Limited Functional Form, Misspecification, and Unreliable Interpretations in Psychology and Social Science

Matthew J. Vowels
Centre for Computer Vision, Speech, and Signal Processing
University of Surrey
Guildford, Surrey, UK
m.j.vowels@surrey.ac.uk

Abstract

The replicability crisis has drawn attention to numerous weaknesses in psychology and social science research practice. In this work we focus on three issues that deserve more attention: The use of models with limited functional form, the use of misspecified causal models, and unreliable interpretation of results. We demonstrate a number of possible consequences via simulation, and provide recommendations for researchers to improve their research practice. We believe it is extremely important to encourage psychologists and social scientists to engage with the debate surrounding areas of possible analytical and statistical improvements, particularly given that these shortfalls have the potential to seriously hinder scientific progress. Every research question and hypothesis may present its own unique challenges, and it is only through an awareness and understanding of varied statistical methods for predictive and causal modeling, that researchers will have the tools with which to appropriately address them.

Introduction

Meta-researchers have increasingly drawn attention to the replicability crisis affecting psychology and social science (Oberauer & Lewandowsky, 2019; Botella & Duran, 2019; Aarts et al., 2015; Stevens, 2017; Marsman et al., 2017; Shrout & Rodgers, 2018; Yarkoni, 2019). A key element of the crisis relates to common and fundamentally problematic analytic and statistical practices, some of which deserve more attention. These problematic practices relate to observational research and modeling in psychology and social science, and may be broadly categorized as issues with (1) the use of statistical/predictive models with limited functional form; (2) the misspecification of causal models; and (3) unreliable interpretations of predictive or causal models. All of these issues affect a researcher's ability to accurately model some aspect of the joint distribution of the data, for the purpose of predicting an outcome, estimating a causal effect, and drawing scientific conclusions. The first issue relates to the ubiquitous use of linear models, and a failure to consider more powerful, possibly data-adaptive techniques for both predictive and causal modeling. The second relates to the use of misspecified implicit (e.g. multiple linear regression) or explicit (e.g., structural equation) causal models which do not sufficiently reflect the true structure in the data. The final issue relates both to how predictive models are often (mis)interpreted as causal models, and vice versa, and also to how these interpretations are likely to be unreliable given the models' underlying limitations and assumptions.

We address the three issues in turn through both didactic illustration and simulation, and make a number of recommendations for improving research practice. While these issues relating to research practice have been previously discussed (e.g., see Claesen et al. 2019; Scheel et al. in press), we believe it is extremely important to continue to encourage and stimulate consideration and engagement with the debate surrounding areas of possible analytical improvement. Furthermore, in spite of researchers having already made important recommendations for improving practice (e.g., Lakens et al. 2016; Scheel et al. in press; Gigerenzer 2018; Jostmann et al. 2016; Lakens & Evers 2014; Orben & Lakens 2020) we see relatively little change in the research communities of psychology and social science (Claesen et al., 2019; Scheel et al., in press). In general, it would seem that researchers in psychology and social science lack some competence in the practice of prediction and causal inference, and these shortfalls have the potential to seriously affect the reliability and interpretation of research and therefore to hinder scientific progress.

Following a review of the literature, the paper is split into three main parts. In Part 1, we describe how the typical models used in psychology are limited by their functional form and discuss the implications of this issue and possible ways to address it. Part 2 is concerned with misspecification in causal modeling, and how the typical models used in psychology and social science do not adequately reflect the true structure of the data. We discuss how this impacts interpretability, how a consideration for causal structure is essential when designing a model, and we identify some challenges associated with undertaking causal modeling. Part 3 introduces the notion of explainability as an alternative to interpretation, and as a means of deriving insight from predictive models. We discuss interpretation, considering the relevant points on limited functional form and misspecification covered in Parts 1 and 2, and discuss how interpretations in psychology and social science tend to be a conflation of causal and predictive interpretations. Finally, we conclude this work with a discussion and by proposing four recommendations for improving practice. A table of relevant terms and their working definitions is provided for convenience in Table 1.

Table 1: Basic working definitions.

Approach	Relating closely to the hypothesis/research question, it describes the broad intention behind research methodology, analysis, and interpretation.
Model	Part of the approach, it is the mathematical relationship between variables, as reflected in the algorithm or technique used for analysis. It may be predictive or causal, or a hybrid.
Predictive	The "study of the association between variables or the identification of the variables which contribute to the prediction of another variable" (Blanca et al., 2018). The word "association" here alludes to the fact that the associations or relationships between variables are not necessarily causal. Prediction may help us to answer questions such as 'when?', 'which?', and 'how much?'.
Causal	The study of cause-effect relationships between variables, which facilitates understanding and answers questions such as such as 'why?', 'how?', and 'what if?' (Pearl, 2009)
Functional Form	The nature of the mathematical function describing the relationship between variables.
Misspecification	When a causal model does not sufficiently reflect the causal structure of the data to render a causal effect identifiable (due to limited functional form and/or incorrect causal structure), the model is misspecified.

Background

A recent article titled 'Declines in religiosity predict increases in violent crime - but not among countries with relatively high average IQ' was retracted from the Journal of Psychological Science on the basis of methodological weaknesses and political sensitivity. The Editor in Chief at the time, Steve Lindsay apologized on multiple grounds, and stated that "In terms of science, Clark et al. may not be worse than some other articles published in Psych Science during my editorship..." (Lindsay, 2020). This may suggest that methodological weakness, as described in terms of "blurred distinctions between psychological constructs versus measures and speculations/extrapolations far removed from the data" is somewhat par for the course in the "young science" (Lindsay, 2020) of psychology. Indeed, over the last ten years, meta-researchers have drawn increasing attention to a purported crisis in the human sciences (particularly psychology) known as the replication crisis. The crisis has been discussed at length by many different meta-researchers (e.g., Oberauer & Lewandowsky 2019; Botella & Duran 2019; Aarts et al. 2015; Stevens 2017; Marsman et al. 2017; Shrout & Rodgers 2018; Yarkoni 2019) who argue that research in the human sciences fails to replicate. For example, only six out of 53 landmark cancer studies were found to replicate (Begley & Ellis, 2012), and between one third and one half of 100 psychology studies in top-ranking journals could be replicated (Aarts et al., 2015; Marsman et al., 2017).

One positive outcome of the widespread awareness of the replicability crisis is the fact that attention has been drawn to many questionable, suboptimal, or problematic aspects associated with the research procedure in general. Indeed, it is only by recognition of these issues, and engagement in relevant constructive debate, that research practice can be improved. A wide range of contributing factors to this crisis have been highlighted and discussed, and include: A lack of understanding about and misuse of p-values and statistical tests (Cassidy et al., 2019; Gigerenzer, 2018; 2004; Colquhoun, 2014; 2017; 2019; McShane et al., 2019); issues relating to the testing of theory (Oberauer & Lewandowsky, 2019; Muthukrishna & Henric, 2017); immature theories (Scheel et al., in press); misunderstandings about statistical power and low sample sizes (Sassenberg & Ditrich, 2019; Baker et al., 2020; Correll et al., 2020); measurement problems (Flake & Fried, in press); a lack of meta-analyses (Schmidt & Oh, 2016); a lack of assumptions testing (Ernst & Albers, 2017); pressure to publish (Shrout & Rodgers, 2018); double-dipping and overfitting (Kassraian-Fard et al., 2016; Kriegeskorte et al., 2009; Mayo, 2013; Yarkoni & Westfall, 2017); a failure to consider the consequences of aggregation and non-ergodicity (Fisher et al., 2018; Peters & Werner, 2017); academia and research being a strategy game with unscientific incentives (Gigerenzer, 2018; DeDeo, 2020); a reluctance of journals to publish replications (Martin & Clarke, 2017; Gernsbacher, 2019); issues with the peer review process (Heesen & Bright, 2020); reporting errors (Nuijten et al., 2016); a lack of research practice standardization (Tong, 2019); overly generous claims and warped interpretations (Yarkoni, 2019; Spellman, 2015; Scheel et al., in press); the conflation of predictive and causal approaches and interpretations (Grosz et al., 2020; Yarkoni & Westfall, 2017; Shmueli, 2010); and general scientific misconduct (Stricker & Günther, 2019).

More specifically, meta-researchers have highlighted how psychologists and social scientists tend to mix causal and predictive language (Grosz et al., 2020). For instance, in a select review of psychology literature, Grosz et al. (2020) explain how “some parts of the articles read as if the entire endeavor were noncausal; yet other parts make sense only in the context of trying to answer a causal research question”. Similarly, meta-researchers have drawn attention to how it is common for psychology and social science researchers to use associational/predictive techniques to test otherwise causal hypotheses (Shmueli, 2010; Yarkoni & Westfall, 2017; Glymour, 1998; Hernan, 2018a; Grosz et al., 2020). Shmueli (2010) explains how “the type of statistical models used for testing causal hypotheses in the social sciences are almost always association-based [i.e., predictive] models.” One can only surmise the possible causes behind this tendency for conflation, but it may relate to the controversial history of causal inference in observational social science and psychology. The conflation may stem from the conflict between recognizing the importance of asking causal questions, without wanting to be seen to be actually using causal methods with observational data. Indeed, the literature on causality in psychology and social science has been described as “one of the oddest literatures in all of academia” (Dowd, 2011), and researchers in these fields are notoriously reluctant to adopt appropriate modeling techniques (Grosz et al., 2020; Hernan, 2018a). Others have mocked the reluctance to undertake causal inference in psychology and the social sciences by referring to causality as “the C-word” (Hernan, 2018a;b), and others refer to its use as “taboo” (Grosz et al., 2020). Indeed, Grosz et al. (2020) explain how causal modeling is only undertaken “implicitly, opaquely, and without an articulation of the underlying assumptions”. The result has been a tendency to use predictive language such as ‘associations’, ‘links, ‘correlations’, ‘relationships’, and to avoid causal language such as ‘causes’, ‘impacts’, ‘effects’ despite designing their models and experiments on the basis of well considered theories about the causal structure of the phenomenon of interest (Shmueli, 2010).

In addition to a general reluctance to adopt clearly articulated causal approaches, one might also argue that the various manifestations of conflation indicate a lack of understanding about the differences between predictive and causal modeling (Yarkoni & Westfall, 2017; Shmueli, 2010; Grosz et al., 2020). For example, there is a relatively well established modeling technique known as Structural Equation Modeling (SEM) which explicitly encodes causal structure (Kline, 2005; Blanca et al., 2018). The point to note about the use of SEM in psychology and social science is that the way the technique is often presented and interpreted obfuscates its causal nature (Grosz et al., 2020). This leads to an awkward conflation of causal modeling with predictive interpretations, resulting in ambiguity and a lack of clarity regarding intentions and assumptions. It may be that researchers are unaware that their SEMs are explicitly causal and fail to sufficiently understand how the results from the analysis are underpinned by a number of restrictive (and often untestable) assumptions.

There is also evidence of a possible lack of understanding relating to the use of predictive models in psychology and social science. Yarkoni & Westfall (2017) provide a number of examples of where researchers seem to have clearly identified that they are adopting a predictive approach but use suboptimal and misguided predictive modeling practice and models which lack predictive power. A wide range of powerful predictive modeling techniques exist, including neural networks (Goodfellow et al., 2016), random forests (Breiman, 2001b), gradient boosting machines (Chen & Guestrin, 2016) etc., many of which derive from developments in machine learning. In spite of the abundance of available options, researchers in psychology and social science most often employ simple linear models when undertaking predictive /associational research (Yarkoni & Westfall, 2017; Blanca et al., 2018). The assumption of linear functional form is often restrictive and has been previously noted to be problematic (van der Laan & Rose, 2011; Asuero et al., 2006; Onwuegbuzie & Daniel, 1999; Achen, 1977; King, 1986; Meehl, 1990; Taleb, 2019) and frequently ignored (Ernst & Albers, 2017). Furthermore, some researchers seem to be unaware of certain basic principles relating to predictive (as well as causal) research, such as those relating to overfitting (Yarkoni & Westfall, 2017; Bishop, 2006; Heman & Slep, 2001) and 'double-dipping' (Kassraian-Fard et al., 2016; Kriegeskorte et al., 2009; Mayo, 2013). Overfitting and double-dipping refer to modeling (mis)practices which increase the fit of a model to the specific data sample being used, and which negatively impact the validity and generalizability of results. Indeed, *any* modeling decision that affects the parameters of the model based on information from the same data sample with which the model is validated results in overfitting, biased effect sizes, and the inflation of p-values and other performance metrics (Bishop, 2006; Yarkoni & Westfall, 2017; Heman & Slep, 2001). Regardless of whether a researcher is undertaking a predictive or causal approach, overfitting inflates the apparent success of the mapping function at the expense of generalizability to new samples, and has been argued to be a major contributor to the current replicability crisis (Shrout & Rodgers, 2018; Gelman & Loken, 2013).

