Introduction to structural causal models in science studies

Thomas Klebel [1] and Vincent Traag [2, a)]
[1)] *Know-Center GmbH, Austria*
[2)] *Centre for Science and Technology Studies (CWTS), Leiden University, the Netherlands*

(Dated: 2025-03-04)

Sound causal inference is crucial for advancing the study of science. Incorrectly interpreting predictive effects as causal might lead to ineffective or even detrimental policy recommendations. Many publications in science studies lack appropriate methods to substantiate causal claims. We here provide an introduction to structural causal models for science studies. Structural causal models, usually represented in a graphical form, allow researchers to make their causal assumptions transparent and provide a foundation for causal inference. We illustrate how to use structural causal models to conduct causal inference using regression models based on simulated data of a hypothetical structural causal model of Open Science. The graphical representation of structural causal models allows researchers to clearly communicate their assumptions and findings, thereby fostering further discussion. We hope our introduction helps more researchers in science studies to consider causality explicitly.

Keywords: Causal inference; Science studies; Scientometrics

I. INTRODUCTION

Causal questions are pervasive in science studies: what are the effects of peer review on the quality of publications (Goodman *et al.*, 1994)? What is the influence of mentorship on protegees success (Malmgren, Ottino, and Nunes Amaral, 2010)? Do incentives to share research data lead to higher rates of data sharing (Woods and Pinfield, 2022)? Yet, answers to such questions rarely consider causality properly. Often, researchers investigate causal questions, but fail to employ adequate methods to make justified causal claims. As an example, there is a burgeoning literature investigating whether publishing Open Access leads to more citations. While the observational evidence seems to suggest such an effect, few studies use methods that would permit justified causal claims (Klebel *et al.*, 2023). Most scientists acknowledge that we should be "thinking clearly about correlation and causation" (Rohrer, 2018), but the implications of causal considerations are often ignored. Similar concerns were raised in the context of biases in science, such as gender bias (Traag and Waltman, 2022).

Uncovering causal effects is a challenge shared by many scientific fields. There are large methodological differences between fields, also with regards to inferring causality. Some fields are experimental, while others are observational. Some fields are historical, examining a single unfolded history, while others are contemporary, allowing to repeat observations. Some fields already have a long tradition with causal inference, while other fields have paid less attention to causal inference. We believe that science studies, regardless of whether that is scientometrics, science of science, science and technology studies, or sociology of science, have paid relatively little attention to questions of causality, with some notable exceptions (e.g., Aagaard and Schneider, 2017; Gläser and Laudel, 2016; Bol, de Vaan, and van de Rijt, 2018; Tomkins, Zhang, and Heavlin, 2017; Simsek, de Vaan, and van de Rijt, 2024; Luc *et al.*, 2021; Davis, 2020; Davis *et al.*, 2008). Quantitative science studies, like most quantitative social sciences, has a long history of working with regression models, and sometimes with more advanced statistical and mathematical models. However, it is not always clear how to interpret results from such approaches, and we believe a more explicit discussion of causality helps in clarifying and strengthening the interpretation.

We here provide an introduction to causal inference for science studies. In particular, we rely on structural causal models, which we believe are easier to communicate and relate to compared to the (formally equivalent) framework of potential outcomes. Multiple introductions to structural causal modelling already exist, typically covering specific fields (Rohrer, 2018; Arif and MacNeil, 2023; Elwert, 2013; Hünermund and Bareinboim, 2023; Deffner, Rohrer, and McElreath, 2022). In addition, Liu *et al.* (2023) discusses several other approaches to causal inference in science studies, while Dong *et al.* (2022) focuses on matching strategies. Beyond these shorter and domain-specific introductions, there are also comprehensive text-books (Huntington-Klein, 2021; Cunningham, 2021; Pearl, 2009) that provide much more detail and explanation than we can provide here. By applying structural causal models to familiar examples from science studies, we aim to make causal inference more accessible to researchers who are unfamiliar with these approaches. We believe structural causal models are relevant to science studies broadly, but they are particularly relevant in quantitative science studies. We avoid technicalities, so that the core ideas can be understood even with little background in statistics. Although we focus on structural causal models, not all causal thinking can necessarily be easily expressed in these models.

The fundamental problem in causal inference is that we never know for sure the answer to the "what-if" question.

[a)] E-mail: v.a.traag@cwts.leidenuniv.nl

For instance, suppose that a professor received tenure. We can observe her publications when she received tenure. Would she also have received tenure, if she had not published that one paper in a high-impact journal? We cannot simply observe the answer, since that situation did not materialize: she in fact did publish that paper in a high-impact journal, and in fact did receive tenure. The so-called counterfactual scenario, where she did not publish that paper and received tenure (or not), is unobservable. This unobservable counterfactual scenario is the fundamental problem.

Experiments are often helpful in getting causal answers. By controlling the conditions, and only actively varying one condition, we can recreate counterfactual scenarios, at least on average, assuming conditions are properly randomised. There are also some experimental studies in science studies, for instance studying the effect of randomly tweeting about a paper or not (Luc *et al.*, 2021; Davis, 2020), making papers randomly openly available (Davis *et al.*, 2008), or studying affiliation effects by experimentally comparing double-anonymous peer review with single-anonymous peer review (Tomkins, Zhang, and Heavlin, 2017). However, there are many questions that do not allow for an experimental setup. For example, randomising scholars' career age or research field is impossible. But even in experimental settings there are limitations to causal inference. For instance, noncompliance in experimental settings might present difficulties (Balke and Pearl, 2012), such as certain types of reviewers being more likely to try to identify authors in a double-anonymous peer review experiment. Additionally, scholars might be interested in identifying mediating factors when running experiments, which further complicates identifying causality (Rohrer *et al.*, 2022). In other words, causal inference presents a continuum of challenges, where experimental settings are typically easiest for identifying causal effects—but certainly no panacea—and observational settings are more challenging—but certainly not impossible.

In this paper we introduce a particular view on causal inference, namely that of structural causal models (Pearl, 2009). Structural causal models are formally equivalent to another causal inference framework known as potential outcomes (Imbens and Rubin, 2015). We believe structural causal models are a relatively straightforward approach to causal inference with a clear visual representation of causality. It should help researchers to reason and discuss their causal thinking more easily, even though not all causal considerations can be expressed as structural causal models. In the next section, we explain structural causal models in more detail. We then cover some case studies based on simulated data to illustrate how causal estimates can be obtained in practice. We close with a broader discussion on causality.

II. STRUCTURAL CAUSAL MODELS

Structural causal models focus, as the name suggests, on the structure of causality, not on the exact details. That is, structural causal models are only concerned with whether a certain factor is causally affected by another factor, not whether that effect is linear, exponential, or an "interaction" with some other effects. Such structural models can be represented by causal diagrams. This graphical approach makes it relatively easy to discuss about causal models and assumptions, because it does not necessarily involve complicated mathematics.

Sometimes, assumptions about specific functional dependencies can be made, and this might help causal inference. For instance, a well-known causal inference strategy is called "difference-in-difference". A key assumption in that strategy is something called "parallel trends". Several such approaches are discussed by Liu *et al.* (2023) in the context of science studies. Not having to deal with such details simplifies the approach and makes it easier to understand the core concepts. But sometimes it also simplifies too much. We can always make stronger assumptions, and sometimes, these stronger assumptions allow us to draw stronger conclusions. But without assumptions, we cannot conclude anything.

The overall approach to causal inference using structural causal models would be the following:

1. Assume a certain structural causal model.
2. Use the assumed structural causal model to understand how to identify causal effects.
3. Identified effects can be interpreted causally *under the assumed structural causal model.*

Whatever structural causal model we construct, it will always be an assumption. Constructing such a structural causal model can be based on domain expertise and prior literature in the area. Whether a structural causal model is realistic or not might be debated. In particular, a common issue is what is sometimes referred to as "omitted-variable bias", which refers to variables that are incorrectly omitted from the structural causal model. By making causal assumptions explicit, we can clarify such discussions, and advance our common understanding.

We cannot always use empirical observations to discern between different structural causal models. That is, different structural causal models can have the same testable implications, and no observations would help discern between them. However, there might also be testable implications that do differ between different structural causal models. We can then put the two (or more) proposed theoretical structural causal models to the test, using empirical evidence to decide which structural causal model is *incorrect*. Note the emphasis on incorrect: we cannot say that a structural causal model is correct, but we can say that a structural causal model is incorrect, if it is inconsistent with the observations. In summary, if we propose a certain structural causal model to try to identify a causal effect, we should make sure that

its testable implications are at least consistent with the empirical evidence we have.

