cortex. *Science*, *293*, 2425–2430.

Haynes, J. D., & Rees, G. (2006). Decoding mental states from brain activity in humans. *Nature Reviews Neuroscience*, *7*(7), 523–534.

Hibar, D., Westlye, L. T., Doan, N. T., Jahanshad, N., Cheung, J., Ching, C. R., Versace, A., Bilderbeck, A., Uhlmann, A., Mwangi, B., Krämer, B., Overs, B., Hartberg, C. B., Abé, C., Dima, D., Grotegerd, D., Sprooten, E., Bøen, E., Jimenez, E., … Andreassen, O. A. (2018). Cortical abnormalities in bipolar disorder: An MRI analysis of 6503 individuals from the enigma bipolar disorder working group. *Molecular Psychiatry*, *23*(4), 932–942.

Huang, B., Zhang, K., Zhang, J., Ramsey, J. D., Sanchez-Romero, R., Glymour, C., & Schölkopf, B. (2020). Causal discovery from heterogeneous/nonstationary data. *Journal of Machine Learning Research*, *21*, 1–53.

Huettel, S. A., Song, A. W., & McCarthy, G. (2004). *Functional magnetic resonance imaging* (Vol. 1). Sinauer Associates.

Kriegeskorte, N., Goebel, R., & Bandettini, P. (2006). Information-based functional brain mapping. *Proceedings of the National Academy of Sciences of the United States of America*, *103*(10), 3863–3868.

Long, Y., Liu, Z., Chan, C. K. Y., Wu, G., Xue, Z., Pan, Y., Chen, X., Huang, X., Li, D., & Pu, W. (2020). Altered temporal variability of local and large-scale resting-state brain functional connectivity patterns in schizophrenia and bipolar disorder. *Frontiers in Psychiatry*, *11*, 422.

Maddox, W. T. (1999). On the dangers of averaging across observers when comparing decision bound models and generalized context models of categorization. *Perception & Psychophysics*, *61*, 354–374.

Malinsky, D., & Danks, D. (2018). Causal discovery algorithms: A practical guide. *Philosophy Compass*, *13*, e12470. https://doi.org/10.1111/phc3.12470

McCaffrey, J., & Danks, D. (2018). Mixtures and psychological inference with resting state fMRI. *The British Journal for the Philosophy of Science*, *73*, 583–611.

Mitchell, T. M., Hutchinson, R., Niculescu, R. S., Pereira, F., Wang, X., Just, M., & Newman, S. (2004). Learning to decode cognitive states from brain images. *Machine Learning*, *5*, 145–175.

Nogueira, A. R., Pugnana, A., Ruggieri, S., Pedreschi, D., & Gama, J. (2022). Methods and tools for causal discovery and causal inference. *WIREs Data Mining and Knowledge Discovery*, *12*(2), e1449.

Norman, K. A., Polyn, S. M., Detre, G. J., & Haxby, J. V. (2006). Beyond mind-reading: Multi-voxel pattern analysis of fMRI data. *Trends in Cognitive Sciences*, *10*(9), 424–430.

O'Toole, A. J., Jiang, F., Abdi, H., Pénard, N., Dunlop, J. P., & Parent, M. A. (2007). Theoretical, statistical, and practical perspectives on pattern-based classification approaches to the analysis of functional neuroimaging data. *Journal of Cognitive Neuroscience*, *19*(11), 1735–1752.

Pearl, J. (2009). *Causality: Models, reasoning, and inference* (2nd ed.). Cambridge University Press.

Peterson, J. C., Bourgin, D., Agrawal, M., Reichman, D., & Griffiths, T. (2021). Using large-scale experiments and machine learning to discover theories of human decision-making. *Science*, *372*(6547), 1209–1214.

Poldrack, R. A., Halchenko, Y. O., & Hanson, S. J. (2009). Decoding the large-scale structure of brain function by classifying mental states across individuals. *Psychological Science*, *20*(11), 1364–1372.

Poldrack, R. A., Mumford, J. A., & Nichols, T. E. (2011). *Handbook of functional MRI data analysis*. Cambridge University Press.

Poline, J. B., Worsley, K. J., Evans, A. C., & Friston, K. J. (1997). Combining spatial extent and peak intensity to test for activations in functional imaging. *NeuroImage*, *5*(2), 83–96.

Power, J. D., Barnes, K. A., Snyder, A. Z., Schlaggar, B. L., & Petersen, S. E. (2012). Spurious but systematic correlations in functional connectivity MRI networks arise from subject motion. *NeuroImage*, *59*(3), 2142–2154.