Part 1: Limited Functional Form - Modeling Relationships Between Variables

In this part we address certain issues that may arise when using modeling techniques that have limited functional form. When we refer to the functional form of a model as being limited we mean that the model does not have the flexibility to sufficiently reflect the complexity of the relationship between variables, possibly resulting in poor predictive ability and biased results. Identifying or deriving an adequately flexible functional form with which to model the relationship between variables, in circumstances where causal relationships are not of concern, is somewhat synonymous with the task of prediction. As such, the majority of this section will be written with consideration of its relevance to predictive modeling, where the goal is to learn a function that optimally maps predictor variables to outcome variables. However, a consideration for functional form is just as important for causal modeling, for which we may be tasked with embedding models representing the relationships between variables into a larger model representing the causal structure of the data generating process. For purposes of prediction alone, it suffices to be solely concerned with finding the optimal mapping function to achieve some desired level of predictive performance. We expect models that reflect the structure of reality to also be good predictors, but this is not necessarily the case the other way around; good predictive functions do not necessarily reflect the structure of reality.

We begin by introducing some of the technical formalism behind predictive modeling, and briefly list some of its wide ranging applications. Following this, we discuss the limitations of undertaking prediction using the two most common and basic methods used in psychology and social science: Correlation and linear regression models. We demonstrate how these methods, in the basic form adopted in psychology and social science, are fundamentally limited in their ability to account for non-linearities present in the data. This motivates a need for more flexible, powerful, potentially data-adaptive predictive methods. Previous research has highlighted that the use of such techniques is rare in psychology and social science, where it is much more usual to use models with restrictive linear functional form (Yarkoni & Westfall, 2017; Blanca et al., 2018). Linear functions may be useful to consider for their computational efficiency and for their tendency to naturally *under-fit* the data, thereby improving generalization particularly when the quantity of data is limited. However, these factors are not sufficient to fully explain the rarity of non-linear, powerful, and/or data adaptive techniques in psychology and social science, and we posit that a possible lack of awareness of these alternative methods is more likely.

Applications and Basic Formalism

The topic of identifying the optimal functional form with which to represent the relationship between variables is vast and well covered by many authors, particularly those in the field of machine learning in the context of prediction (Bishop, 2006; Duda et al., 2001; Murphy, 2012). Prediction has been described as "the study of the association between variables or the identification of the variables which contribute to the prediction of another variable" (Blanca et al., 2018) and therefore relates closely to the more general task of identifying the optimal function that maps between sets of variables. The applications for predictive models are wide ranging, and include personalized medicine (Rahbar et al., 2020), time series forecasting (Makridakis et al., 2020), facial and object recognition (Krizhevsky et al., 2012; Jonsson et al., 2000), and many others. Such techniques are therefore extremely valuable and influential in shaping our modern world.

The basic formalism for predictive modeling is as follows: Researchers may be confronted with a dataset comprising samples from a population $(\mathbf{x}_i, \mathbf{y}_i) \in \mathcal{X} \times \mathcal{Y}$. In words, we have a set of samples of predictors or random variables[1], which take on values in the set $\mathcal{X}$ and which are related to some outcome variables[2] which take on values in the set $\mathcal{Y}$. If the outcome is binary or categorical, the task of prediction becomes equivalent to one of classification. The goal of prediction usually involves finding a mapping function $f : \mathcal{X} \rightarrow \mathcal{Y}$. We will use the terms *predictive function* and *predictive model* to refer to the mapping function used to make predictions.

The Common Assumption of Linear Functional Form

Variations on simple measures of correlation and linear models (including linear SEMs) were found to be the most frequently used modeling techniques in psychology research in recent years (Blanca et al., 2018; Bolger et al., 2019).[3] The principal assumption associated with these models is that the true relationships between the variables are sufficiently represented as linear. Such models therefore have a limited functional form that can only represent linear relationships. In other words, they describe relationships between predictor and outcome variables that can be summarized in terms of a weighted sum. Of course, in reality the true relationships between variables may be highly complex and nonlinear. Indeed, assuming our dataset is sampled from a 'true' population distribution, there exists a 'true' functional form describing the functional relationships between the variables. Figure 1 illustrates how traditional methods (including linear regression) have the most limited capacity (owing to strong restrictions on the functional form) to model complex real-world phenomena (Coyle et al., 2020; van der Laan & Rose, 2018; van der Laan & Starmans, 2014).

Correlation is generally used to measure the association or statistical dependence between variables (i.e., to identify variables which may be good predictors). As one of the most common ways to measure dependence, there are two important aspects relating to correlation to bear in mind. Figure 2 shows a number of bivariate distributions along with their correlation coefficient. The first thing to note from the upper six plots is that correlation itself is a non-linear metric for dependence. Lower values of the Pearson Correlation Coefficient (PCC) are associated with a disproportionately lower dependence than higher values (and this is also reflected visually in the plots). The second thing to note from the lower four plots is that the PCC catastrophically fails to capture non-linear dependence.

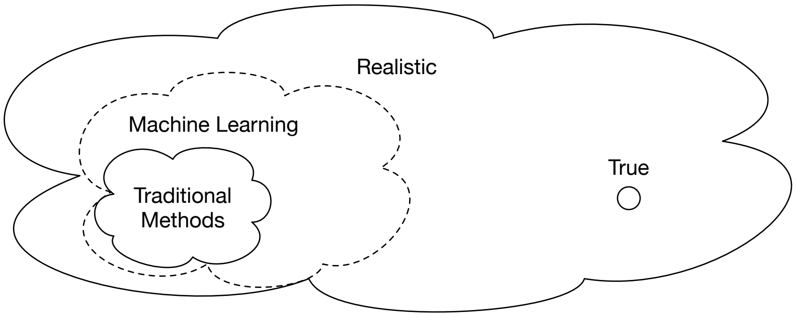
The first issue is important for researchers to understand when drawing conclusions about relative levels of correlation. For example, the difference between $PCC = 0.1$ and $PCC = 0.2$ is less dramatic than, say the difference between $PCC = 0.8$ and $PCC = 0.9$, in spite of the former describing a much higher proportionate increase. The second issue relates to an assumption of linearity: If the relationship between the two variables is linear, then correlation provides a measure of linear

[1] These variables are sometimes called 'independent variables', but due to the fact that they are usually non-independent, we avoid this potentially unhelpful terminology.

[2] These variables are sometimes called 'dependent variables', but due to the fact that many dependencies exist we also avoid this terminology.

[3] It might be argued that any arbitrary function can be represented as some linear sum of features, and that therefore all models are fundamentally linear. However, using such a broadly encompassing definition of the term 'linear model' makes discussion pedantic. As such, we use the term to describe the typical linear regression model where the outcome is modeled as a linear sum of raw variables or low-order functions of these variables (such as exponents: $\mathbf{x}^1, \mathbf{x}^2$; and interactions: $\mathbf{x}_1\mathbf{x}_2$ etc.).

Figure 1: Approximating Realistic Data Distributions



Note. Traditional techniques such as linear regression may be severely limited in their capacity to model highly complex, non-linear data. Machine Learning methods may help to expand the coverage of realistic data distributions, but the true distribution may still lie outside. Combining flexible function approximation techniques from machine learning, with an incorporation of domain knowledge and model structure, can help us get as close as possible to modelling the true data distribution (van der Laan & Rose, 2011).

dependence; if the relationship is non-linear, then correlation may provide meaningless measures of dependence. In cases where the relationship is non-linear, researchers will need to either linearize the relationship (e.g., by creating a new variable that accounts for this non-linearity), or consider using an alternative measure of dependence. One such alternative to correlation is Shannon Mutual Information (M.I.), which gives us a measure for how much information one variable contains about another (Cover & Thomas, 2006; Kraskov et al., 2004; Steeg & Galstyan, 2012; 2013; Gao et al., 2015; Kinney & Atwal, 2014). The estimates for M.I. are also shown in Figure 2, and it can be seen that M.I. not only handles non-linear relationships between variables, but also increases linearly with the degree of dependence of the variables. Note that M.I. ranges between $[0, H]$ where H is the entropy of either distribution when the two distributions are identical (i.e., $I(\mathbf{x}, \mathbf{y}) = H(\mathbf{x}) = H(\mathbf{y})$ when $\mathbf{x} = \mathbf{y}$).[4] M.I. cannot be negative, and as such it is not able to indicate the 'direction' of the association in the way that correlation can. However, this is an acceptable limitation given that many non-linear relationships are non-monotonic (i.e. they are not always either increasing or decreasing) and in such cases a notion of positive or negative direction is unhelpful.

Linear regression is another very common modeling technique used for both predictive and causal modeling and constitute a relatively small sub-class in the class of Generalized Linear Models. There is one principal assumption for linear regression which is important for achieving both successful causal *and* predictive modeling. Namely, that the outcome can be well approximated using a weighted linear sum of the input variables. Indeed, the linearity imposes a strong functional constraint that restricts the function's flexibility and is, therefore, an assumption about *functional form* (van der Laan & Rose, 2011). Linear methods are unlikely to match the functional form of realistic data distributions, and to get closer to the true functional form, researchers should consider using more flexible predictive methods.

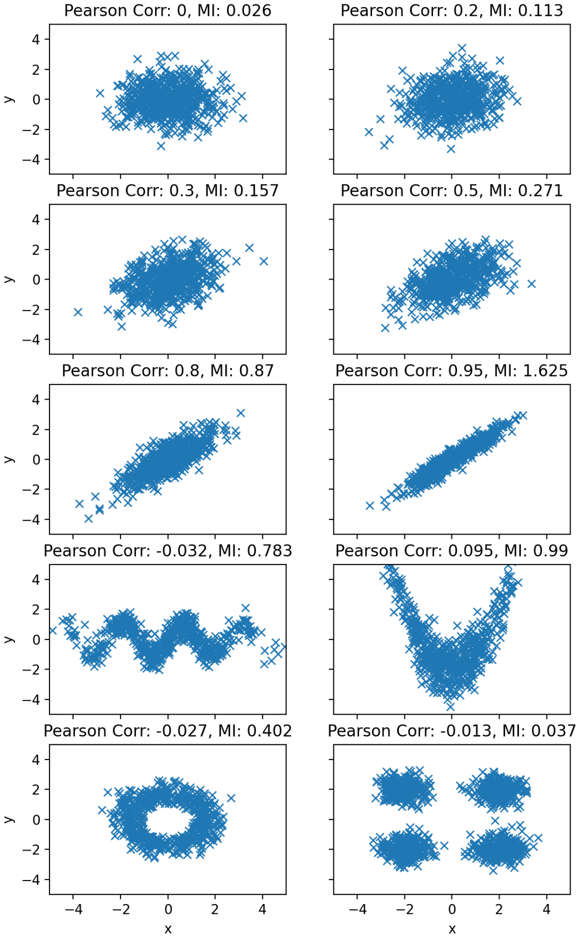
Improving on the Functional Form of Linear Models

In order to improve the predictive or associational performance of a predictive function, researchers may need to explore either *feature engineering* approaches, or other functional approximation techniques such as those commonly used in machine learning. Introducing hierarchical structure within linear functions can improve the fit (Yarkoni, 2019; Gelman & Hill, 2007; Bolger et al., 2019), but even hierarchical linear models are constrained according to linear functional associations.

Feature engineering involves the substitution of raw input variables with functions of these raw variables called *features*. Depending on the functional form used to derive these features, the features themselves may then be linearly related to the outcome, facilitating better overall functional approximation. For instance, researchers may include more exotic basis functions such (e.g., sinusoidal functions; Vowels et al. 2018, or kernels; Scholkopf 2019), or simply combine features to form new

[4] Readers are pointed to Cover & Thomas (2006) for an introduction to information theoretic concepts such as entropy and mutual information.

Figure 2: Pearson Correlation and Shannon Mutual Information.

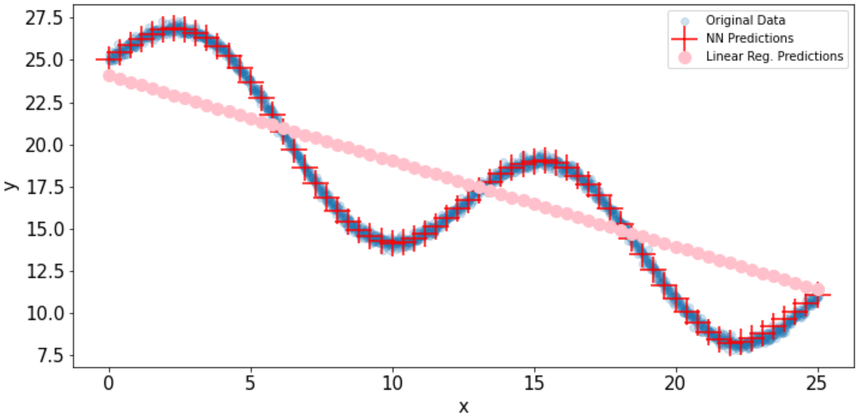


Note. Simulations demonstrating the relationships between the Pearson measure of correlation, and the Mutual Information metric for measuring statistical dependence. The upper six plots depict linear bivariate relationships, whereas the lower four plots are non-linear.

ones (e.g., interaction features which are composed by multiplying two variables together). Feature engineering may thereby help to account for the non-linearities of the data in the features themselves, but in doing so, each feature may need to be carefully chosen or designed. For example, in Figure 2, the plot in the third row on the right has a simple basis function which is $\mathbf{x}^2$. While the raw values of $\mathbf{x}$ could not be used to model the outcome as part of a linear sum, the squared values could be used to essentially linearize the predictor in question. However, in real-world applications (i.e., research scenarios with real data) we will not know the functional form *a priori* and it may be difficult to ascertain. For instance, the function may not be an exact quadratic function $\mathbf{x}^2$, but some other, arbitrarily complex function. The feature engineering process may or may not be guided by knowledge about the domain of interest. For example, in the case of a time series with known seasonal variation (e.g., financial data exhibiting fluctuation due to the business cycle) the use of sinusoidal basis functions may be well justified and aid prediction and generalization (Hamilton, 1994; Vowels et al., 2018).