Nonetheless, any structural causal model always remains a simplification of reality, and is usually designed for a specific causal question. For example, a structural causal model of the entire academic system, containing each and every detail about potential effects, is overly detailed and likely not useful for the majority of empirical studies. For most studies, a simpler structural causal model is probably more productive. In general, it is advisable to aim for structural causal models that are as simple as possible, but as complex as necessary. In some cases, problems of causal identification might emerge in simple structural causal models, and are not heavily dependent on specific details. Adding more nuance to a structural causal model may then not solve the problem that was identified in a simpler structural causal model. However, sometimes problems might only become apparent with more complex structural causal models, and additional nuance might reveal that identifying a causal effect is more challenging. We encounter and discuss this in some examples later.

The main challenge then is to use a given structural causal model to identify a causal effect: what factors should be controlled for and, equally important, what factors should not be controlled for? We introduce an answer to that question in the next subsection. The introduction we provide here only covers the basics. We explicitly provide an introduction that is as simple as possible, in order to be understandable to a broad audience. Our introduction covers many typical situations that can be encountered, but there are other cases that cannot be understood without using a more formal logic known as do-calculus (Pearl, 2009).

To provide an introduction useful to readers and scholars in science studies, we consider the case of Open Science, a movement and practice of making research processes more open and transparent (Fecher and Friesike, 2014). Many studies have been conducted on the potential impacts Open Science might have on academia, society, and the economy (Klebel *et al.*, 2023; Tennant *et al.*, 2016). However, studies on specific types of Open Science impact, such as those on the Open Access citation advantage, often lack a clear understanding of causal pathways and thus fail to develop a meaningful strategy for estimating causal effects. Our introduction shows how causal inference could be leveraged to improve these and similar studies.

A. Introducing DAGs

It is convenient to represent a structural causal model using a directed acyclic graph (DAG). A DAG is a directed graph (sometimes called a network) where the nodes (sometimes called vertices) represent variables, and the links (sometimes called edges) represent causal effects. A DAG is acyclic, meaning there are no directed cycles, so that if $X \to Z \to Y$, there cannot be a link $Y \to X$ (or $Y \to Z$ or $Z \to X$). See Table I for an overview of some of the concepts related to DAGs.

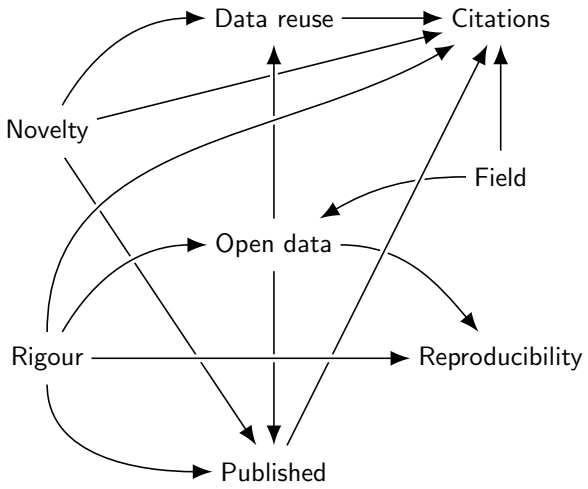

Figure 1. Hypothetical structural causal model on Open Science

If $X \to Y$, it means that Y directly depends on X, that is, Y is a function of X. We do not specify what function exactly, so it can be a linear function, an exponential function, or any other function. Interactions between variables, moderators, hurdles, or any other functional specifications are not indicated separately, and all can be part of the function. A variable X that has a direct causal effect on Y is called a parent of Y.

Throughout this introduction, we work with a single example DAG on Open Science (see Figure 1). In this DAG, *Novelty* and *Rigour* are both assumed to affect the number of *Citations* and whether something will be *Published* or not. Although preprints or working papers can also be considered published, for the sake of simplicity we here use the term *Published* to refer to journal publications only. Unlike *Novelty*, *Rigour* influences whether data is made available openly: scholars that are doing more rigorous research may be more likely to share their data openly. Unlike *Rigour*, *Novelty* affects *Data reuse*; open data from a more novel study may be more likely to be reused by other researchers. If data is reused, the original study might be cited again, so *Data reuse* is assumed to affect *Citations*. In some cases, *Open data* will be mandated by a journal, and so whether something will be *Published* may also depend on *Open data*. Whether something is *Reproducible* is assumed to be affected by the *Rigour* of the study, and also by *Open data* itself: studies that share data might make it easier for other scientists to reproduce their findings. Finally, *Citations* are also influenced by the *Field* of study (some fields are more citation intensive), as is *Open data* (data sharing is more common in some fields).

As explained earlier, this DAG is a simplification, and we can debate whether it should be changed in some way.

Table I. Overview of concepts for Directed Acyclic Graphs (DAG).

Concept	Explanation
Node, vertex	Represents a variable in a DAG
Link, edge	Represents a causal effect from one node on another in a DAG.
Acyclic	No cycles (e.g. $X \to Y \to Z \to X$) are present.
Parents	The parents of a node Y are the nodes that point to Y.
Path	A series of nodes connected through links. Can be directed, when respecting the direction of the link (e.g. $X \to Y \to Z$) or undirected, when ignoring the direction of the link (e.g. $X \to Y \leftarrow Z$).
Causal path	A path from X to Y is causal if it is directed, i.e. all links respect the direction (e.g. $X \to Z \to Y$)
Non-causal path	A path between X and Y is non-causal if it is undirected, i.e. some links do not respect the direction (e.g. $X \to Z \leftarrow Y$).
Open path	When a path between two nodes X and Y is open, there is an association between X and Y. See also Figure 2.
Closed path	When all paths between two nodes X and Y are closed, there is no association between X and Y. See also Figure 2.

Although the DAG is perhaps not fully realistic, it is plausible and, by and large, consistent with the Open Science literature. This DAG is constructed without one particular causal question in mind. Instead, we illustrate all the necessary concepts using this example, and use this DAG for multiple possible causal questions. For a particular study, it might be best to construct a particular DAG for the specific causal question. A reasonable starting point for constructing a DAG for a particular causal question of X on Y might be the following: (1) consider all factors that affect and are affected by X and/or Y; (2) consider how these factors are causally related between each other. There might be additional relevant considerations, but it should provide a reasonable simplification to start with.

A useful tool for working with DAGs is called `dagitty` (Textor *et al.*, 2016), which is available from the website http://dagitty.net, which also contains many useful pointers to additional introductions and tutorials.

B. Using DAGs to identify causal effects

We are interested in the causal effect of one variable X on another variable Y. As the popular adage goes, correlation does not imply causation. That is, X and Y might be correlated, even if X does not affect Y. For instance, in Figure 1 *Reproducibility* and *Published* are correlated because both are affected by *Open data*, but *Reproducibility* does not have any causal effect on *Published* or vice versa.

Most scholars will be acquainted with problems of confounding effects, and that we somehow need to "control" for confounders. But there are also other factors besides confounders. Most scholars will also be acquainted with mediators. Fewer scholars will be acquainted with colliders. Controlling for a collider often leads to incorrect causal inferences. Hence, the question of what variables to control for is more complicated than just controlling for confounders. In particular, colliders raise the question what we should *not* control for. In this section, we explain confounders, mediators and colliders in more detail, and use DAGs to understand which factors we should control for, and which factors we should *not* control for.

1. Paths in DAGs

In DAGs, we think of correlation and causation in terms of paths between variables. In a graph, a path between two nodes consists of a series of connected nodes. That is, we can move from one node to another across the links between the nodes to reach another part of the graph. For example, in Figure 1 we can move from *Novelty* to *Data reuse* to *Citations*. In this example, the path follows the direction of the links. Paths that follow the direction of the links resemble the flow of causality, and we refer to them as causal paths. That is, *Novelty* affects *Data reuse*, which in turn affects *Citations*. This is an indirect causal effect of *Novelty* on *Citations*, mediated by *Data reuse*. There is another indirect causal effect of *Novelty* on *Citations*, mediated by *Published*. In addition, there is also a link directly from *Novelty* to *Citations*, which represents a direct causal effect. The combination of the two indirect effects and the direct effect is known as the total causal effect.

In addition, there are also paths that do not follow the direction of the links. This can be most easily done by simply ignoring the directions, and also allowing to traverse links upstream, so to speak. There is then a path between *Open data* and *Citations* through *Field*. There is not a single direction that we follow, and the path looks like *Open data* $\leftarrow$ *Field* $\rightarrow$ *Citations*. Paths that do not follow a single direction do not represent a causal effect, and we refer to them as non-causal paths.

The key insight is that two variables that are connected through certain paths are correlated, even if they are not connected through any causal paths. We discern two types of paths. One type of path, through which two

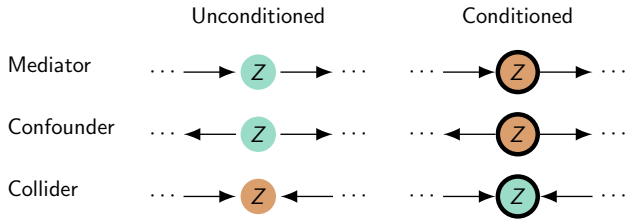

Figure 2. Overview of open and closed nodes. Open nodes are marked in green, closed nodes are marked in orange. If all nodes on a path are open, the path is open. If any node on a path is closed, the path is closed.

variables are correlated, is called an *open* path. Another type of path, through which two variables are not correlated, is called a *closed* path. If there are no open paths between two variables, the two are not correlated. Both causal paths and non-causal paths can be open or closed. Indeed, if there is a non-causal path that is open, two variables are correlated, but this "correlation does not imply causation".