Ramsey, J. D., Hanson, S. J., Hanson, C., Halchenko, Y. O., Poldrack, R. A., & Glymour, C. (2010). Six problems for causal inference from fMRI. *NeuroImage*, *49*(2), 1545–1558.

Ramsey, J. D., Spirtes, P., & Glymour, C. (2011). On meta-analyses of imaging data and the mixture of records. *NeuroImage*, *57*(2), 323–330.

Rashid, B., Arbabshirani, M. R., Damaraju, E., Cetin, M. S., Miller, R., Pearlson, G. D., & Calhoun, V. D. (2016). Classification of schizophrenia and bipolar patients using static and dynamic resting-state fMRI brain connectivity. *NeuroImage*, *134*, 645–657.

Rashid, B., Damaraju, E., Pearlson, G., & Calhoun, V. (2014). Dynamic connectivity states estimated from resting fMRI identify differences among schizophrenia, bipolar disorder, and healthy control subjects. *Frontiers in Human Neuroscience*, *8*, 897.

Rathore, S., Habes, M., Iftikhar, M. A., Shacklett, A., & Davatzikos, C. (2017). A review on neuroimaging-based classification studies and associated feature extraction methods for Alzheimer's disease and its prodromal stages. *NeuroImage*, *155*, 530–548.

Roebroeck, A., Formisano, E., & Goebel, R. (2005). Mapping directed influence over the brain using Granger causality and fMRI. *NeuroImage*, *25*, 230–242.

Roebroeck, A., Formisano, E., & Goebel, R. (2011). The identification of interacting networks in the brain using fMRI: Model selection, causality and deconvolution. *NeuroImage*, *58*, 296–302.

Rubin, D. B. (2005). Causal inference using potential outcomes: Design, modeling, decisions. *Journal of the American Statistical Association*, *100*(469), 322–331.

Samuel Schwarzkopf, D., & Rees, G. (2011). Pattern classification using functional magnetic resonance imaging. *WIREs Cognitive Science*, *2*(5), 568–579.

Sanchez-Romero, R., Ramsey, J. D., Zhang, K., Glymour, M. R., Huang, B., & Glymour, C. (2019). Estimating feedforward and feedback effective connections from fMRI time series: Assessments of statistical methods. *Network Neuroscience*, *3*(2), 274–306.

Seth, A. K., Chorley, P., & Barnett, L. C. (2013). Granger causality analysis of fMRI BOLD signals is invariant to hemodynamic convolution but not downsampling. *NeuroImage*, *65*, 540–555.

Smucny, J., Shi, G., & Davidson, I. (2022). Deep learning in neuroimaging: Overcoming challenges with emerging approaches. *Frontiers in Psychiatry*, *13*.

Spirtes, P., Glymour, C., & Scheines, R. (1993). *Causation, prediction, and search*. The MIT Press.

Tom, S. M., Fox, C. R., Trepel, C., & Poldrack, R. A. (2007). The neural basis of loss aversion in decision-making under risk. *Science*, *315*(5811), 515–518.

van Erp, T. G., Walton, E., Hibar, D. P., Schmaal, L., Jiang, W., Glahn, D. C., Pearlson, G. D., Yao, N., Fukunaga, M., Hashimoto, R., Okada, N., Yamamori, H., Bustillo, J. R., Clark, V. P., Agartz, I., Mueller, B. A., Cahn, W., de Zwarte, S. M. C., Hulshoff Pol, H. E., ... Turner, J. A. (2018). Cortical brain abnormalities in 4474 individuals with schizophrenia and 5098 control subjects via the enhancing neuro imaging genetics through meta analysis (enigma) consortium. *Biological Psychiatry*, *84*(9), 644–654.

VanderWeele, T. (2015). *Explanation in causal inference: Methods for mediation and interaction*. Oxford University Press.

Vilares, I., Wesley, M. J., Ahn, W. Y., Bonnie, R. J., Hoffman, M., Jones, O. D., Morse, S. J., Yaffe, G., Lohrenz, T., & Montague, P. R. (2017). Predicting the knowledge–recklessness distinction in the human brain. *Proceedings of the National Academy of Sciences of the United States of America*, *114*(12), 3222–3227.

Weinberger, D. R., & Harrison, P. (2011). *Schizophrenia*. John Wiley & Sons.

Wen, D., Wei, Z., Zhou, Y., Li, G., Zhang, X., & Han, W. (2018). Deep learning methods to process fMRI data and their application in the diagnosis of cognitive impairment: A brief overview and our opinion. *Frontiers in Neuroinformatics*, *12*, 23.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

How to cite this article: Danks, D., & Davis, I. (2023). Causal inference in cognitive neuroscience. *WIREs Cognitive Science*, *14*(5), e1650. https://doi.org/10.1002/wcs.1650