Besides generalized linear models with feature engineering, there exist many alternative and much more powerful function approximation techniques, such as those common in machine learning. These techniques are able to *learn* functional relationships from the data themselves and can be used instead of, or in combination with, feature engineering. For instance, random forests (Breiman, 2001b) comprise a group of decision trees that are capable of learning highly non-linear relationships and interactions between variables, without these interactions needing to be pre-specified. The mapping learned by the forest adapts to the data in order to minimize a performance objective (e.g., mean squared error). One of the advantages of random forests is that they employ bootstrapping and thereby mitigate problems with learned functions overfitting the data. Neural networks are an alternative approach to function approximation which are also data-adaptive and are highly parameterized (Goodfellow et al., 2016). They learn by iteratively updating their parameters according to an error signal until some criterion for convergence is met. An example of predictions from a simple neural network compared with those of a linear regressor on a bivariate problem is shown in Figure 3. It can be seen the neural network has fit the data almost perfectly, whilst the linear regression approximates the mean slope of the line, ignoring the cycling fluctuation. While prior knowledge may enable one to employ sinusoidal basis functions with linear regression in order to achieve a similar degree of fit, the advantage with neural networks is that such prior feature engineering is not required, and any arbitrary function can be approximated (Hornik et al., 1989).

Figure 3: Neural network versus linear regression function predictions.



Note. Demonstrates how linear functional forms cannot capture the non-linear relationships. In contrast, non-linear, data-adaptive techniques such as neural networks, can.

Overfitting and Double-Dipping

As described previously overfitting and double-dipping refer to the consequences of various modeling practices which increase the fit of a model to a specific data sample, but which negatively impact the validity and generalizability of results. An awareness of overfitting becomes increasingly crucial when attempting to model non-linear functional relationships between variables. Overfitting and double-dipping have been extensively covered elsewhere, particularly in the machine learning literature (where overfitting is sometimes associated with what is known as the bias-variance trade-off) (Belkin et al., 2019; Bishop, 2006; Murphy, 2000; Yarkoni & Westfall, 2017; Mayo, 2013). Prior research has highlighted how modeling practices that result in overfitting are common in psychology and social science, as well as a number of other fields, and have been noted for their possible contribution to the replicability crisis (Shrout & Rodgers, 2018; Gelman & Loken, 2013; Yarkoni & Westfall, 2017). Even the common forward and backward method for variable inclusion constitutes data-driven overfitting practices which have the potential to significantly impact model generalizability and interpretability, and yet these practices are routinely included as part of standard statistical education and practice in psychology (e.g., see Field (2009)). We mention such (mis)practice again here because, when using powerful function approximation techniques, a consideration for overfitting is even more important. There are numerous techniques for mitigating issues with overfitting, including regularization, cross-validation, train-test splits etc. and it is important that researchers in psychology and social science familiarize themselves with these fundamental concepts, especially when accounting for complex, non-linear associations between variables.

Part 1 Summary

In Part 1, we presented how models with limited functional form may be unable to represent the complex relationships between variables. The typical analyses used in psychology and social science include simple measures of correlation, and various manifestations of linear regression. While such modeling techniques are limited in their predictive capacity, there are many algorithms used in the field of machine learning which can learn an appropriately flexible functional form from the data themselves. When using more powerful techniques, it is especially important to validate models on an out-of-sample test set (e.g., by using a cross-validation method, or train/test splitting) in order to avoid overfitting. However, it is worth noting that overfitting (and the related problem of double-dipping) is also possible with simple linear models, and prior meta-research suggests that researchers may be unaware of these issues. Finally, the rarity of modeling techniques with powerful, data-adaptive functional form represents a possible missed opportunity in psychology and social science, and we encourage researchers to consider the functional form of their models, and familiarize themselves with the associated pitfalls and limitations (e.g., overfitting), in order that they can get closer to modeling the true relationships underpinning the phenomenon under study.

Part 2: Causal/Structural Misspecification

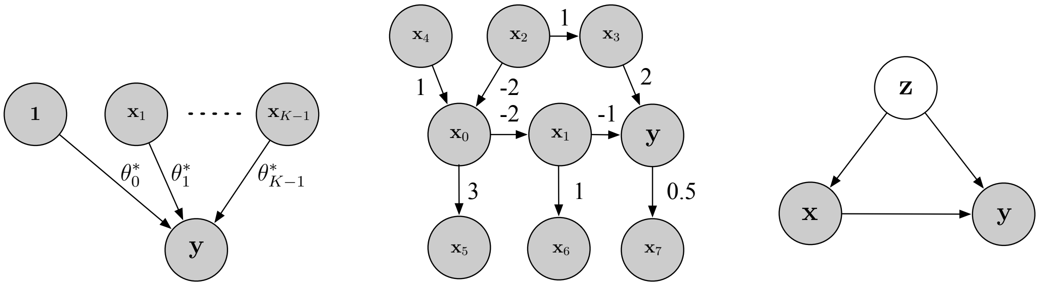
As described in the Introduction, prior research has highlighted a reluctance to adopt explicit causal approaches (Grosz et al., 2020; Hernan, 2018a). Causal techniques provide the means to answer fundamental questions that help us to develop an understanding of the world (Pearl, 2009; van der Laan & Rose, 2011). To the best of our knowledge, we are not aware of a well-established theory in psychology or social science which does not incorporate at least some level of consideration for cause and effect, and, if there is one, we would question its utility in so far as it can help us understand the world. Models which sufficiently align with the structure of reality may facilitate causal inference, even with observational (as opposed to experimental) data (Glymour, 2001; Pearl, 2009; Pearl et al., 2016; Grosz et al., 2020) and have wide ranging applications including advertisement (Bottou et al., 2013), policy making (Kreif & DiazOrdaz, 2019), the evaluation of evidence within legal frameworks (Pearl, 2009; Siegerink et al., 2016), and the development of medical treatments (Petersen et al., 2017; van der Laan & Rose, 2011). There are a number of challenges associated with adopting a causal approach.

Misspecification represents one of the principal challenges associated with causal inference, and arises when the true causal structure and/or the functional form of the relationships between variables in the data generating process are not sufficiently reflected in a causal model. Misspecification results in biased effect size estimates which are not meaningfully interpretable. In this Part, we primarily focus on misspecification stemming from problems associated with structure and to do so, we consider misspecfication in restricted linear settings. As we will show, even in this restricted setting, it is extremely important that the model sufficiently accounts for the true structure of the data in order that the resulting model is interpretable. This section is not intended as a technical guide to undertaking causal inference in general (for more information on causal inference see e.g., Pearl 2009; Petersen et al. 2017; Pearl et al. 2016; Glymour 2001; Angrist & Krueger 2001; Rubin 2005; Gelman & Hill 2007).

Recovering Causal Effects

Given the frequency with which psychologists and social scientists adopt linear regression methods to test causal theories (Shmueli, 2010; Blanca et al., 2018), it is extremely important that researchers understand the structural bias associated with the use of such models. In this section, we demonstrate how typical linear regression models used in psychology and social science impose a strong implicit causal/structural form which is unlikely to reflect the true causal structure of the data, even when the functional form is linear, and are therefore likely to be misspecified. We show that, through a consideration of the causal structure of the phenomenon under study, one can nonetheless use linear regression to recover causal effects under a number of restrictive assumptions.

Figure 4: Simple Directed Acyclic Graphs



(a) (b) (c)

Note. Example causal Directed Acyclic Graphs (c-DAGs). Example (a) depicts the case where all 'predictor' or causal variables are exogenous (i.e., they have no causal parents and are independent of each other). This corresponds with the causal structure of a simple multiple regression, where the dependent outcome $\mathbf{y}$ is a linear sum of the $\mathbf{x}$ variables. The empirical causal effect of each variable is equivalent to the multiple regression coefficient estimates. Example (b) is adapted from Peters et al. 2017. Example (c) depicts a graph with an unobserved confounding variable $\mathbf{z}$.

Multiple Regression Without Misspecification

In this section we demonstrate the strong, implicit structural form associated with multiple regression. We begin by demonstrating that multiple linear regression (in its basic form) is not misspecified with respect to the true data generating process when all predictors are exogenous (see structure in Figure 4(a)). In such a scenario, the resulting model is interpretable.

If the true data generating process could be described as a weighted sum of a set of input variables, then our goal of prediction within the Ordinary Least Squares multiple linear regression framework would also be adequate for causal modeling, causal parameter estimation, or causal inference. Such a model might be depicted graphically as in Figure 4(a). In this scenario, there would exist parameters $\boldsymbol{\theta}^*$ (also known as effect sizes) which represent the true causal parameters, and our OLS-derived parameters would represent empirical/sample estimations thereof.

The graphs in Figure 4 are known as causal Directed Acyclic Graphs (c-DAGs), and they represents a generalization of the graphical representation often used in Structural Equation Modelling (SEM) (Pearl, 2009; Koller & Friedman, 2009; Rohrer, 2018). The arrows indicate causal directional relationships between variables, parameterized by θ, and the grey nodes indicate observed variables. The *acyclicity* pertains to the restriction that there can be no closed loops (i.e., feedback) in the graph. Graph terminology (e.g., 'parent', 'ancestor', 'descendant', 'child') is useful in describing the top-level relationships between variables. For example, a node with an incoming arrow is a child of its parent variable, and further upstream or downstream variables are ancestors or descendants respectively.

In general, the arrows in a c-DAG indicate causal dependencies, and there is no implied functional form that prescribes how the variables are combined at a node (i.e., there could be highly non-linear, adaptive functions with interactions). Furthermore, the nodes represent variables which may or may not be univariate or parametric. In other words, a node labelled $\mathbf{x}$ does not restrict the dimensionality or (non-)parameterization of $\mathbf{x}$ itself. For instance, a node $\mathbf{x}$ could comprise multiple predictors which do not conform to a parameterized distribution. Hence, c-DAGs encode the fundamental essence of the causal structure, without imposing potentially irrelevant restrictions. We have included some extra information in the c-DAG of Figure 4(a) for the sake of demonstration. This particular c-DAG represents the intercept parameter of a multiple linear regression as a vector of ones multiplied by the parameter θ_0^*. The structural equations for this graph may be represented in Equation 1:

$$
\begin{aligned}
\mathbf{x}_{k=0} &:= \mathbf{1} \\
\mathbf{x}_k &:= \mathbf{U}_k(0,1) \text{ for } k = 1, ..., (K-1) \\
\mathbf{y} &:= \sum_{k=0}^{K-1} \theta_k^* \mathbf{x}_k + \mathbf{U}_y(0,1) \text{ for } k = 0, ..., (K-1)
\end{aligned}
\tag{1}
$$

Let us assume that $\mathbf{U}_k$ and $\mathbf{U}_y$ are N-dimensional vectors of identically and independently distributed (i.i.d.) normally distributed random noise. The ':=' symbol (endearingly referred to as the walrus operator in the Python programming world) denotes *assignment* rather than equality. This distinction between assignment and equality is useful in reflecting the structural/causal direction of the arrows in the c-DAG. For example, the outcome $\mathbf{y}$ is a function of its inputs, and the equation should not be rearranged to imply that the inputs are a function of the outcome (the arrows point in one direction). Equation 1 encode the fact that all the input variables are exogenous (i.e. completely independent of each other and determined only by i.i.d. noise) and that the outcome is determined by a weighted linear combination of these variables. In this setting we might understandably refer to the input variables as the independent variables, and the outcome as the dependent variable. As mentioned, these equations correspond with a simple multiple linear regression and can be solved to find $\boldsymbol{\theta}$ using OLS. We demonstrate this by undertaking a simulation for $K = 4$ with $\theta_0^* = 3.3$, $\theta_1^* = 0.1$, $\theta_2^* = 0.3$ and $\theta_3^* = 0.5$. We set $N = 5000$ so that we do not have to be concerned about the stochastic variability associated with small samples, and the results are shown in Table 2.

Table 2: Estimated parameters for DAG in Figure 4(a).

	$\hat{\theta}_0$	$\hat{\theta}_1$	$\hat{\theta}_2$	$\hat{\theta}_3$
$\mathbf{y}$	3.31	0.11	0.31	0.50

From this demonstration it can be seen that the OLS regression successfully recovered $\hat{\boldsymbol{\theta}}$ close to $\boldsymbol{\theta}^*$. In this case, the data generating process directly matched the model we used to estimate the parameters and was therefore *not* misspecified. When there is no misspecification, the estimated parameters may be interpreted as *causal* parameters that tell us about the phenomenon (in this case, a simple, simulated phenomenon). Indeed, the parameters here can be interpreted as 'one unit increase in $\mathbf{x}_1$ yields a θ_1 increase in $\mathbf{y}$', as is common practice in psychology and social science.

The interpretability of the model was only possible because the structure of the data matched the structure of a multiple linear regression, equivalent to Figure 4(a), where all 'predictors' are exogenous. However, this is an unrealistic scenario, and in most real-world cases, the predictors will not be exogenous.