Formalising this slightly, two variables X and Y are correlated if there is an open path between X and Y. If there are no open paths between X and Y, they are not correlated. We can identify a causal effect of X on Y by *closing* all non-causal paths between X and Y and by *opening* all causal paths from X to Y. Whether a path is open or closed depends on the types of variables on a path, and whether those variables are conditioned on. We explain this in more detail in the next subsection, and provide a visual summary of the explanation in Figure 2.

As explained, all paths between X and Y need to be considered, regardless of their direction. That is, $X \rightarrow Z \rightarrow Y$ is a type of path that we should consider, but also $X \leftarrow Z \rightarrow Y$ and $X \rightarrow Z \leftarrow Y$. Going back to the paths we considered earlier: if we are interested in the causal effect of *Open data* on *Citations*, there is a directed, causal path from *Open data* to *Data reuse* to *Citations*, but there is also a non-causal path between *Open data* and *Citations* that runs through *Field*.[1]

We call a path open when all the nodes, i.e. variables, on the path are open. If there is a single closed variable on a path, the entire path is closed. You can think of this as a sort of information flow: if all nodes are open, information can flow through, but a single closed node blocks the flow of information. We can change whether a variable should be considered open or closed by conditioning on it. By closing a variable, we can therefore close a path. By opening a variable, we can potentially open a path, unless the path is still closed by another variable.

[1] Note that there are many additional paths in this example: *Open data* $\leftarrow$ *Rigour* $\rightarrow$ *Citations*, *Open data* $\rightarrow$ *Reproducibility* $\leftarrow$ *Rigour* $\rightarrow$ *Citations*, etc.

There are many ways in which we can condition on a variable. A common approach in quantitative analysis is to include such a variable in a regression analysis. But another way is to analyse effects separately for various categories of some variable. For example, we can condition on *Field* by performing an analysis for each field separately. This can be thought of as comparing cases only within these categories. Other approaches include for example so-called matching procedures. When matching cases on a certain variable, we only compare cases which are the same (or similar) on that variable. Finally, in science studies, indicators are frequently "normalised", especially citation indicators (Waltman and van Eck, 2019), which amounts to conditioning on the variables used for the normalisation.

2. *Confounders, colliders and mediators*

We can discern three types of variables: a confounder, a collider and a mediator. Whether a variable Z is a confounder, a collider or a mediator depends on how Z is connected on a path between X and Y. Below we consider each type of variable in more detail.

The first type of variable that we consider is a confounder. A confounder Z is always connected like $X \leftarrow Z \rightarrow Y$. Here Z is the common cause for both X and Y. A confounder is open when not conditioned on. If we condition on a confounder, it is closed. Usually, we want to close paths with confounders, as the paths do not represent a causal effect. For example, in Figure 1, *Field* plays the role of a confounder on the path between *Open data* and *Citations*. That path is open; we can close it by conditioning on *Field*.

The second type of variable that we consider is a collider. A collider Z is always connected like $X \rightarrow Z \leftarrow Y$. Here Z is affected by both X and Y. A collider is closed when not conditioned on. If we condition on a collider, it is opened. Usually, we want to keep paths with a collider closed, as the paths do not represent a causal effect. For example, in Figure 1, *Published* plays the role of a collider on the path between *Rigour* and *Novelty*. That path is closed; we can open it by conditioning on *Published*.

Finally, the third type of variable that we consider is a mediator. A mediator Z is always connected like $X \rightarrow Z \rightarrow Y$. Here, Z is affected by X and in turn Z affects Y, so that Z mediates the indirect causal effect of X on Y. A mediator is open when not conditioned on. If we condition on a mediator, it is closed. Usually, we want to keep paths with mediators open, as the paths represent a causal effect. However, it might be that we are interested in the *direct* effect of X on Y, instead of the total effect of X on Y. By controlling for a mediator Z we can close the indirect path $X \rightarrow Z \rightarrow Y$, and estimate the direct path $X \rightarrow Y$ (assuming there are no other indirect paths left). For example, in Figure 1, *Open data* is a mediator between *Rigour* and *Reproducibility*. That path is open; we can close it by conditioning on *Open data*. This is

relevant if we try to identify the direct causal effect of *Rigour* on *Reproducibility*.

Note that the same variable can play different roles in different paths. For example, in Figure 1, *Open data* plays the role of a confounder in the path *Reproducibility* ← *Open data* → *Data Reuse* → *Citations*. At the same time, *Open data* plays the role of a collider in the path *Reproducibility* ← *Rigour* → *Open data* ← *Field* → *Citations*. The former path is open, while the latter path is closed. If we are interested in the causal effect, both paths should be closed, since neither represents a causal effect. However, if we condition on *Open data*, we close the path where *Open data* is a confounder, while we open the path where *Open data* is a collider. Hence, we cannot close both paths by conditioning on *Open data*. If we cannot condition on other variables, for example because we did not collect such variables for a study, we have no way of identifying the causal effect of *Reproducibility* on *Citations*.

Our assumed DAG implies that there should be no causal effect of *Reproducibility* on *Citations*. If we condition on *Open data* and *Rigour* all non-causal paths are closed, meaning that we then expect to find no correlation. If, in contrast, we still find a non-zero correlation after conditioning on *Open data* and *Rigour*, it means our DAG is incorrect, and we need to revise it. This is only one of the testable implications of our DAG, but there are also others, and each offers an opportunity to falsify our DAG. Even if all testable implications find empirical support, this does not imply the DAG is correct, since other DAGs might have the same testable implications.

III. CASE STUDIES

In this section, we apply the concepts introduced above to potential research questions, demonstrating how to estimate causal effects. We show how a researcher can use a hypothesised causal model of the phenomenon under study to estimate causal effects. We use the DAG introduced earlier (Figure 1) to illustrate our estimation strategies.

For the purposes of these hypothetical examples, we simulate data according to the DAG in Figure 1. As explained, a DAG only specifies that a variable is affected by another variable, but it does not specify how. In Appendix A we provide more details of how we simulate data.

Regression analysis is the common workhorse of quantitative analysis, also in science studies. We use regression analysis to illustrate how a researcher might analyse their data to provide causal estimates[2]. Of course, more complex analytical approaches, such as Bayesian models or non-linear models can also be used. Such models might have great scientific, philosophical, or practical benefits, but they are certainly no prerequisite for sound causal inference. Moreover, having complex models is no substitute for sound causal inference, and wrong causal conclusions can still be drawn from complex models. From that point of view, using simpler methods while paying proper attention to causality might be preferred over using complex methods while ignoring issues of causality.

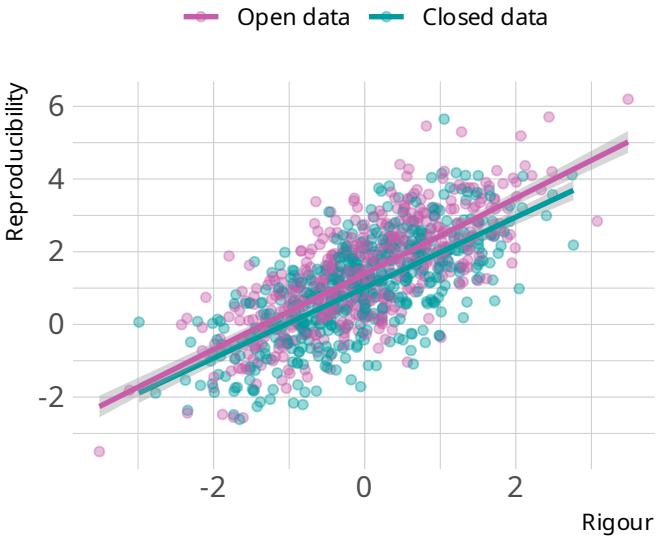

Figure 3. Effect of *Rigour* and *Open data* on *Reproducibility*

A. The effect of Rigour on Reproducibility

To provide a first impression of the simulated data, and some intuition of how we can estimate causal effects, we first analyse the effect of *Rigour* on *Reproducibility* (see Figure 3). *Rigour* and *Reproducibility* are clearly positively correlated: higher *Rigour* is associated with higher *Reproducibility*. We also see that the overall level of reproducibility tends to be higher if there is *Open data*.

Following our model (Figure 1), *Rigour* and *Open data* are the only variables influencing *Reproducibility*. Let us consider the total causal effect of *Rigour* on *Reproducibility*. There are several paths between *Rigour* and *Reproducibility*, some causal, some non-causal. The model shows two causal paths: a direct effect *Rigour* → *Reproducibility* and an indirect effect *Rigour* → *Open data* → *Reproducibility*, where the effect is mediated by *Open data*. The non-causal paths are more convoluted: all run through *Citations* and/or *Published*, with both variables acting as colliders on these paths. The non-causal paths are hence all closed, unless we condition on any of the colliders.

Since the causal paths are open, and the non-causal paths are closed, we do not have to control for anything. We can estimate the total causal effect of *Rigour* on *Re-*

[2] We will write the equation in the typical style of R. For example, $Y \sim X + A$, refers to the linear equation $Y = \alpha + \beta_X X + \beta_A A$, where we are interested in estimating the coefficients α, β_X and β_A.

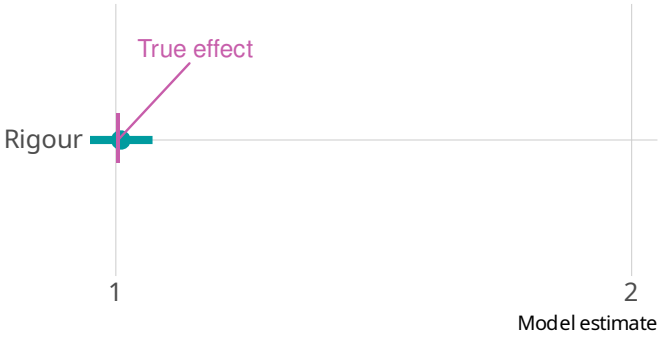

Figure 4. Effect of *Rigour* on *Reproducibility*, estimated with a simple linear regression.

producibility simply with a regression of the form

$$\text{Reproducibility} \sim \text{Rigour}$$

Since we simulated the data, we can calculate the "true" causal effect, which in this case is 1 (see Appendix B for details). We can hence validate our regression approach and see if it is capable of correctly inferring the true causal effect. Figure 4 shows that the regression approach is capable of retrieving the correct result. We deliberately chose a moderate sample size of 1000 for our simulation. Point estimates derived from the simulated data thus only approximate the theoretical values.

The example serves to highlight two points. First, it can be helpful to plot the data to gain an intuitive understanding of what the assumed relationship looks like. Second, sound causal inference does not necessarily involve controlling for many variables. In some cases, a simple regression might be all that is needed. Not all causal effects are equally straightforward to measure, as the next examples show.

B. The effect of Open data on Citations

Suppose we are interested in the total causal effect of *Open data* on *Citations*. Previous research on the topic indicates that articles sharing data tend to receive more citations (Piwowar, Day, and Fridsma, 2007; Piwowar and Vision, 2013; Kwon and Motohashi, 2021). According to our model (Figure 1), there are multiple pathways from *Open data* to *Citations*. To estimate the causal effect, we need to make sure that all causal paths are open, and all non-causal paths are closed (see panel A in Figure 5).

There are two causal paths, both indirect: one mediated by *Data reuse* and one mediated by *Published*. To estimate the total causal effect of *Open data* on *Citations* we should not control for either *Data reuse* or *Published*. In contrast, typical approaches in scientometrics examine only the literature published in journals and thereby implicitly condition on *Published*. This implicit conditioning closes the causal path, and thus biases our estimate of the total causal effect of *Open data* on *Citations*.

The non-causal paths pass through *Rigour*, *Field*, *Reproducibility* or *Novelty*. On all paths passing through *Rigour*, it acts as a confounder, and we can hence close all these non-causal paths by controlling for *Rigour*. There is only one non-causal path passing through *Field*, where it acts as a confounder, and we can close the path by conditioning on it. The non-causal paths that pass through *Reproducibility* are already closed, because it acts as a collider on those paths. Finally, all non-causal paths passing through *Novelty* are already closed because *Data reuse* and *Published* act as colliders on those paths. In summary, we should control for *Rigour* and *Field*.

The final regression model to estimate the causal effect of *Open data* on *Citations* is thus as follows:

$$\text{Citations} \sim \text{Open data} + \text{Field} + \text{Rigour}$$

Panel B of Figure 5 shows the effect estimates from our regression, alongside the true effect of *Open data* on *Citations*, which is 5.39. We can see that our model is indeed able to estimate the causal effect of *Open data* on *Citations*.

This example highlights key components of causal inference: controlling for confounders (*Rigour* and *Field*), not controlling for mediators (*Data reuse* and *Published*), and not controlling for colliders (*Reproducibility*). This shows that constructing an appropriate DAG is very helpful when aiming to draw causal conclusions. Without making assumptions explicit via a DAG, it would be much more difficult to discuss which variables should be controlled for and which not.

Some researchers might be tempted to defer the decision of what variables to control for to the data (for example via stepwise regression) or not make any decision at all by simply including all available variables, an approach termed "causal salad" by McElreath (2020). However, neither approach is able to correctly identify the correct variables to control for. Stepwise regression would in this case suggest including the mediating variables (and even excluding *Open data*), leading to wrong causal conclusions (see Appendix D). Including all variables could similarly lead the researcher to conclude that *Open data* has no effect on *Citations* (see Appendix E).

The example highlights that relatively simple DAGs are often sufficient to uncover limitations to identifying causal effects. For instance, if we had not measured *Field*, controlling for it and identifying the causal effect would become impossible. In that case, it is irrelevant whether there are any other confounding effects between *Citations* and *Open data*, since those effects do not alleviate the problem of being unable to control for *Field*.

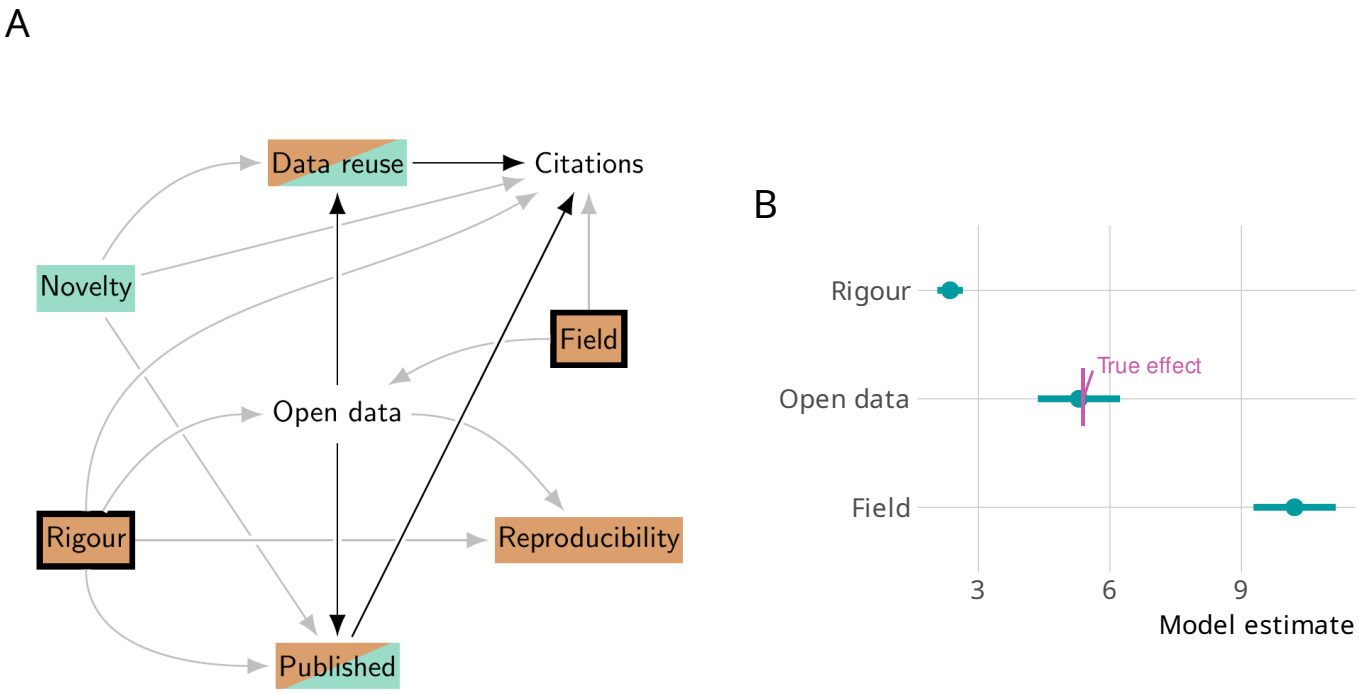

Figure 5. Effect of Open data on Citations. A: DAG illustrating which variables to condition on (or not). Open nodes are marked in green, closed nodes are marked in orange, and nodes that are open in one path but closed in another are marked semi-green and semi-orange. Nodes that are controlled for are marked by a thick outline. B: Effect estimate (regression coefficients with 95%-CI).