Multiple Regression - Misspecified for Realistic Structure

In the previous section we showed how a simple multiple regression can be used to recover meaningful, causal parameter estimates, so long as the true causal structure of the data corresponds with the implicit causal structure implied by the multiple regression. However, the implicit causal structure of a linear regression is extremely restrictive and, when modeling real-world data, it is likely to be misspecified. In this section we demonstrate what happens when such misspecification occurs.

Let us see what happens when we follow the same procedure to try to estimate some parameters for another simple data generating process which follows the example in Figure 4(b). We assume the following data generating structural equations (adapted from Peters et al. 2017):

$$
\begin{aligned}
\mathbf{x}_4 := \mathbf{U}_4, \quad \mathbf{x}_2 := 0.8\mathbf{U}_2, \quad \mathbf{x}_0 := \mathbf{x}_4 - 2\mathbf{x}_2 + 0.2\mathbf{U}_0, \quad \mathbf{x}_1 := -2\mathbf{x}_0 + 0.5\mathbf{U}_1, \\
\mathbf{x}_3 := \mathbf{x}_2 + 0.1\mathbf{U}_3, \quad \mathbf{x}_5 := 3\mathbf{x}_0 + 0.8\mathbf{U}_5, \quad \mathbf{x}_6 := \mathbf{x}_1 + 0.5\mathbf{U}_6, \\
\mathbf{y} := 2\mathbf{x}_3 - \mathbf{x}_1 + 0.2\mathbf{U}_y, \quad \mathbf{x}_7 := 0.5\mathbf{y} + 0.1\mathbf{U}_7
\end{aligned}
\tag{2}
$$

For these equations we have simplified the notation to make things clearer: $\mathbf{U}_k \sim \mathcal{N}(0,1)$. The structural process is still linear and the additive noise is Gaussian, so we do not yet need to worry about utilizing flexible function approximation techniques (such as those discussed in Part 1).

It is worth studying these equations to understand their implications. Note that, for instance, $\mathbf{x}_3$ is only determined by $\mathbf{x}_2$, as well as its own exogeneous noise $\mathbf{U}_3$. This means that, if we perform surgery on these equations by, for example, setting $\mathbf{x}_3$ to a different value or distribution, we have cut off its dependence to its parent. Such graph surgery enables us to explore a range of causal queries such as interventions and counterfactuals, and is formalized by Pearl's *do*-calculus (Pearl, 2009).

Given the simple linear form in Equation 2 for Figure 4(b), it is possible to traverse the paths in the c-DAG and to combine the effects multiplicatively. Such a process should be familiar to those who have studied path diagrams and SEM (Kline, 2005). For instance, the effect of $\mathbf{x}_0$ on $\mathbf{y}$ is the multiplication of the effect of $\mathbf{x}_0 \rightarrow \mathbf{x}_1$ with the effect of $\mathbf{x}_1 \rightarrow \mathbf{y}$. Together, we have the mediated path: $\mathbf{x}_0 \rightarrow \mathbf{x}_1 \rightarrow \mathbf{y}$. According to Equation 2 and Figure 4, the effect of $\mathbf{x}_0$ on $\mathbf{y}$ therefore corresponds with $-2 \times -1 = 2$. In this case, $\mathbf{x}_1$ is *mediating* the effect of $\mathbf{x}_0$ on $\mathbf{y}$. Readers may already be aware of the issues relating to the inclusion of mediators in a regression analysis (see e.g., Cinelli et al. 2020; Rohrer 2018; Pearl 2009), and this is trivially demonstrated by comparing the regressions of $\mathbf{y}$ onto $\mathbf{x}_0$ whilst (a) adjusting for $\mathbf{x}_1$ and (b) and not adjusting for $\mathbf{x}_1$. Here, adjusting for a variable is equivalent to *controlling* for it, but the adjustment terminology is more appropriate for structural scenarios (Pearl, 2009). First, the data are simulated according to Eq. 2, with $N = 5000$. The bivariate correlations and p-values for each of these variables are shown in Table 3.

Table 3: Bivariate Pearson correlations and p-values for the DAG in Figure 4(b).

$r(p)$	$\mathbf{x}_0$	$\mathbf{x}_1$	$\mathbf{x}_2$	$\mathbf{x}_3$	$\mathbf{x}_4$	$\mathbf{x}_5$	$\mathbf{x}_6$	$\mathbf{x}_7$
$\mathbf{y}$	.92(.00)	-.92(.00)	-.58(.00)	-.56(.00)	.76(.00)	.91(.00)	-.93(.00)	1.00(.00)

The results in Table 3 demonstrate a strong and statistically significant bivariate correlation between each predictor and the outcome. Now, when only using $\mathbf{x}_0$ as a predictor, we estimate the effect of $\mathbf{x}_0$ on $\mathbf{y}$ to be $\hat{\theta}_0 = 1.28$. Recall that the true effect of $\mathbf{x}_0$ on $\mathbf{y}$ is 2. In spite of the large sample size, the output estimate is highly biased and does not seem to correspond with any of the parameters in the original simulation. Indeed, regardless of how large the sample size is, this coefficient estimate will converge to a value that is far from the true estimand. This is because the structure of the data generating process was not considered: We simply applied a linear regression to the data without accounting for the fact that the implicit structure of a linear regression does not match the structure in the data. In this situation, the multiple regression model might still have some limited utility as a purely *predictive* function, but its parameters should not be interpreted as anything relevant to the causal structure of the phenomenon of interest because it is *misspecified.*

When confronted with the dilemma of multiple observed variables, typical practice in psychology and social science might involve using the forward or backward method for variable inclusion (Field, 2009). Besides the problems associated with such practice (i.e., potential overfitting, as described in Part 1), including variables according to some predictive/associational heuristic is likely to result in misspecification. Another approach might be to simply include all variables in the model. Indeed, all the $\mathbf{x}_k$ variables are highly and statistically significantly correlated with the outcome $\mathbf{y}$, so if we were not already aware of the implicit causal structure of linear regression, this might seem like a sensible thing to do. When we include all variables in the model, this results in $\hat{\theta} = -0.01$. Recall again that the true effect of $\mathbf{x}_0$ on $\mathbf{y}$ is 2. The estimate of -0.01 is highly biased. This is because including all the variables in the model imposes the structure shown in Figure 4(a), where all variables are exogenous.

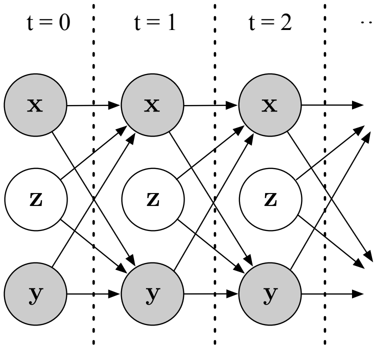
Including $\mathbf{x}_0$ and the mediating variable $\mathbf{x}_1$ confirms that including mediating variables is problematic: The regression including both $\mathbf{x}_0$ and $\mathbf{x}_1$ yields $\hat{\theta} = -.94$. As expected, the effect of $\mathbf{x}_0$ on the outcome is highly biased, and of the opposite sign (i.e., negative rather than positive) to the true causal effect. It should now be clear that the use of what might be called naive multiple regression cannot yield meaningfully interpretable parameters unless the model corresponds with Figure 4(a), and this is highly unlikely. Indeed, it is arguable as to whether the interpretation of this parameter (and even its direction) is of any scientific value at all. Utilizing hierarchical or Bayesian approaches will not help in cases where the structure of the model is misspecified.

Addressing Misspecification Using Causal Inference Techniques

We have seen that using naive multiple regression is inadequate when trying to estimate a causal effect from data with a non-trivial structure, even when the underlying functional form of the relationships is linear. Even where the structure is of relatively low complexity, the resulting coefficient estimates can be wildly biased. This illustrates that, regardless of whether the functional form matches the true functional form of the data (and in the linear simulations above, it did), it is impossible to recover meaningful effect size estimations with a misspecified model. In order to recover an unbiased estimate of the true effect, we need to understand techniques from the field of causal inference.

Structural Equation Modelling (SEM) is reported to be one of the most common methods used in psychology and social science (Blanca et al., 2018), and enables unbiased estimation of the parameters, so long as the structure of the SEM model matches or at least subsumes the structure of the data generating process, and so long as a number of restrictive assumptions are met (Peters et al., 2017). These assumptions apply to causal inference in general. The subsumption point relates to the fact that researchers, when faced with uncertainty about the structure of the data generating process, should choose to expand their model class rather than restrict it. In other words, researchers should, in general, choose to include an extra arrow in their SEM rather than remove one. The choice to expand the model allows for the possibility of a particuarl cause and effect relationship in the data, whereas a removal of a causal link enforces an absence of dependency and thereby represents a strong model restriction that needs to be well justified before its imposition.

Figure 5: Example Directed Acyclic Graph for Time Series



Note. c-DAG for a time series setting, highlighting the complexity associated with identifying a particular causal effect, especially when there may be unobserved confounding (Peters et al., 2017).

In practice, we rarely have access to the true model when we create an SEM (D'Amour, 2019; Wang & Blei, 2019; Tenenbaum & Griffiths, 2002). Indeed, as the SEM grows in complexity and/or its causal constraints, the chance of it becoming misspecified increases. If certain assumptions are made, and we reduce our goal to the estimation of a specific and restricted set of effects (e.g., just the effect of $\mathbf{x}_0$ on $\mathbf{y}$), it may be sufficient to leverage domain knowledge and causal inference techniques to acquire a reliable estimate without having to correctly specify the full graph. Such techniques have been extensively covered elsewhere (Peters et al., 2017; Pearl, 2009; Imbens & Rubin, 2015; Pearl et al., 2016; Angrist & Krueger, 2001) and include the use of instrumental variables, propensity score matching, and regression discontinuity designs (Blossfeld, 2009), but we briefly cover one particular technique known as *backdoor adjustment* below (Pearl, 2009).

Backdoor adjustment involves identifying what are known as *backdoor paths*. An example of a backdoor path between $\mathbf{x}_0$ and $\mathbf{y}$ in Figure 4(b) is $\mathbf{x}_0 \leftarrow \mathbf{x}_2 \rightarrow \mathbf{x}_3 \rightarrow \mathbf{y}$. $\mathbf{x}_2$ and $\mathbf{x}_3$ are therefore part of what is known as the backdoor adjustment set; a set of variables which, if adjusted for, block the backdoor path. We can adjust for all the backdoor variables, or the minimal set sufficient to block the path (in our case, either $\mathbf{x}_2$ or $\mathbf{x}_3$ will do). Including $\mathbf{x}_0$ and $\mathbf{x}_3$ yields $\hat{\theta} = 2.00$.

We have now recovered an unbiased estimate of the effect of $\mathbf{x}_0$ on $\mathbf{y}$ (which was approximately equal to two), and we only needed to regress $\mathbf{y}$ onto two variables, despite our world knowledge dictating that at least eight were involved in the data generating processes as a whole (indeed, all variables in this simulation are highly and significantly correlated with the outcome). If we are also interested in the mediation through $\mathbf{x}_1$ then we can undertake separate regressions to break the problem down. The estimated parameters are then meaningfully interpretable insofar as they correspond with the parameters in the true data generating process. In other words, if $\theta = 2$, then every unit increase in $\mathbf{x}_0$ results in two units increase in $\mathbf{y}$.

Does Time Help?

Researchers may believe that the inductive bias imposed with the directionality of time is helpful in identifying the causal effect and correctly specifying a causal model. Indeed, the fact that time cannot flow backwards does constrain the possible directions of our arrows in our c-DAG, and therefore reduces the complexity of a time series model. However, in spite of the fact that a time series model may be the only way to answer a certain causal question, time series models may be far more complex than cross-sectional models, owing to the introduction of the additional time dimension. Therefore, certain causal questions may only be answerable by considering time, but the causal effect of interest may be considerably harder to identify as a result. Figure 5 depicts a simple scenario with two variables, $\mathbf{x}$ and $\mathbf{y}$, and a hidden confounder $\mathbf{z}$. Each variable influences its own future as well as the future of the other variable. In the presence of the unobserved confounder the causal effect between $\mathbf{x}$ and $\mathbf{y}$ (however this might be defined) is *unidentifiable*. The complexity of this graph could grow further still if we include causal arrows between $\mathbf{x}$ and $\mathbf{y}$ (and potentially $\mathbf{z}$) for the same time point (i.e., $\mathbf{x}$ and time one influences $\mathbf{y}$ at time one), or if we add any additional (un)observed variables. In spite of the restriction that the arrows cannot flow backwards, this structure therefore has the potential to be immensely troublesome from the point of view of identifiability. Indeed, the use of causal inference with time series phenomena is a very current and ongoing research topic in the fields of causal inference and machine learning (Peters et al., 2017; Krishnan et al., 2017; Lohmann et al., 2012). Interested readers are pointed to an accessible introduction of the topic, and its use in psychology, by Gische et al. (2020).

Challenges, Assumptions, and Limitations of Causal Modeling

It is worth emphasizing that, with only naive multiple linear regression models, we were unable to acquire a meaningful effect size estimate for non-trivial data generating process. Indeed, we used a relatively simplistic synthetic simulation to demonstrate that multiple linear regression yields meaningless estimates, but in real-world applications the graph may actually be significantly more complex which makes it extremely challenging to correctly specify the structure of the c-DAG, and therefore to use techniques such as backdoor adjustment.