C. The effect of Open data on Reproducibility

Suppose we are interested in the causal effect of *Open data* on *Reproducibility*. Such an effect is often assumed in debates on how to increase the reproducibility across the scholarly literature (Molloy, 2011). The empirical evidence so far is less convincing (Nuijten *et al.*, 2017; Hardwicke *et al.*, 2018, 2021; Nosek *et al.*, 2022, p. 721). In our DAG in Figure 1, we assume there is a causal effect of *Open data* on *Reproducibility*. The causal effect is direct, there is no indirect effect of *Open data* on *Reproducibility*. Although the DAG does not specify these parametric assumptions, in our simulation, the effect is positive.

1. *Conditioning on a collider may bias estimates*

Many bibliometric databases predominantly cover research published in journals or conferences. Science studies frequently relies on such bibliometric databases for analysis. By only considering the literature published in journals, we implicitly condition on *Published*. On the path *Open data* → *Published* ← *Rigour* → *Reproducibility*, *Published* acts as a collider. As discussed in Section II, conditioning on a collider can bias our estimates.

We show the level of *Reproducibility* for *Open data* when considering only research that is *Published* in Figure 6. The level of *Reproducibility* is then lower with

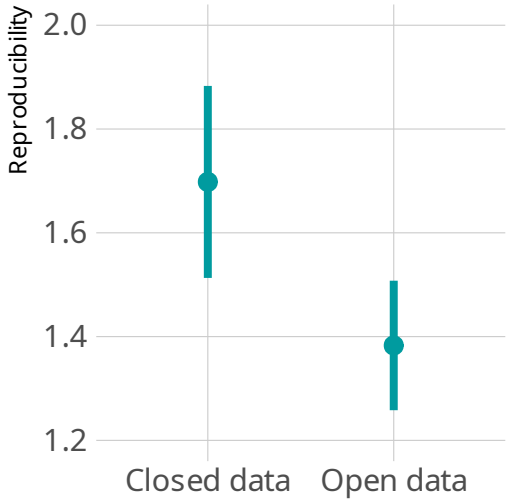

Figure 6. Reproducibility of research published in journals with and without *Open data*. Displaying means with 95%-CI.

Open data than without *Open data*. This is rather counterintuitive, since the causal effect of *Open data* on *Reproducibility* is in fact positive in our model.

The apparent negative effect is due to the fact that we conditioned on *Published*, by analysing only the published research. If we condition on a collider, we open that path; in this case we open the path *Open data* → *Published* ← *Rigour* → *Reproducibility*. How condition-

ing on a collider biases the estimates is difficult to foresee, especially in more complicated cases. In this case, however, there is a reasonably intuitive explanation. In our model, *Published* depends on both *Open data* and *Rigour* (and *Novelty*, but that is not relevant here): research is more likely to be published in a journal if it has *Open data* and if it is more rigorous. As a result, research that is published in a journal without *Open data* tends to have higher *Rigour*. If research had neither *Open data* nor sufficiently high *Rigour*, it would be less likely to be published in a journal at all[3]. Therefore, published research without *Open data* has higher *Rigour*. This higher *Rigour* in turn affects *Reproducibility*, leading to higher *Reproducibility* for published research without *Open data*.

The example shows how we can draw wrong, and even opposite, conclusions if we do not use clear causal thinking. Based on the results in Figure 6, some might incorrectly conclude that *Open data* has a negative causal effect on *Reproducibilty*. However, in our model, *Open data* has a positive causal effect on *Reproducibility*. Hence, we should take great care in interpreting empirical results without causal reflection.

Sometimes, when determining what variables to control for, scholars are inclined to think in terms of ensuring that cases are "comparable", or to make sure that we compare "like with like". Although the intuition is understandable, its application is only limited, and at times can be misleading. That is, using the "like with like" intuition, we might be inclined to condition on *Published*, because we then compare published papers with other published papers. If we do so, we bias the estimation of the causal effect of *Open data* on *Reproducibility*, as explained above. Comparing "like with like" may thus create problems.

2. Identifying the causal effect

As explained, conditioning on the collider *Published* opens the non-causal path *Open data* → *Published* ← *Rigour* → *Reproducibility*. This non-causal path is open because *Published* is open (because it is a collider that is conditioned on), and because *Rigour* is open (because it is a confounder that is not conditioned on). We can hence close this non-causal path by conditioning on *Rigour*. In addition, *Rigour* acts as a confounder on the non-causal path *Open data* ← *Rigour* → *Reproducibility*. To identify the causal effect, we hence also need to close this non-causal path by conditioning on *Rigour*. In short, we close both non-causal paths by conditioning on *Rigour*.

[3] In our DAG, *Published* is also affected by *Novelty*, and the same reasoning applies there. In this case, research that is published in a journal, but that does not share *Open data* and has low *Rigour*, is then more likely to have high *Novelty*. Otherwise, it again would most likely not have been published in a journal at all.

Panel A in Figure 7 shows the DAG for this question. There are no other non-causal paths that are open, and no causal paths that are closed. The regression model is thus

$$\text{Reproducibility} \sim \text{Open data} + \text{Rigour}$$

but still restricted to only published research.

The true effect of *Open data* on *Reproducibility* in our simulation is 0.4 (see Table II). After controlling for *Rigour*, our regression model is able to estimate this parameter correctly (panel B of Figure 7), although we are only considering research published in journal articles, therefore "conditioning on a collider".

The reason we can estimate the parameter correctly is that conditioning on *Rigour* closes the path *Open data* → *Published* ← *Rigour* → *Reproducibility*. Whether *Published* is conditioned on is then irrelevant for the identification of the causal effect. If we consider all research instead of only research published in journal articles, our estimate only changes from 0.369 to 0.379.

In identifying the causal effect of *Open data* on *Reproducibility*, we do not need to control for other variables, such as *Novelty*. If there were an additional confounder between *Published* and *Data reuse*, this would not change anything in terms of what variables we should control for to identify the effect of *Open data* on *Reproducibility*. This shows how making the DAG richer and more nuanced does not necessarily change the identification. Of course, other changes to the DAG do change the identification: if there were another confounder between *Open data* and *Reproducibility*, we would need to control for it.

3. Interpreting regression coefficients and measurement problems

Often, researchers not only interpret the coefficient that is the subject of their main research question, but also interpret the other coefficients. However, it is easy to draw wrong conclusions for those other coefficients, and interpret them incorrectly as causal effects. Since these other effects are often represented in the second table in an article, this was referred to as the "Table 2 fallacy" by Westreich and Greenland (2013).

Let us briefly consider the coefficient for the factor that we controlled for, namely *Rigour*. We estimated the coefficient for *Rigour* in our regression model to be about 1. What does this estimate represent? From the point of view of the effect of *Rigour* on *Reproducibility* there are two causal paths: one directly from *Rigour* to *Reproducibility* and one indirectly, mediated by *Open data* (we illustrated this earlier in Figure 4). Since we controlled for *Open data* in our regression model, it means we closed the indirect causal path. All other non-causal paths are also closed, and so there is only one path that is still open, which is the direct causal path from *Rigour* to *Reproducibility*. Hence, our estimate of 1 should represent the direct causal effect of *Rigour* on *Reproducibility*, and

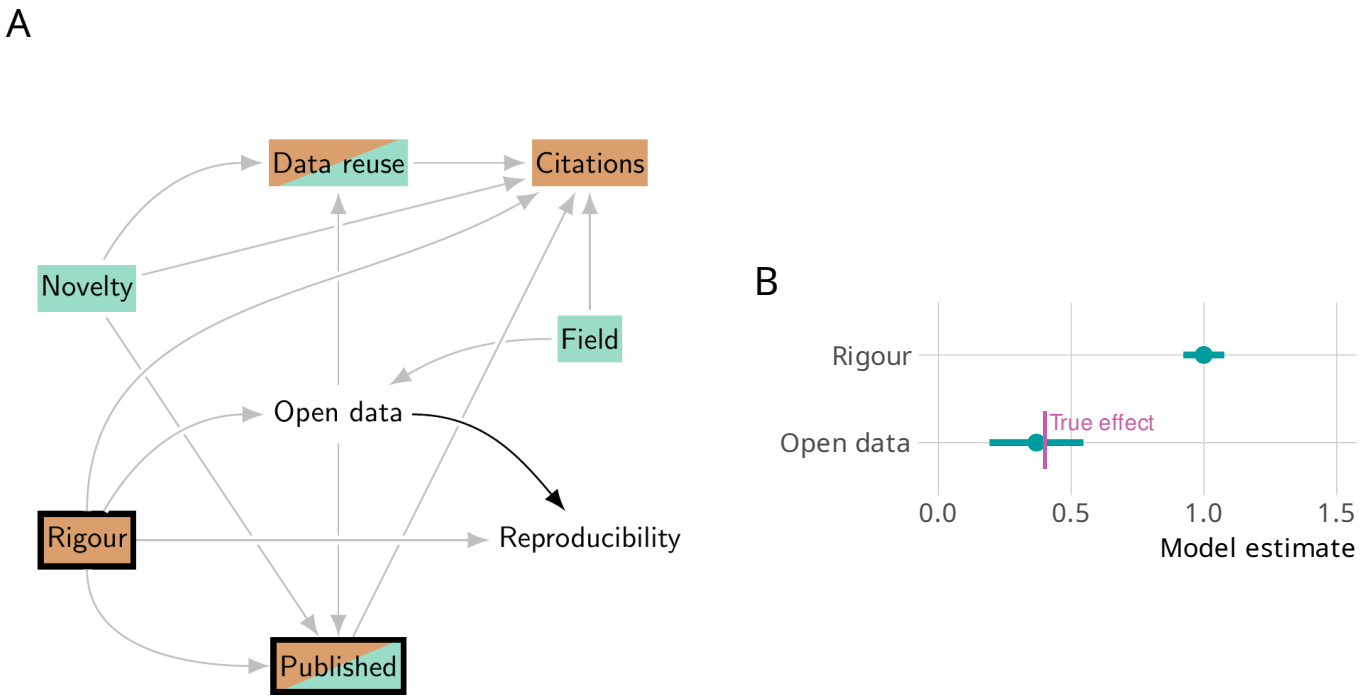

Figure 7. Effect of *Open data* on *Reproducibility*. A: DAG illustrating which variables to condition on (or not). Open nodes are marked in green, closed nodes are marked in orange, and nodes that are open in one path but closed in another are marked semi-green and semi-orange. Nodes that are controlled for are marked by a thick outline. B: Effect estimate (regression coefficients with 95%-CI).

indeed this corresponds with the coefficient we used in our simulation (see again Table II).

In the example above, we should interpret the estimate of the effect of *Rigour* on *Reproducibility* as a direct causal effect, not as a total causal effect. In other cases, coefficients for the controlled factors might not correspond to any causal effect. Indeed, we should carefully reason about any effect we wish to identify, and not interpret any estimates for controlled variables as causal without further reflection.

Additionally, most empirical studies will suffer from measurement problems. That is, the concept of interest is often not observed directly, but measured indirectly through some other proxies or indicators. These issues can be readily incorporated in structural causal models, and might make certain limitations explicit. For example, in the analysis above we controlled for *Rigour* to infer the causal effect of *Open data* on *Reproducibility*, but in reality, we most likely cannot control for *Rigour* directly. Instead, we are controlling for the measurement of *Rigour*, for example as measured by expert assessment of the level of rigour. We could include this in the structural causal model as *Rigour* → *Rigour measurement*. We cannot directly control for *Rigour*, and we can only control for *Rigour measurement*, which does not (fully) close the backdoor path between *Open data* and *Reproducibility*, and might hence still bias the estimate of the causal effect. If *Rigour measurement* would additionally be affected by other factors, such as *Published*, this might introduce additional complications. Taking measurement seriously can expose additional challenges that need to be addressed (McElreath, 2020, Chapter 15; Esterling, Brady, and Schwitzgebel, 2025).

IV. DISCUSSION

The study of science is a broad field with a variety of methods. Academics have employed a range of perspectives to understand science's inner workings, driven by the field's diversity in researchers' disciplinary backgrounds (Sugimoto *et al.*, 2011; Liu *et al.*, 2023). In this paper we used structural causal models to highlight why causal thinking is important for the study of science, in particular for quantitative approaches. In doing so, we do not mean to suggest that we always need to estimate causal effects. Descriptive research is valuable in itself, providing context for uncharted phenomena. Likewise, studies that predict certain outcomes are very useful. However, neither descriptive nor predictive research should be interpreted causally. Both descriptive and predictive work might be able to inform discussions about possible causal mechanisms, and may provide some insight about what might be happening. However, without making causal thinking explicit, they can easily lead to wrong interpretations and conclusions. We covered several related potential issues in data analysis, such as the Table 2 fallacy (see Section III C) or the "causal salad"

approach (see Section III B).

A. The case for causal thinking

Quantitative research in science studies should make a clear distinction between prediction and causation. For example, if we observe that preregistered studies are more likely to be reproducible, we might use this information to predict which studies are more likely to be reproducible. This might be a perfectly fine predictive model. But is this also a causal effect, where preregistering a study *causes* the results to be more reproducible? Or is the observed relation between preregistration and reproducibility due to an unobserved confounding factor, such as methodological rigour? Only with an adequate causal model can we try to answer such questions.

The difference between prediction and causation becomes critical when we make policy recommendations. Should research funders mandate open data, in an attempt to improve reproducibility? Besides the problems that such a one-size-fits-all approach might have (Ross-Hellauer *et al.*, 2022), the crucial question is whether or not such an intervention would increase reproducibility. In our simulation, we have assumed that *Open data* has a moderate but positive effect on *Reproducibility*. As discussed in Section III C, naively analysing the published literature might lead one to incorrectly conclude that *Open data* is detrimental to *Reproducibility*. It is therefore imperative that policy recommendations are grounded in careful causal analysis of empirical findings to avoid serious unintended consequences.

More fundamentally, causal thinking is a useful device to connect theories to empirical analyses. Many studies in the social sciences suffer from a vague connection between their theoretical or verbal description and their empirical approach (Yarkoni, 2019). A key issue is to translate theoretically derived research questions into estimands (statements about what we aim to estimate), and subsequently, strategies for estimating those estimands (Lundberg, Johnson, and Stewart, 2021). In other words, we have to link our statistical models and estimates clearly to our theoretical concepts and research questions (McElreath, 2020). Without causal thinking, it is impossible to improve our theoretical understanding of how things work. While building increasingly rich causal diagrams is important in revealing underlying assumptions, this might also reveal deeper problems with our theoretical accounts (Nettle, 2023). Deciding on which parts of the system under study to include and which to omit (Smaldino, 2023, 318), as well as resisting the urge to add nuance on every turn (Healy, 2017), need to accompany any empirical attempt of inferring causality.

Methodologically, structural causal models do not make any assumptions about the functional relationship between variables. If identifying a certain causal effect based on a structural causal model is not possible, stronger assumptions might still allow to identify causal effects. As we have outlined in Section II, well-known causal inference techniques, such as instrumental variables, difference-in-difference, and regression discontinuity, rely on stronger assumptions, making assumptions about the functional form of the relationships (e.g. linear, or parallel trends), or about thresholds or hurdles. That is the essence of causal inference: we make assumptions to build a causal model, and use these assumptions to argue whether we can identify the causal effect given the observations we make.

Any claims of causal effects derived via causal inference will always depend on the assumptions made. Often, we cannot verify the assumptions empirically, but they might have *implications* that we can verify empirically. If we find no empirical support for these testable implications, we might need to go back to the drawing board. Finding empirical support for testable implications still does not imply that our assumptions are correct; other assumptions might have similar testable implications. Indeed, we already emphasised this in the context of the DAGs: we cannot say whether a DAG is correct, but we might be able to say whether a DAG is incorrect.

Not everything can be modelled as a structural causal model, particularly phenomena involving complex sociocultural, political, and epistemic dynamics, as well as feedback loops and specific parametric assumptions that resist representation as causal paths. There is also broader thinking and reflection of causality beyond structural models, from historical and ethnographic approaches to narrative and critical perspectives on how science develops. Nonetheless, we believe that structural causal models are a very helpful tool to communicate causal thinking among (quantitative) science study researchers, even if not the only one.

B. Going beyond—why causal thinking is useful even if causal inference is impossible

In practice, it might not always be possible to estimate a causal effect, because some variables are not observed in a study, or might even be unobservable (Rohrer, 2018). We believe that making causal thinking explicit is still highly beneficial to the broader research community in such cases. First, the process of having gone through the exercise of trying to construct a causal model is not wasted, as the model itself might be useful. Researchers might be able to build on the model in subsequent studies, and refine or revise it.

Secondly, causal models make explicit researchers' beliefs of how specific causal mechanisms work. Other researchers might disagree with those causal models. This is a feature, not a bug. By making disagreement visible, it might be possible to uncover whether the various suggested causal models have different empirically testable implications. Empirical evidence for these testable implications could then adjudicate between these various causal models. Directing research towards studying such

adjudicating empirically testable implications advances research further, and helps to build a more cumulative evidence base.

Thirdly, causal models make explicit why causal estimates might be impossible in a given study. Often, researchers state in their conclusion that there might be missing confounders and that they therefore cannot draw causal conclusions (but they may nonetheless proceed to provide advice that implicitly assumes causality). Simply stating that confounders may exist is not enough. If we, as researchers, believe that we have missed confounders, we should make explicit what we believe we missed. We can of course never be sure that we considered all relevant aspects, but that should not prevent us from trying to be as explicit as possible.