More generally, it is extremely difficult to obtain reliable effect size estimates from observational data concerning complex real-world social phenomena using these techniques. Indeed, the infamous 'crud' factor, which describes the fact that "everything [in social science] correlates to some extent with everything else" makes causal inference in social science and psychology particularly challenging (Meehl, 1990; Orben & Lakens, 2020).[5] One challenge relates to the identification of suitable backdoor adjustment variables, as well as other structural entities such as colliders, mediators, instrumental variables, proxy variables etc. in order to facilitate the *identification* of the causal effect using the observed data (for techniques, see e.g., Cinelli et al. 2020; Rubin 2005; Imbens & Rubin 2015; Angrist & Krueger 2001; Pearl 2009; Wang & Blei 2019; D'Amour 2019). Another challenge relates to the fact that social scientists are often concerned with the study of complex social systems with dynamic interdependencies. Such systems may not exhibit readily identifiable cause and effect pairs (Blossfeld, 2009).

In the same way that we chose to identify a *single* causal effect using the backdoor adjustment method, it may be beneficial for researchers to attempt to simplify their causal research questions. For example, in contrast with the typical use of SEM in psychology and social science (where the researcher attempts to derive multiple effect estimates simultaneously), targeted learning adopts the philosophy by 'targeting' a specific causal effect of interest, and orienting the analysis around its estimation using machine learning to reduce misspecification (van der Laan & Rose, 2011). The 'no free lunch theorem' familiar to machine learners applies here: causal inference yields the most information, but it is not easy (Wolpert & Macready, 1997). Attempting to undertake inference across

[5] The crud factor also results in an abundance of meaningless statistical significance, owing to the fact that null-effects are practically non-existent in social phenomena (Meehl, 1990).

multivariate, complex, linear SEM graphs is therefore extremely ambitious in light of its limited functional form and likely misspecification, and is highly unlikely to yield meaningful estimates. That said, exploratory work can still be highly valuable (Shrout & Rodgers, 2018). Part of the development process for SEMs (or, more generally, the underlying theory about the phenomenon) could involve causal directionality tests and validation via causal discovery techniques from machine learning (Peters et al., 2017; Scholkopf, 2019). Such techniques, at least in restricted circumstances, may be able to test the directionality of the causal effects (Goudet et al., 2019; Mooij et al., 2010), identify backdoor adjustment set variables (Gultchin et al., 2020), estimate the magnitude of causal effects using flexible function approximation techniques (Yoon et al., 2018; Shi et al., 2019), or infer hidden confounders from proxy variables using variational inference (Louizos et al., 2017a; Vowels et al., 2020). We recommend both Targeted Learning (van der Laan & Rose, 2011) as well as deep latent variable neural network models (Louizos et al., 2017a; Vowels et al., 2020) as possible approaches to the significant problem of causal effect size estimation, although many others exist (Gultchin et al., 2020; Shalit et al., 2017; Shi et al., 2019; Zhang et al., 2020; Yao et al., 2018).

Even once a researcher believes that they have accounted for the difficulties described above, and have simplified their research question or hypothesis, their consequent estimations then rest on the assumption known as *ignorability*; that there are no further latent/unobserved factors that have yet to be accounted for. Figure 4(c) depicts the presence of an unobserved confounder $\mathbf{z}$. Particularly in cases where researchers are dealing with observational (as opposed to experimental) data, the assumption of ignorability may be strong, untestable, and unrealistic. Other assumptions may also be relevant, depending on the causal question being asked, such as the stable unit treatment value assumption and the positivity assumption for estimating treatment effects. It is important researchers familiarize themselves with all relevant assumptions and limitations before undertaking causal inference, and make them explicit in their work (e.g., when they use SEM) (Grosz et al., 2020).

Finally, the simulations here assumed linear and additive structural equations of the form: $\mathbf{x}_1 := \theta_0\mathbf{x}_0 + \mathbf{U}_1$. However, and as discussed earlier, c-DAGs are general and do not restrict the functional forms relating the variables. Indeed, in real-world scenarios the assumption of linearity may impair the capacity of the model to estimate unbiased coefficients, in much the same way as it limited predictive models (Coyle et al., 2020; van der Laan & Rose, 2011; van der Laan & Starmans, 2014; van der Laan & Rose, 2018). The difficulties of effect estimation are therefore compounded by the difficulties associated with identifying an appropriate functional form for the dependencies between variables (i.e., identifying what Blossfeld 2009 calls "effect shapes"). Unless the structure of the model *and* its functional form sufficiently match those of the true data generating process, *and* we have an identifiable causal effect, the model may be misspecified and uninterpretable.

Part 2 Summary

We described how difficult it is to obtain reliable causal effect size estimates, and we have also demonstrated how a failure to consider the causal structure may yield biased, meaningless effect sizes, regardless of whether the researcher adopts a predictive or causal approach. We provided one example of a causal inference technique known as backdoor adjustment, as a way to identify the causal effect of interest. Doing so enabled us to simplify the analytical problem from one of estimating all path coefficients in a complex graph, to one of estimating a specific effect by identifying variables from an adjustment set. In practice, identifying these backdoor variables represents a significant challenge, because it requires sufficient causal knowledge. Causal inference rests on a number of strong assumptions, perhaps the strongest of all being that of ignorability: That there are no unobserved confounders. Finally, researchers must also consider the functional form used to represent the causal dependencies between the variables. As such, problems with identifiability, ignorability, misspecification due to incorrect structure, and misspecification due to limited functional form have the potential to compound each other. In summary, it is important that researchers recognize the significant difficulties associated with estimating meaningful causal effects with observational data.[6]

[6] Given the complexity associated with avoiding misspecification, on top of considering functional form, readers may come to the conclusion that causal inference should be reserved for Randomized Controlled Trial (RCT) and experimental contexts. The common view is that RCTs represent the "gold standard" of research. However, a growing literature highlights the limitations of RCTs, and how observational studies may, at least in certain circumstances, represent a promising alternative, particularly in terms of lower cost, reduced ethical implications, and larger sample size (Frieden, 2017; Deaton & Cartwright, 2018; Bothwell et al., 2016; Jones

PART 3: UNRELIABLE INTERPRETATIONS

In this part, we introduce explainability and interpretability, and describe how misspecified models with limited functional form may be neither explainable, nor interpretable. When the complexity of a model is increased to mitigate the issue of limited functional form it may be nonetheless explainable in spite of possible misspecification due to incorrect structure. We discuss a range of problems relating to conflated and unreliable interpretations in psychology and social science. In our view, the conflation arises not just as a result of the alleged taboo against causal inference (Grosz et al., 2020), but also due to an apparent lack of understanding concerning the limitations associated with the interpretability of misspecified models with limited functional form and/or incorrect causal structure.

EXPLAINABILITY AND INTERPRETABILITY

Scrutinizing the parameters of a model in a predictive sense is referred to as *explaining*, in that we are explaining the behavior of the model, rather than *interpreting* the model's parameters in relation to some external real-world causal phenomenon (Rudin, 2019). We therefore distinguish *interpretability* from *explainability*. In this paper we use the term interpretation to describe the process of using a model to understand something about the structure in the data or phenomenon, and the term is therefore of particular relevance to causal approaches. As we will show, linear models are not immune to problems affecting interpretability both for reasons of limited functional form as well as misspecification (see Parts 1 and 2). Explainability, on the other hand, refers to the capacity to explain why a model makes a certain prediction or classification, based on its functional form or algorithmic rules (Rudin, 2019), and is therefore a term particularly relevant to predictive approaches. As the complexity of a model's functional form increases, it becomes increasingly difficult to either interpret or explain a model (Rudin, 2019).

THE (UN)INTERPRETABILITY OF LINEAR MODELS

Linear models are deceptively simple to *explain* because their model coefficients seem to provide a direct means to understand why the model made a certain prediction. If the model is not misspecified (i.e., it has adequate functional form and causal structure), then this parameter may be interpreted in a causal sense as well as in a predictive/explainable sense. In other words, the parameter not only tells us something about how the model's output changes with respect to a change in its input, but also something about the external phenomenon being modeled. However, if the model is misspecified due to incorrect structure, then the parameter may only be used to explain the behavior of the model, and will not correspond meaningfully with some external causal quantity.

Perhaps surprisingly, if the model is misspecified *both* in terms of its functional form and its structure, then the model may be neither interpretable nor explainable. In this scenario, complex cancellation effects may render the coefficients of linear models meaningless (Lundberg et al., 2020; Breiman, 2001a; Haufe et al., 2014). Just because a predictive model (e.g., multiple linear regression) indicates that variable $\mathbf{x}_1$ has statistically significant association with an outcome, does not imply that it is meaningful to interpret this coefficient either in terms of a specific quantified value, or in terms of an ordinal level of variable importance. The problems are caused both by the function's inability to account for non-linear relationships and by the mismatch of the function's implicit structural (i.e., causal) form with the true form of the data. We demonstrated the latter issue in Part 2. For the former issue, we generate a synthetic example, closely following that of Lundberg et al. (2020).[7] Essentially, the relationship between the outcome and two particular features in a semi-synthetic dataset is modified to include an increasing amount of non-linearity following the relationships in Equation 3.

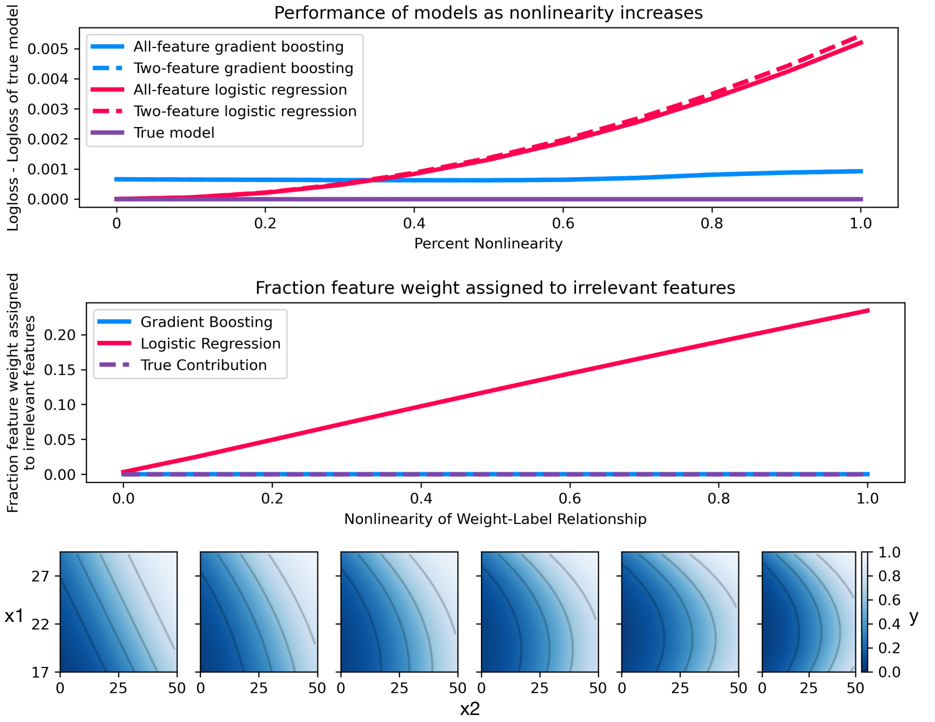
& Podolsky, 2015). Furthermore, in a social science context, randomized experiments may be practically infeasible and potentially unethical (Blossfeld, 2009). To clarify, we do not wish to engage in a debate about the merits and pitfalls associated with undertaking causal inference on experimental versus observational data, but we do note that the perception of RCTs as representing a gold standard is potentially limiting and scientifically unhelpful.

[7] Full code for the original example can be found here: `https://github.com/suinleelab/treeexplainer-study/`.

$$\mathbf{y} = \sigma((1-q)(0.388\mathbf{x}_1 - 0.325) + q(1.714\mathbf{x}_1^2 - 1) + 1.265\mathbf{x}_2 + 0.0233) \qquad (3)$$

Here, σ is the logistic link function, q is the degree of non-linearity, which is varied between zero (describing a linear relationship) and one (describing a model with a quadratic relationship), $\mathbf{y}$ is the outcome, and $\mathbf{x}_1$ and $\mathbf{x}_2$ are the two predictor variables. The choice of the factors (e.g., 0.388) and intercepts (e.g., -0.325) are arbitrary, and derive from the classic NHANES I dataset (Launer, 1994; Fang & Alderman, 2000) from which the predictors and outcome are drawn. The relationship between the predictors and the outcome as q is increased from zero to one is shown in the lowest plot of Figure 6. Two models were fit to these synthetic data: a linear logistic regressor, and a machine learning algorithm known as *XGBoost* (Chen & Guestrin, 2016). The upper plot in Figure 6 shows how the logistic regressor's error increases as the non-linearity increases. In contrast, the XGBoost model's prediction error remains low. Notably, when q is close to zero (i.e., the percent non-linearity is low), the linear model outperforms the XGBoost model, and has the potential to directly match the data generating process. The middle plot shows how the contribution of irrelevant features to the outcome changes as the non-linearity increases. For the XGBoost model, any irrelevant features are ignored regardless of the degree of non-linearity, and their weights remain at zero (which is in line with the true model). On the other hand, the linear model assigns weight (i.e., the coefficients of the model change) to irrelevant features as the non-linearity increases. This is highly problematic for explainability and interpretability - it results in irrelevant features being indicated to be of predictive importance even when they are not.