By making explicit how a causal effect is not identifiable, we might be able to suggest variables that we should try to collect in the future. Additionally, by making explicit how our estimates deviate from a causal effect, we might make informed suggestions of the direction of this deviation, e.g. whether we are under- or overestimating the causal effect. Possibly, we might even use some form of sensitivity analysis (Cinelli *et al.*, 2019) to make more informed suggestions.

The social sciences have a distinct advantage over other scientific disciplines when causal inference is challenging: we can talk to people. In case quantitative methods struggle to identify causal relationships, qualitative methods might still provide insight into causal effects, for instance because in interviews people can point out what they believe to be a causal effect. For example, suppose we are interested in the effect of *Open data* on research efficiency but struggle to quantify the causal effect. We could talk to researchers who have reused openly available datasets, asking whether and how publicly available data has helped them to conduct their research more efficiently. Responses like these might uncover causal evidence where quantitative methods encounter more difficulties.

Finally, developing explicit causal models can benefit qualitative research as well. For example, when developing an interview guide to study a particular phenomenon, it could be helpful to first develop a clear understanding of the potential causal pathways related to that phenomenon. Furthermore, even if qualitative data cannot easily quantify the precise strength of a causal relationship, it may corroborate the structure of a causal model. Ultimately, combining quantitative causal identification strategies with direct qualitative insights on mechanisms can lead to more comprehensive evidence (Munafò and Smith, 2018; Tashakkori, Johnson, and Teddlie, 2021), strengthening and validating our collective understanding of science.

ACKNOWLEDGEMENTS

We thank Ludo Waltman, Tony Ross-Hellauer, Jesper W. Schneider, Nicki Lisa Cole, and three anonymous reviewers for valuable feedback on an earlier version of the manuscript. TK used GPT-4 and Claude v2.1 to assist in language editing during the final revision stage.

AUTHOR CONTRIBUTIONS

Thomas Klebel: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing. Vincent Traag: Conceptualization, Data curation, Formal analysis, Funding acquisition, Investigation, Methodology, Project administration, Software, Validation, Visualization, Writing - original draft, and Writing - review & editing.

COMPETING INTERESTS

The authors have no competing interests.

FUNDING INFORMATION

The authors received funding from the European Union's Horizon Europe framework programme under grant agreement Nos. 101058728 and 101094817. Views and opinions expressed are however those of the authors only and do not necessarily reflect those of the European Union or the European Research Executive Agency. Neither the European Union nor the European Research Executive Agency can be held responsible for them. The Know-Center is funded within COMET—Competence Centers for Excellent Technologies—under the auspices of the Austrian Federal Ministry of Transport, Innovation and Technology, the Austrian Federal Ministry of Economy, Family and Youth and by the State of Styria. COMET is managed by the Austrian Research Promotion Agency FFG.

DATA AND CODE AVAILABILITY

All data and code, as well as a reproducible version of the manuscript, are available at (Klebel and Traag, 2024).

REFERENCES

Aagaard, K.and Schneider, J. W., "Some considerations about causes and effects in studies of performance-based research fund-

ing systems," J. Informetr. **11**, 923–926 (2017) 10.1016/j.joi.2017.05.018.

Arif, S.and MacNeil, M. A., "Applying the structural causal model framework for observational causal inference in ecology," Ecological Monographs **93**, e1554 (2023) 10.1002/ecm.1554.

Balke, A.and Pearl, J., "Bounds on treatment effects from studies with imperfect compliance," Journal of the American Statistical Association (2012).

Bol, T., de Vaan, M., and van de Rijt, A., "The matthew effect in science funding," Proceedings of the National Academy of Sciences **115**, 4887–4890 (2018) 10.1073/pnas.1719557115.

Cinelli, C., Kumor, D., Chen, B., Pearl, J., and Bareinboim, E., "Sensitivity Analysis of Linear Structural Causal Models," in *Proceedings of the 36th International Conference on Machine Learning* (PMLR, 2019) pp. 1252–1261.

Cunningham, S., *Causal Inference* (Yale University Press, 2021).

Davis, P. M., "Reanalysis of Tweeting Study Yields No Citation Benefit," (2020).

Davis, P. M., Lewenstein, B. V., Simon, D. H., Booth, J. G., and Connolly, M. J. L., "Open access publishing, article downloads, and citations: Randomised controlled trial," BMJ **337**, 343–345 (2008) 10.1136/bmj.a568.

Deffner, D., Rohrer, J. M., and McElreath, R., "A causal framework for cross-cultural generalizability," Advances in Methods and Practices in Psychological Science **5**, 25152459221106366 (2022) 10.1177/25152459221106366.

Dong, X., Xu, J., Bu, Y., Zhang, C., Ding, Y., Hu, B., and Ding, Y., "Beyond correlation: Towards matching strategy for causal inference in Information Science," Journal of Information Science **48**, 735–748 (2022) 10.1177/0165551520979868.

Elwert, F., "Graphical causal models," (Springer Netherlands, Dordrecht, 2013) pp. 245–273.

Esterling, K. M., Brady, D., and Schwitzgebel, E., "The necessity of construct and external validity for deductive causal inference," Journal of Causal Inference **13** (2025) 10.1515/jci-2024-0002.

Fecher, B.and Friesike, S., "Open science: One term, five schools of thought," (Springer International Publishing, Cham, 2014) pp. 17–47.

Gläser, J.and Laudel, G., "Governing Science: How Science Policy Shapes Research Content," European Journal of Sociology **57**, 117–168 (2016) 10.1017/S0003975616000047.

Goodman, S. N., Berlin, J., Fletcher, S. W., and Fletcher, R. H., "Manuscript Quality before and after Peer Review and Editing at Annals of Internal Medicine," Annals of Internal Medicine **121**, 11–21 (1994) 10.7326/0003-4819-121-1-199407010-00003.

Hardwicke, T. E., Bohn, M., MacDonald, K., Hembacher, E., Nuijten, M., Peloquin, B. N., deMayo, B. E., Long, B., Yoon, E. J., and Frank, M. C., "Analytic reproducibility in articles receiving open data badges at the journal psychological science: an observational study," Royal Society Open Science **8**, 201494 (2021) 10.1098/rsos.201494.

Hardwicke, T. E., Mathur, M. B., MacDonald, K., Nilsonne, G., Banks, G. C., Kidwell, M. C., Hofelich Mohr, A., Clayton, E., Yoon, E. J., Henry Tessler, M., Lenne, R. L., Altman, S., Long, B., and Frank, M. C., "Data availability, reusability, and analytic reproducibility: evaluating the impact of a mandatory open data policy at the journal *cognition*," Royal Society Open Science **5**, 180448 (2018) 10.1098/rsos.180448.

Healy, K., "Fuck nuance," Sociological Theory **35**, 118–127 (2017) 10.1177/0735275117709046.

Huntington-Klein, N., *The Effect: An Introduction to Research Design and Causality* (CRC Press, 2021).

Hünermund, P.and Bareinboim, E., "Causal Inference and Data Fusion in Econometrics," 10.48550/arXiv.1912.09104 (2023).

Imbens, G. W.and Rubin, D. B., *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction* (Cambridge University Press, Cambridge, 2015).

Klebel, T., Cole, N. L., Tsipouri, L., Kormann, E., Karasz, I., Liarti, S., Stoy, L., Traag, V., Vignetti, S., and Ross-Hellauer, T., "PathOS - D1.2 Scoping Review of Open Science Impact," (2023).

Klebel, T.and Traag, V., "Code for "introduction to causality in science studies"," 10.5281/zenodo.10639143 (2024).

Kwon, S.and Motohashi, K., "Incentive or disincentive for research data disclosure? A large-scale empirical analysis and implications for open science policy," International Journal of Information Management **60**, 102371 (2021) 10.1016/j.ijinfomgt.2021.102371.

Liu, L., Jones, B. F., Uzzi, B., and Wang, D., "Data, measurement and empirical methods in the science of science," Nature Human Behaviour **7**, 1046–1058 (2023) 10.1038/s41562-023-01562-4.

Luc, J. G. Y., Archer, M. A., Arora, R. C., Bender, E. M., Blitz, A., Cooke, D. T., Hlci, T. N., Kidane, B., Ouzounian, M., Varghese, T. K., and Antonoff, M. B., "Does Tweeting Improve Citations? One-Year Results From the TSSMN Prospective Randomized Trial," The Annals of Thoracic Surgery **111**, 296–300 (2021) 10.1016/j.athoracsur.2020.04.065.

Lundberg, I., Johnson, R., and Stewart, B. M., "What is your estimand? Defining the target quantity connects statistical evidence to theory," American Sociological Review **86**, 532–565 (2021) 10.1177/00031224211004187.

Malmgren, R. D., Ottino, J. M., and Nunes Amaral, L. A., "The role of mentorship in protégé performance," Nature **465**, 622–626 (2010) 10.1038/nature09040.

McElreath, R., *Statistical rethinking: a Bayesian course with examples in R and Stan*, 2nd ed., CRC texts in statistical science (Taylor and Francis, CRC Press, Boca Raton, 2020).