Figure 6: The Uninterpretability of Linear Models in the Presence of Non-Linearity



Note. Demonstrates how the predictive performance of a logistic regressor drops as non-linearity increases, whereas the XGBoost (Chen & Guestrin, 2016) model does not (top); shows how irrelevant feature attribution increases with non-linearity for the linear regressor, but for XGBoost it does not (middle); the relationship between variables in the dataset for these experiments becomes increasingly non-linear. These experiments were close adaptations of those by Lundberg et al. (2020).

The (Un)Interpretability of Models with Complex Functional Form - Camels in the Countryside

In Part 1 we suggested that researchers explore machine learning methods which facilitate the modeling of complex, non-linear relationships between variables. These techniques are applicable to predictive as well as causal approaches. In spite of their flexible functional form, powerful predictive approaches are explainable but not necessarily interpretable. We now describe a famous example which highlights how using powerful function approximation circumvents limitations in functional form does not yield interpretable models. This is one of the principal limitations of purely predictive approaches and closely relates to misspecification (see Part 2). The example involves the classification of images of cows and camels, where images of cows frequently feature countryside backgrounds and images of camels tend to feature sandy or desert regions (Arjovsky et al., 2020). A predictive function will not respect the orthogonality and semantics of the animal or background, and the background provides a convenient cue, albeit one which is irrelevant and *confounding*, with which to classify the animal. Hence, a cow in a desert may be wrongly classified as a camel, and a camel with a countryside background may be wrongly classified as a cow. This issue may never become problematic in practice, so long as the function is not exposed to a new distribution of images, where the joint distribution of backgrounds and animals changes. This highlights how predictive models, owing to their misspecification, are sensitive to what is known as covariate or distributional shift. Given a change in the number of photographs of cows in desert regions, or camels in the countryside, the performance of the classifier may suffer considerably.

This example concerning issues relating to classification of high-dimensional image data may appear somewhat unrelated to the typical data that psychologists are concerned with, but actually the problem of confounding is just as important in the low-dimensional setting (Cinelli et al., 2020; Rohrer, 2018). Indeed, predictive models are usually fit by minimizing an error criterion (e.g., mean squared error or binary cross entropy), and there is therefore nothing to restrict these models from leveraging any or all statistical correlations present in the data. The use of predictive model explainability techniques (discussed in more detail below) can be used to help identify whether the model might be leveraging factors which have the potential to be confounding, and can provide considerable insight. Unfortunately, if the confounders are latent/unobserved, then it may be very difficult to identify and avoid such problems. Consequentially, predictive models are rarely interpretable.

Limited Functional Form and Misspecification Results in Conflated and Unreliable Interpretations

The examples above highlighted that when the functional form of a model is limited in its capacity to model the relationships between variables, the model coefficients become meaningless and the model is unexplainable. A further problem arises when the model is misspecified for structural reasons. The issues associated with limited functional form and causal misspecification therefore compound to yield model coefficients that are (doubly) uninterpretable. Treating them otherwise would be to interpret these coefficients as being causally meaningful, and this is an example of conflated and unreliable interpretation. If the functional form of the model were correct (i.e., both the model as well as the relationships between variables were linear), then a linear model would be explainable, but not interpretable. This is because the outcome predicted by the model would indeed be changing according to a β_k change in the input variable $\mathbf{x}_k$, but owing to misspecification, this β_k would still not correspond with any causal quantity. As such, it is only when linear models are neither misspecified due to limited functional form (compared with the true relationship in the data) nor structurally misspecified, that they are interpretable.

Explainability Techniques

The ability to interrogate and explain our predictive models is important, particularly given that the deployment of such models for automated decision making processes have the potential to seriously impact individuals' lives (Hardt et al., 2016; Kilbertus et al., 2017; Locatello et al., 2019; Cao & Daume III, 2019; Liu et al., 2019; Howard & Borenstein, 2018; Rose, 2010; Louizos et al., 2017b; Moyer et al., 2018; Buolamwini & Gebru, 2018). Indeed, the European Union has recently decreed that the use of machine learning algorithms (which includes the use of predictive functions) be undertaken in such a way that any individual affected by an automated decision has the right to an explanation regarding that decision (Aas et al., 2019; European Union, 2016). In the previous section we described the camels in the countryside problem, whereby powerful predictive models with flexible functional form do not respect causal structure in the data. However, complex models (often called *black box* models) are more difficult to explain than linear models, and we therefore need explainability techniques to do the explaining for us.

Model explainability is a burgeoning area of machine learning, in which commendable strides have been made in recent years (e.g., Alaa & van der Schaar 2019; Wachter et al. 2018; Lundberg et al. 2020). The techniques facilitate a form of *meta-modeling*, whereby a simpler, human-interpretable and thereby explainable model is used to represent the more complex, underlying model (Rudin, 2019). One popular explainability technique derives from a game theoretic approach to quantifying the contribution of multiple players in a collaborative game; namely, Shapley values (Shapley, 1953). Recently, Shapley values have been adapted to yield meaningful explanations of models that correspond well with human intuition (Lundberg & Lee, 2017; Lundberg et al., 2017; 2020; Sundararajan & Najmi, 2020; Chen et al., 2020). Indeed, these methods were used with XGBoost in the experiments demonstrating the problems with linear model interpretability above (Figure 6). The family of Shapley methods provide breakdowns which indicate how much each input variable or feature contributes to a model's prediction for any individual datapoint. Such individualized prediction and explainability is particularly important for individualized treatment assignments (for example), and thereby mitigates concerns regarding the use of aggregation in psychology and social science (Bolger et al., 2019; Fisher et al., 2018). The methods can be used equally for complex functions (such as neural networks) as well as for simple linear functions. By combining powerful function approximation with explainability techniques, we may be able to achieve accurate forecasts and outcome predictions, while maintaining the capacity to understand what our model is actually doing when it makes a prediction.

From a research standpoint, explainability techniques allow researchers to understand, in a purely associational sense, which variables and interactions between variables are important when making a prediction. For example, if one identifies that a variable, previously considered to be important, contributes negligible predictive value then one might investigate whether this variable does or does not fit into a particular theoretical framework. We would therefore argue that researchers should consider a combination of predictive methods with explainability tools as a useful means to contribute new knowledge, particularly during the early and/or exploratory stages of investigation. It is, however, worth emphasizing that just because a predictive model finds a particular feature (ir-)relevant to making a prediction, does not mean that this association is meaningful outside of the function/model (as with camels in the countryside). Furthermore, an explainability technique represents a form of model in its own right, and the process of modeling a model brings its own difficulties (see e.g., Rudin 2019; Kumar et al. 2020). Indeed, if the explanation model is good at explaining the data in a simple, human-readable form, then the explanation model provides evidence that a simpler, more explainable model was possible to begin with. These difficulties notwithstanding, the explainability techniques provide a valuable means to leverage predictive model for exploratory research.

Part 3 Summary

In Part 3, we have described how either limited functional form, or model misspecification, or both, result in uninterpretable models. In such cases, any attempt to interpret the models in spite of these limitations results in conflation and unreliability. The interpretations are conflated because a misspecified model cannot be interpreted causally, and they are unreliable because predictive models can only be explained. This distinction is important because, if a misspecfication has occurred (perhaps because we intentionally adopted a predictive/non-causal approach), one can restrict the purview of scientific conclusions to the specific mathematics of the algorithm used for prediction. In other words, powerful function approximation techniques may be able to accurately predict outcomes and have the flexibility to match the functional form of the true data distribution, but they do not necessarily respect or reflect the *causal* structure in the data generating process. Does this mean that predictive techniques cannot generate understanding? Not entirely. There are many scenarios, particularly during the exploratory stages of a research project, for which researchers may not yet have a strong, empirically supported inductive bias or theory about the data generating process. Rather than testing specific theoretical hypotheses during these early stages, it may be pertinent to ask more general research questions. The goal may then be to amass varied evidence (e.g., by using predictive models) to gradually uncover a basis for the development of an increasingly refined theory

(Gelman, 2014; Shrout & Rodgers, 2018; Oberauer & Lewandowsky, 2019; Tong, 2019). Of course, researchers should be transparent about whether this is their goal, and carefully consider how they interpret (or indeed explain) predictive models. Model explainability techniques may be useful in building up an intuition about 'what is important' in the phenomenon of interest. However, these techniques are not without their own limitations, and we urge researchers to engage broadly with experts in the practice of these techniques to ensure that (a) their approaches are optimal for their research, and (b) that their interpretations (or explanations) are tempered according to the limitations of their models.

Recommendations and Conclusion

The replicability crisis has drawn attention to numerous weaknesses in typical psychology and social science research practice. However, in our view, issues relating to limited functional form, model misspecification, and unreliable interpretations have not been sufficiently addressed in prior work. Indeed, while it is difficult, if not impossible, to quantitatively apportion the crisis according to its myriad causes, in our view the issues covered in this work represent significant contributing factors.

Our view is that, in general, researchers in psychology and social science lack some competence in the practice of prediction and causal inference. If researchers were more competent at prediction, they would avoid interpreting linear model parameters using implicit causal language (Grosz et al., 2020), avoid using naive linear models to test causal hypotheses derived from causal theories, and instead be using varied and flexible function approximation techniques, model explainability tools, and train/test data splitting and/or cross-validation techniques (Yarkoni & Westfall, 2017). On the other hand, if researchers were more competent in causal inference, they would be less ambitious about specifying and interpreting large (causal) SEM graphs (which are almost invariably accepted as valid *a priori*; VanderWeele 2020; Ropovik 2015), more restrained when it comes to interpreting the coefficients of misspecified models, more transparent about assumptions when defining explanatory models (Grosz et al., 2020), use more explicitly causal language and terminology (Grosz et al., 2020), more clearly distill and identify the specifics of their causal questions or hypotheses, and be less likely to worsen the bias and generalizability of their inferences by adopting *ad hoc*, data driven variable model manipulation techniques during the analysis stage. Finally, if researchers had a clearer understanding about the differences between predictive and causal approaches, then we would also see more delineation between the two. Typical practice therefore involves a combination of unreliable interpretations regarding models with limited functional form and causal misspecification.

1. We recommend that psychologists and social scientists give more consideration to predictive approaches, particularly during the exploratory stages of a research project.

The inherent complexity and non-linearity of the typical phenomena of interest to psychologists and social scientists may make the goal of causal inference arbitrarily complex (Meehl, 1990). This may partly explain why researchers in psychology and social science are generally discouraged from drawing causal conclusions from observational data, despite them doing so implicitly anyway (Grosz et al., 2020; Dowd, 2011). Indeed, the use of SEM could be taken as evidence of an explicit intention to undertake causal research, as the very structure of the model is an imposition of the researcher's view on the data generating process. The use of an explicit causal graph with opaque predictive interpretations represents a further example of the conflation of predictive and causal approaches. In cases where the models themselves are misspecified both in terms of linear functional form and untestable structural assumptions, the interpretation of such models becomes unreliable.

When researchers wish to model the relationships between variables, either as part of a causal model, or for purposes of prediction, then it may be extremely advantageous for them to consider techniques common in machine learning, particularly in combination with model explainability techniques. Indeed, Yarkoni & Westfall (2017) have previously made a similar recommendation. Powerful function approximation techniques including feature engineering or data-adaptive techniques such as neural networks or random forests, can be used to leverage as many associations present in the data sample as possible. In the case of predictive modeling, a consideration for the causal structure of the data is possible but not necessary. Incorporating causal inductive bias may aid in generalization, but it is not strictly necessary to achieve good predictive performance. Unfortunately, the use of techniques with potentially data-adaptive, flexible functional form is extremely rare in psychology and social science, where the use of models with restrictive linear functional form is ubiquitous (Yarkoni & Westfall, 2017; Blanca et al., 2018).

2. We recommend that psychologists and social scientists seek collaboration with statisticians and machine learning engineers/researchers, whose principal focus is to understand, practice, and develop function approximation and causal inference techniques. Given that there exist entire fields dedicated to the study of relevant modeling approaches (e.g., statistics, machine learning, causal inference), independently of the empirical human sciences, it is perhaps unrealistic to expect an expert in psychology or social science, to have equal expertise in the practice of predictive and explanatory modeling, particularly when the mathematical knowledge required to understand these techniques is both significant and rare in these fields (Boker & Wenger, 2007). Furthermore, new methods are continually developed and updated in the fields of statistics and machine learning. As well as encouraging researchers to make themselves more familiar with the topics of predictive and causal modeling, we also recommend they seek collaboration with experts in the practice of their chosen analytical approach. Note that this recommendation has been made by researchers previously in various contexts (e.g., Lakens et al. 2016).

3. We recommend researchers be transparent about whether they are adopting a predictive or causal approach and to qualify their interpretations. We have discussed how unreliable interpretations may stem from issues of limited functional form and causal misspecification, and how these issues may be common in the fields of psychology and social science. We encourage researchers to ask themselves what an interpretation of an effect size or parameter derived using a naive (i.e., misspecified) model actually means: Is it actually an explanation for how much the output of the *model* changes with respect to a change in the input; or is it being interpreted causally (e.g., this childhood intervention increased well-being by θ-amount)? In either case, researchers need to be transparent and clearly articulate whether they are adopting a predictive or causal approach. Each approach is associated with assumptions and limitations which need to be clearly stated in order to contextualize any explanations or interpretations which are made. Predictive model explainability tools have their own limitations and may actually contradict the results from undertaking causal inference: While the inclusion of a mediator in a regression can completely block a causal path reducing the estimated effect to zero, a strong effect might be indicated by an explanation of a predictive model. Similarly to Grosz et al. (2020), we therefore recommend that researchers clearly state their approach as well as its associated assumptions and limitations, and moderate their explanations, interpretations, and conclusions accordingly.

4. We recommend that researchers distill their research questions and hypotheses. It may be pertinent for researchers to attempt to distill and simplify causal questions so that they are both minimal and sufficient. For example, in our discussion of causal inference, we chose to identify a single causal effect, and for this it was sufficient to identify the minimal backdoor adjustment set necessary to render this causal effect identifiable. As such, a full graph did not need to be specified, even though it may need to be considered in order to find the backdoor adjustment variables. van der Laan & Rose (2011) recommend a similar “targeted” approach. More generally, by distilling our research questions and hypotheses, we may be able to increase the chance that our modeling attempts are successful, and that we have realistic expectations of the level of understanding that can be achieved. This recommendation therefore overlaps with the recommendation for transparency in so far as distilling a research question or hypothesis will make it easier to be transparent.

While we have focused on the fields of psychology and social science, we feel the highlighted issues are relevant to all empirical human sciences fields. There is little doubt in our minds that the lack of understanding about the assumptions, limitations, and pitfalls associated with predictive and explanatory modeling has contributed to the replicability crisis, and we implore researchers to address these shortfalls, lest they hinder scientific progress. Every research question and hypothesis may present its own unique challenges, and it is only through an awareness and understanding of varied statistical methods for predictive and causal modeling, that researchers will have the tools with which to appropriately address them.

References

A. A. Aarts et al. Estimating the reproducibility of psychological science. *Science*, 349(6251), 2015.

K. Aas, M. Jullum, and A. Loland. Explaining individual predictions when features are dependent: More accurate approximations to Shapley values. *arXiv:1903.10464*, 2019.

C.H. Achen. Measuring representation: perils of the correlation coefficient. *American Journal of Political Science*, 21(4):805–815, 1977.

A.M. Alaa and M. van der Schaar. Demystifying black-box models with symbolic metamodels. *33rd Conference on Neural Information Processing Systems*, 2019.

J.D. Angrist and A.B. Krueger. Instrumental variables and the search for identification: From supply and demand to natural experiments. *Journal of Economic Perspectives*, 15(4):69–85, 2001.

M. Arjovsky, L. Bottou, I. Gulrajani, and D. Lopez-Paz. Invariant risk minimization. *arXiv:1907.02893v3*, 2020.

A.G. Asuero, A. Sayago, and A.G. Gonzalez. The correlation coefficient: An overview. *Critical Reviews in Analytical Chemistry*, 36(1), 2006.

D.H. Baker, G. Vilidaite, F.A. Lygo, A.K. Smith, T.R. Flack, A.D. Gouws, and T.J. Andrews. Power contours: optimising sample size and precision in experimental psychology and human neuroscience. *Psychological Methods*, 2020.

C.G. Begley and L.M. Ellis. Raise standards for preclinical cancer research. *Nature*, 483:531–533, 2012.

M. Belkin, D. Hsu, S. Ma, and S. Mandal. Reconciling modern machine-learning practice and the classical bias-variance trade-off. *PNAS*, 116(32):15849–15854, 2019. doi: 10.1073/pnas.1903070116.

C. M. Bishop. *Pattern Recognition and Machine Learning*. Springer, New York, 2006.

M.J. Blanca, R. Alarcon, and R. Bono. Current practices in data analysis procedures in psychology: what has changed? *Frontiers in Psychology*, 2018.

H.P. Blossfeld. *Causal analysis in population studies*, chapter Causation as a generative process. The elaboration of an idea for the social sciences and an application to an analysis of an interdependent dynamic social system. Spinger Science and Business Media, 2009.

S. M. Boker and M. J. Wenger. *Data analytic techniques for dynamical sytems*. Lawrence Erlbaum Associates, New Jersey, 2007.

N. Bolger, K.S. Zee, M. Rossignac-Milon, and Hassin R.R. Causal processes in psychology are heterogeneous. *Journal of Experimental Psychology General*, 148(4):601–618, 2019.

J. Botella and J.I. Duran. A meta-analytical answer to the crisis of confidence of psychology. *Anales de psicologioa*, 35(2):350–356, 2019.

L.E. Bothwell, J.A. Greene, and S.H. Podolsky. Assessing the gold standard - lessons from the history of RCTs. *N. Engl. J. Med.*, 374:2175–81, 2016.

L. Bottou, J. Peters, Quinonero-Candela J., D. X. Charles, D. M. Chickering, E. Portugaly, D. Ray, P. Simard, and E. Snelson. Counterfactual reasoning and learning systems: the example of computational advertising. *Journal of Machine Learning Research*, 14, 2013.

L. Breiman. Statistical modeling: The two cultures. *Statistical Science*, 16(3):199–215, 2001a.

L. Breiman. Random forests. *Machine Learning*, 45(1):5–32, 2001b.

J. Buolamwini and T. Gebru. Gender Shades: Intersectional accuracy disparities in commercial gender classification. *Proc. of Machine Learning Research*, 81:1–15, 2018.

Y. T. Cao and H. Daume III. Toward gender-inclusive coreference resolution. *arXiv:1910.13913v2*, 2019.

S.A. Cassidy, R. Dimova, B. Giguere, J.R. Spence, and D.J. Stanley. Failing grade: 89 percent of introduction-to-psychology textbooks that define or explain statistical significance do so incorrectly. *Advances in Methods and Practices in Psychological Science*, 2(3), 2019.

H. Chen, J.D. Janizek, S.M. Lundberg, and S-I. Lee. True to the model or true to the data? *arXiv:2006.16234v1*, 2020.

T. Chen and C. Guestrin. XGBoost: A scalable tree bosting system. *KDD Conference for Knowledge Discovery and Data Mining*, 2016.

C. Cinelli, A. Forney, and J. Pearl. A crash course in good and bad controls. *Technical Report R-493*, 2020.

A. Claesen, S.L.B.T. Gomes, F. Tuerkinckx, and W. Vanpaemel. Preregistration: Comparing dream to reality. *PsyArXiv*, 2019. doi: 10.31234/osf.io/d8wex.

D. Colquhoun. An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science*, 1(3), 2014.

D. Colquhoun. The reproducibility of research and the misinterpretation of p-values. *Royal Society Open Science*, 4(12), 2017.

D. Colquhoun. The false positive risk: a proposal concerning what to do about p-values. *The American Statistician*, 73, 2019.

J. Correll, C. Mellinger, G.H. McCelland, and C.M. Judd. Avoid Cohen's 'small', 'medium', and 'large' for power analysis. *Trends in Cognitive Sciences*, 24(3), 2020. doi: 10.1016/j.tics.2019.12.009.

T. M. Cover and J. A. Thomas. *Elements of information theory*. John Wiley and Sons Inc., New York, 2006.

J.R. Coyle, N.S. Hejazi, I. Malenica, and R.V. et al. Phillips. Targeted learning: Robust statistics for reproducible research. *arXiv2006.07333*, 2020.

A. D'Amour. On multi-cause causal inference with unobserved confounding: Counterexamples, impossibility and alternatives. *Proceedings of the 22nd International Conference on Artificial Intelligence and Statistics*, 89, 2019.

A. Deaton and N. Cartwright. Understanding and misundertstanding randomized controlled trials. *Social Science and Medicine*, 210:2–21, 2018.

S. DeDeo. When science is a game. *arXiv:2006.05994v2*, 2020.

B. E. Dowd. Separated at birth: statistcians, social scientists, and causality in health services research. *Health Research and Educational Trust*, 46(2):397–420, 2011.

R. O. Duda, P. E. Hart, and D. G. Stork. *Pattern Classification*. John Wiley and Sons Inc., New York, 2001.

A. F. Ernst and C. J. Albers. Regression assumptions in clinical psychology research practice - a systematic review of common misconceptions. *PeerJ*, 5, 2017.

European Union. Regulation (eu) 2016/679 of the European Parliament and of the Council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/ec (gdpr). *Official Journal of the European Union*, 59, 2016.

J. Fang and M.H. Alderman. Serum uric acid and cardiovascular mortality: the NHANES I epidemiologic follow-up study. *JAMA*, 283(18):2404–2410, 2000.

A. Field. *Discovering statistics using SPSS*. Sage, Los Angeles, 3rd edition, 2009.

A.J. Fisher, J.D. Medaglia, and B.F. Jeronimus. Lack of group-to-individual generalizability is a threat to human subjects research. *PNAS*, 115(27), 2018.

J. Flake and E. Fried. Measurement schmeasurement: questionable measurement practices and how to avoid them. *Advances in Methods and Practices in Psychological Science*, in press.

T.R. Frieden. Evidence for health decision making - beyond randomized, controlled trials. *N. Engl. J. Med.*, 377:465–475, 2017.

S. Gao, G.V. Steeg, and A. Galstyan. Efficient estimation of mutual information for strongly dependent variables. *AISTATS*, 2015.

A. Gelman. Correlation does not even imply correlation. *Statistical Modeling, Causal Inference, and Social Science (BLOG)*, 2014.

A. Gelman and J. Hill. *Data analysis using regression and multilevel/hierarchical models*. Cambridge University Press, Cambridge, UK, 2007.

A. Gelman and E. Loken. The garden of forking paths: why multiple comparisons can be a problem even when there is no 'fishing expedition' or 'p-hacking' and the research hypothesis was posited ahead of time, 2013. URL `http://www.stat.columbia.edu/~gelman/research/unpublished/p_hacking.pdf`.

M.A. Gernsbacher. Three ways to make replication mainstream. *Behav. Brain. Sci.*, 41, 2019. doi: 10.1017/S0140525X1800064X.

G. Gigerenzer. Mindless statistics. *Journal of Socio-Economics*, 33:587–606, 2004.

G. Gigerenzer. Statistical rituals: the replication delusion and how we got there. *Advances in Methods and Practices in Psychological Science*, 1(2):198–218, 2018.

C. Gische, S.G. West, and M.C. Voelkle. Forecasting causal effects of interventions versus predicting future outcomes. *Structural Equation Modeling: A Multidisciplinary Journal*, 2020. doi: 10.1080/10705511.2020.1780598.

C. Glymour. What went wrong? reflections on science by observation and the bell curve. *The University of Chicago Press on behalf of the Philosophy of Science Association*, 1998.

C. Glymour. *The mind's arrows: Bayes nets and graphical causal models in psychology*. MIT Press, 2001.

I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, Cambridge, Massachusetts, 2016.

O. Goudet, D. Kalainathan, M. Sebag, and I. Guyon. *Cause effect pairs in machine learning*, chapter Learning bivariate functional cusal model. Springer, 2019.

M.P. Grosz, J.M. Rohrer, and F. Thoemmes. The taboo against explicit causal inference in nonexperimental psychology. *Perspectives on Psychological Science*, pp. 1–13, 2020.

L. Gultchin, M.J. Kusner, V. Kanade, and R. Silva. Differentiable causal backdoor discovery. *Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics*, 108, 2020.

J. D. Hamilton. *Time Series Analysis*. Princeton University Press, New Jersey, 1994.

M. Hardt, E. Price, and N. Srebro. Equality of opportunity in supervised learning. *arXiv:1610.02413v1*, 2016.

S. Haufe, F. Meinecke, K. Gorgen, S. Dahne, J-D. Haynes, B. Blankertz, and F. Biebmann. On the interpretation of weight vectors of linear models in multivariate neuroimaging. *NeuroImage*, 87 (15):96–110, 2014.

R. Heesen and L.K. Bright. Is peer review a good idea? *The British Journal for the Philosophy of Science*, 2020. doi: 10.1093/bjps/axz029.

R.E. Heman and A.M.S. Slep. The hazards of predicting divorce without crossvalidation. *J Marriage Fam.*, 63(2):473–479, 2001.

M. Hernan. The c-word: scientific euphemisms do not improve causal inference from observational data. *American Journal of Public Health*, 108(5):625–626, 2018a.

M. Hernan. The c-word: the more we discuss it, the less dirty it sounds. *American Journal of Public Health*, 108(5):625–626, 2018b.

K. Hornik, M. Stinchcombe, and H. White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2:359–366, 1989.

A. Howard and J. Borenstein. The ugly truth about ourselves and our robot creations: the problem of bias and social inequity. *Science and engineering ethics*, 24(5):1521–1536, 2018.

G.W. Imbens and D.B. Rubin. *Causal inference for statistics, social, and biomedical sciences. An Introduction.* Cambridge University Press, New York, 2015.

D.S. Jones and S.H. Podolsky. The history and fate of the gold standard. *Lancet*, 385(1502-3), 2015.

K. Jonsson, J. Matas, J. Kittler, and Y.P. Li. Learning support vectors for face verification and recognition. *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition*, 2000.

N.B. Jostmann, D. Lakens, and T.W. Schubert. A short history of the weight-importance effect and a recommendation for pre-testing: commentary on ebersole et al. (2016). *JESP*, 67, 2016. doi: 10.1016/j.jesp.2015.12.001.

P. Kassraian-Fard, C. Matthis, J.H. Balsters, M.H. Maathuis, and N. Wenderoth. Promises, pitfalls, and basic guidelines for applying machine learning classifiers to psychiatric imaging data, with autism as an example. *Frontiers in Psychology*, 7(177), 2016. doi: 10.3389/fpsyt.2016.00177.

N. Kilbertus, M. Rojas-Carulla, G. Parascandolo, M. Hardt, D. Janzing, and B. Schölkopf. Avoiding discrimination through causal reasoning. *31st Conference on Neural Information Processing Systems*, 2017.

G. King. How not to lie with statistics: Avoiding common mistakes in quantitative political science. *American Journal of Political Science*, 30(3):666–687, 1986.

J.B. Kinney and G.S. Atwal. Equitability, mutual information, and the maximal information coefficient. *PNAS*, 111(9):3354–3359, 2014.

R.B. Kline. *Principles and practice of structural equation modeling*. Guilford Press, 2005.

D. Koller and N. Friedman. *Probabilistic Graphical Models: Principles and Techniques*. MIT Press, Cambridge, Massachusetts, 2009.

A. Kraskov, H. Stogbauer, and P. Grassberger. Estimating mutual information. *Physical Review E*, 69, 2004.

N. Kreif and K. DiazOrdaz. Machine learning in policy evaluation: new tools for causal inference. *arXiv:1903.00402v1*, 2019.

N. Kriegeskorte, W.K. Simmons, P.S. Bellgowan, and C.I. Baker. Circular analysis in systems neuroscience: the danges of double dipping. *Nat. Neurosci*, 12, 2009. doi: 10.1038/nn.2303.

R. G. Krishnan, U. Shalit, and D. Sontag. Structured inference networks for nonlinear state space models. *Association for the Advancement of Artificial Intelligence*, 2017.

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. *Proceeding NIPS Proceedings of the 25th International Conference on Neural Information Processing Systems*, pp. 1097–1105, 2012.

E.I. Kumar, S. Venkatasubramanian, C. Scheidegger, and S.A. Friedler. Problems with Shapley-value-based explanations as feature importance measures. *Proceedings of the 37th International Conference on Machine Learning*, 2020.

D. Lakens and E.R.K. Evers. Sailing from the seas of chaos into the corridor of stability: practical recommendations to increase the informational value of studies. *Perspectives on Psychological Science*, 9(3):278–292, 2014. doi: 10.1177/1745691614528520.

D. Lakens, Hilgard J., and J. Staaks. On the reproducibility of meta-analyses: six practice recommendations. *BMC Psychology*, 4(24), 2016. doi: 10.1186/s40359-016-0126-3.

L.J. et al. Launer. Body mass index, weight change, and risk of mobility disability in middle-aged and older women: the epidemiologic follow-up study of NHANES I. *JAMA*, 271(14):1093–1098, 1994.

S. Lindsay. Apology re Clark et al., 2020. URL `https://onlineacademiccommunity.uvic.ca/lindsaylab/2020/06/26/apology-re-clark-et-al/`.

H. Liu, J. Dacon, W. Fan, H. Liu, and J. Liu, Z.and Tang. Does gender matter? towards fairness in dialogue systems. *arXiv:1910.10486v1*, 2019.

F. Locatello, G. Abbati, T. Rainforth, T. Bauer, S. Bauer, B. Scholkopf, and O. Bachem. On the fairness of disentangled representations. *arXiv:1905.13662v1*, 2019.

G. Lohmann, K. Erfurth, K. Muller, and R. Turner. Critical comments on dynamic causal modelling. *Neuroimage*, 59(3):2322–9, 2012.

C. Louizos, U. Shalit, J. Mooij, D. Sontag, R. Zemel, and M. Welling. Causal effect inference with deep latent-variable models. *31st Conference on Neural Information Processing Systems*, 2017a.

C. Louizos, K. Swersky, Y. Li, M. Welling, and R. Zemel. The variational fair autoencoder. *arXiv:1511.00830*, 2017b.

S.M. Lundberg and S-I. Lee. A unified approach to interpreting model predictions. *31st Conference on Neural Information Processing Systems*, 2017.

S.M. Lundberg, G.G. Erion, and S-I. Lee. Consistent individualized feature attribution for tree ensembles. *Proceedings of the 34th International Conference on Machine Learning, Sydney, Australia*, 2017.

S.M. Lundberg, G. Erion, H. Chen, A. DeGrave, J.M. Prutkin, B. Nair, R. Katz, J. Himmelfarb, N. Bansal, and S-I. Lee. From local explanations to global understanding with explainable AI for trees. *Nature Machine Intelligence*, 2:56–67, 2020.

S. Makridakis, E. Spiliotis, and V. Assimakopoulos. The M4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020.

M. Marsman, F.D. Schonbrodt, R.D. Morey, Y. Yao, A. Gelman, and E-J. Wagenmakers. A bayesian bird's eye view of 'replications of important results in social psychology'. *R. Soc. open sci.*, 4, 2017.

G.N. Martin and R.M. Clarke. Are psychology journals anti-replication? a snapshot of editorial practices. *Frontiers in Psychology*, 8, 2017.

D. Mayo. Some surprising facts about (the problem of) surprising facts. *Studies in the History and Philosophy of Science*, 2013. doi: 10.1016/j.shpsa.2013.10.005.

B.B. McShane, D. Gal, A. Gelman, C. Robert, and J.L. Tackett. Abandon statistical significance. *The American Statistician*, 73:235–245, 2019.

P.E. Meehl. Why summaries of research on psychological theories are often uninterpretable. *Psychological Reports*, 66:195–244, 1990.

J. M. Mooij, O. Stegle, D. Janzing, K. Zhang, and B. Scholkopf. Probabilistic latent variable models for distinguishing between cause and effect. *NIPS*, pp. 1687–1695, 2010.

D. Moyer, S. Gao, R. Brekelmans, G. V. Steeg, and A. Galstyan. Invariant representations without adversarial training. *NeurIPS*, 2018.

K. P Murphy. *Machine Learning: A probabilistic Perspective*. MIT Press, Cambridge, Massachusetts, 2012.

R. R. Murphy. *Introduction to AI robotics*. MIT Press, Cambridge, Massachusetts, 2000.

M. Muthukrishna and Henric. What constitutes strong psychological science? the (neglected) role of diagnosticity and a priori theorizing. *Perspectives on Psychological Science*, 2017.

M.B. Nuijten, C.H.J. Hartgerink, M.A.L.M. van Assen, S. Epskamp, and J.M. Wicherts. The prevalence of statistical reporting errors in psychology (1985-2013). *Behavior Research Methods*, 48: 1205–1226, 2016.

K. Oberauer and S. Lewandowsky. Addressing the theory crisis in psychology. *Psychonomic Bulletin and Review*, 26:1596–1618, 2019.

A.J. Onwuegbuzie and L.G. Daniel. Uses and misuses of the correlation coefficient. *MSERA*, 1999.

A. Orben and D. Lakens. Crud (re)defined. *Advances in Methods and Practices in Psychological Science*, 3(2):238–247, 2020. doi: 10.1177/2515245920917961.

J. Pearl. *Causality*. Cambridge University Press, Cambridge, 2009.

J. Pearl, M. Glymour, and N.P. Jewell. *Causal inference in statistics: A primer*. Wiley, 2016.

J. Peters, D. Janzing, and B. Scholkopf. *Elements of Causal Inference*. MIT Press, Cambridge, Massachusetts, 2017.

O. Peters and M.J. Werner. A recipe for irreproducible results. *arXiv:1706.07773v1*, 2017.

M. Petersen, L. Balzer, D. Kwarsiima, N. Sang, G. Chamie, J. Ayieko, J. Kabami, A. Owaraganise, T. Liegler, F. Mwangwa, and K. Kadede. Association of implementation of a universal testing and treatment intervention with HIV diagnosis, receipt of antiretroviral therapy, and viral suppression in East Africa. *Journal of American Medical Association*, 2017.

H. Rahbar, D.S. Hippe, A. Alaa, S.H. Cheeney, and M. van der Schaar. The value of patient and tumor factors in predictive preoperative breast MRI outcomes. *Radiology: imaging cancer*, 2(4), 2020.

J.M. Rohrer. Thinking clearly about correlations and causation: Graphical causal models for observational data. *Advances in Methods and Practices in Psychological Science*, 2018. doi: https://doi.org/10.1177/2515245917745629.

I. Ropovik. A cautionary note on testing latent variable models. *Frontiers in Psychology*, 6, 2015. doi: 10.3389/fpsyg.2015.01715.

A. Rose. Are face-detection cameras racist? *Time Business*, 2010.

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

C. Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1:206–215, 2019.

K. Sassenberg and L. Ditrich. Research in social psychology changed between 2011 and 2016: larger sample sizes, more self-report measures, and more online studies. *Advances in Methods and Practices in Psychological Science*, 2(2):107–114, 2019.

A.M. Scheel, L. Tiokhin, P.M. Isager, and D. Lakens. Why hypothesis testers should spend less time testing hypotheses. *Perspectives on Psychological Science*, in press.

F.L. Schmidt and I-S. Oh. The crisis of confidence in research findings in psychology: is lack of replication the real problem? or is it something else? *Archives of Scientific Psychology*, 4:32–37, 2016.

B. Scholkopf. Causality for machine learning. *arXiv:1911.10500v1*, 2019.

U. Shalit, F. D. Johansson, and D. Sontag. Estimating individual treatment effect: generalization bounds and algorithms. *arxiv:1606.03976v5*, 2017.

L.S. Shapley. A value for n-person games. *Contributions to the Theory of Games*, 2(28):307–317, 1953.

C. Shi, D. M. Blei, and V. Veitch. Adapting neural networks for the estimation of treatment effects. *33rd Conference on Neural Information Processing Systems*, 2019.

G. Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

P.E. Shrout and J.L. Rodgers. Psychology, science, and knowledge construction: broadening perspectives from the replication crisis. *Annual Review of Psychology*, 69:487–510, 2018.

B. Siegerink, W. den Hollander, M. Zeegers, and R. Middelburg. Causal inference in law: an epidemiological perspective. *European Journal of Risk Regulation*, 7(1):175–186, 2016.

B.A. Spellman. A short (personal) future history of revolution 2.0. *Perspectives on Psychological Science*, 10(6):886–899, 2015. doi: 10.1177/1745691615609918.

G. V. Steeg and A. Galstyan. Information transfer in social media. *WWW*, 2012.

G.V. Steeg and A. Galstyan. Information-theoretic measures of influence based on content dynamics. *WSDM*, 2013.

J.R. Stevens. Replicability and reproducibility in comparitive psychology. *Frontiers in Psychology*, 8, 2017.

J. Stricker and A. Günther. Scientific misconduct in psychology. *Zeitschrift für Psychologie*, 227, 2019. doi: 10.1027/2151-2604/a000356.

M. Sundararajan and A. Najmi. The many Shapley values for model explanation. *arXiv:1908.08474v2*, 2020.

N. Taleb. Fooled by correlation: Common misinterpretations in social science. *Academia Online*, 2019.

J. B. Tenenbaum and T.L. Griffiths. Theory-based causal inference. *Neural Information Processing Systems*, 2002.

C. Tong. Statistical inference enables bad science; statistical thinking enables good science. *The American Statistician*, 73:246–261, 2019.

M. J. van der Laan and S. Rose. *Targeted Learning - Causal Inference for Observational and Experimental Data*. Springer International, New York, 2011.

M. J. van der Laan and S. Rose. *Targeted Learning in Data Science*. Springer International, Switzerland, 2018.

M. J. van der Laan and R. J. C. M. Starmans. Entering the era of data science: targeted learning and the integration of statistics and computational data analysis. *Advances in Statistics*, 2014.

T.J. VanderWeele. Causal inference and constructed measures: towards a new model of measurement for psychological constructs. *arXiv:2007.00520*, 2020.

M. J. Vowels, K. Mark, L. M. Vowels, and N. Wood. Using spectral and cross-spectral analysis to identify patterns and synchrony in couples' sexual desire. *PLoS One*, 2018. doi: 10.1371/journal.pone.0205330.

M. J. Vowels, N.C. Camgoz, and R. Bowden. Targeted VAE: Structured inference and targeted learning for causal parameter estimation. *Under Review*, 2020.

S. Wachter, B. Mittelstadt, and C. Russell. Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard Journal of Law and Technology*, 2018.

Y. Wang and D. M. Blei. The blessings of multiple causes. *arXiv:1805.06826v3*, 2019.

D.H. Wolpert and W.G. Macready. No free lunch theorems for optimization. *IEEE Transacions on Evolutionary Computation*, 1(67), 1997.

L. Yao, S. Li, Y. Li, M. Huai, J. Gao, and A. Zhang. Representation learning for treatment effect estimation from observational data. *32nd Conference on Neural Information Processing Systems (NeurIPS)*, 2018.

T. Yarkoni. The generalizability crisis. *PsyArXiv*, 2019. doi: https://doi.org/10.31234/osf.io/jqw35.

T. Yarkoni and J. Westfall. Choosing prediction over explanation in psychology: lessons from machine learning. *Perspectives on Psychological Science*, 2017.

J. Yoon, J. Jordan, and M. van der Schaar. GANITE: Estimation of individualized treatment effects using generative adversarial nets. *ICLR*, 2018.

W. Zhang, L. Liu, and J. Li. Treatment effect estimation with disentangled latent factors. *arXiv:2001.10652*, 2020.