Molloy, J. C., "The Open Knowledge Foundation: Open data means better science," PLoS Biology **9**, e1001195 (2011) 10.1371/journal.pbio.1001195.

Munafò, M. R.and Smith, G. D., "Robust research needs many lines of evidence," Nature **553**, 399–401 (2018) 10.1038/d41586-018-01023-3.

Nettle, A. D., "It probably is that bad," (2023).

Nosek, B. A., Hardwicke, T. E., Moshontz, H., Allard, A., Corker, K. S., Dreber, A., Fidler, F., Hilgard, J., Kline Struhl, M., Nuijten, M., Rohrer, J. M., Romero, F., Scheel, A. M., Scherer, L. D., Schönbrodt, F. D., and Vazire, S., "Replicability, robustness, and reproducibility in psychological science," Annual Review of Psychology **73**, 719–748 (2022) 10.1146/annurev-psych-020821-114157.

Nuijten, M., Borghuis, J., Veldkamp, C. L. S., Dominguez-Alvarez, L., van Assen, M. A. L. M., and Wicherts, J. M., "Journal data sharing policies and statistical reporting inconsistencies in psychology," Collabra: Psychology **3**, 31 (2017) 10.1525/collabra.102.

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

Piwowar, H., Day, R., and Fridsma, D., "Sharing detailed research data is associated with increased citation rate," PLoS ONE **2**, e308 (2007) 10.1371/journal.pone.0000308.

Piwowar, H.and Vision, T. J., "Data reuse and the open data citation advantage," 10.7287/peerj.preprints.1v1, Tech. Rep. (2013).

Rohrer, J. M., "Thinking clearly about correlations and causation: Graphical causal models for observational data," Advances in Methods and Practices in Psychological Science **1**, 27–42 (2018) 10.1177/2515245917745629.

Rohrer, J. M., Hünermund, P., Arslan, R. C., and Elson, M., "That's a lot to process! Pitfalls of popular path models," Advances in Methods and Practices in Psychological Science **5**, 25152459221095827 (2022) 10.1177/25152459221095827.

Ross-Hellauer, T., Klebel, T., Bannach-Brown, A., Horbach, S. P. J. M., Jabeen, H., Manola, N., Metodiev, T., Papageorgiou, H., Reczko, M., Sansone, S.-A., Schneider, J., Tijdink, J., and Vergoulis, T., "TIER2: Enhancing trust, integrity and efficiency in research through next-level reproducibility," Research Ideas and Outcomes **8**, e98457 (2022) 10.3897/rio.8.e98457.

Simsek, M., de Vaan, M., and van de Rijt, A., "Do grant proposal texts matter for funding decisions? a field experiment," Scientometrics **129**, 2521–2532 (2024) 10.1007/s11192-024-04968-7.

Smaldino, P. E., *Modeling social behavior: mathematical and agent-based models of social dynamics and cultural evolution* (Princeton University Press, Princeton, 2023).

Sugimoto, C. R., Ni, C., Russell, T. G., and Bychowski, B., "Academic genealogy as an indicator of interdisciplinarity: An examination of dissertation networks in library and information science," Journal of the American Society for Information Science and Technology **62**, 1808–1828 (2011) 10.1002/asi.21568.

Tashakkori, A., Johnson, R. B., and Teddlie, C., *Foundations of mixed methods research: integrating quantitative and qualitative approaches in the social and behavioral sciences*, second edition ed. (SAGE, Los Angeles London New Delhi Singapore Washington DC Melbourne, 2021).

Tennant, J. P., Waldner, F., Jacques, D. C., Masuzzo, P., Collister, L. B., and Hartgerink, C. H. J., "The academic, economic and societal impacts of open access: an evidence-based review," 10.12688/f1000research.8460.3, Tech. Rep. (2016).

Textor, J., van der Zander, B., Gilthorpe, M. S., Liśkiewicz, M., and Ellison, G. T., "Robust causal inference using directed acyclic graphs: the R package 'dagitty'," International Journal of Epidemiology **45**, 1887–1894 (2016) 10.1093/ije/dyw341.

Tomkins, A., Zhang, M., and Heavlin, W. D., "Reviewer bias in single- versus double-blind peer review," Proc. Natl. Acad. Sci. U. S. A. **114**, 12708–12713 (2017) 10.1073/pnas.1707323114.

Traag, V. A.and Waltman, L., "Causal foundations of bias, disparity and fairness," 10.48550/arXiv.2207.13665 (2022).

Waltman, L.and van Eck, N. J., "Field Normalization of Scientometric Indicators," Springer Handbook of Science and Technology Indicators , 281–300 (2019) 10.1007/978-3-030-02511-3_11.

Westreich, D.and Greenland, S., "The table 2 fallacy: Presenting and interpreting confounder and modifier coefficients," American Journal of Epidemiology **177**, 292–298 (2013) 10.1093/aje/kws412.

Woods, H. B.and Pinfield, S., "Incentivising research data sharing: a scoping review," Wellcome Open Research **6**, 355 (2022) 10.12688/wellcomeopenres.17286.2.

Yarkoni, T., "The generalizability crisis," (2019) 10.31234/osf.io/jqw35.

Appendix A: Simulations of DAG

As explained, a DAG only specifies that a variable is affected by another variable, but it does not specify how. For simulating data, we do need to specify our DAG in Figure 1 in more detail. In particular, we sample *Field* uniformly from two fields; we sample *Rigour* and *Novelty* from standard normal distributions (i.e. with a mean of 0 and a standard deviation of 1); we sample *Open data* and *Published* from Bernoulli distributions (i.e. Yes or No); and we sample *Data reuse*, *Reproducibility* and *Citations* again from standard normal distributions. The effects of some variables on other variables are represented by simple linear equations (using a logistic specification for the Bernoulli distributions), with particular coefficients for the effects (see Table II). These distributions are not necessarily realistic. Yet, our aim is not to provide a realistic simulation, but to illustrate how causal inference can be applied. Relying on standard normal distributions and linear equations simplifies the simulation model and the analyses of the simulated data.

Table II: Coefficients for simulated data. They represent the 'true' effects for each path in the DAG, as well as basic properties for the equations (intercepts, sigma).

From	To	Coefficient
intercept	open_data	-3
rigour	open_data	0.1
field	open_data	1, 5
intercept	published	-1
novelty	published	1
rigour	published	2
open_data	published	8
intercept	data_reuse	-1
open_data	data_reuse	2
novelty	data_reuse	1
intercept	reproducibility	1
open_data	reproducibility	0.4
rigour	reproducibility	1
intercept	citations	-1
novelty	citations	2
rigour	citations	2
published	citations	2
data_reuse	citations	2
field	citations	10, 20
sigma	none	1

Appendix B: Theoretical effect of Rigour on Reproducibility

There is a direct effect of *Rigour* on *Reproducibility* and a indirect effect, mediated by *Open data*. Let X be *Rigour*, Z *Open data* and Y *Reproducibility*. We then have

$$X \sim \text{Normal}(0, 1)$$

$$Z \sim \text{Bernoulli}(\text{logistic}(\alpha_Z + \beta X + \phi_F))$$

$$Y \sim \text{Normal}(\alpha_Y + \gamma X + \theta Z, \sigma)$$

If we try to estimate a simple OLS $Y = \hat{\alpha} + \hat{\beta} X$, then

$$\hat{\beta} = \frac{\text{Cov}(X, Y)}{\text{Var(X)}}.$$

Working out $\text{Cov}(X, Y)$, we can use that $Y = \alpha_Y + \gamma X + \theta Z + \epsilon_\sigma$ where $\epsilon_\sigma \sim \text{Normal}(0, \sigma)$, and obtain that

$$\text{Cov}(X, Y) = \gamma \text{Cov}(X, X) + \theta \text{Cov}(X, Z) + \text{Cov}(X, \epsilon_\sigma),$$

where $\text{Cov}(X, X) = \text{Var}(X, X) = 1^2 = 1$ and $\text{Cov}(X, \epsilon_\sigma) = 0$, because ϵ_σ is independent of X. Hence, we obtain

$$\text{Cov}(X, Y) = \gamma + \theta \text{Cov}(X, Z).$$

Writing out $\text{Cov}(X, Z)$, we find that $\text{Cov}(X, Z) = E(XZ)$ because $E(X) = 0$. Then elaborating $E(XZ) = E(E(XZ|F))$, we can expand $E(XZ|F)$ as a sum

$$E(XZ|F) = \int_x \sum_{z=0}^{1} xz P(Z = z \mid X = x, F) P(X = x) \mathrm{d}x$$

Obviously, $xz = 0$ when $z = 0$, while $xz = x$ when $z = 1$. Hence, this simplifies to only the $z = 1$ part, such that

$$E(XZ|F) = \int_x x P(Z = 1 \mid X = x, F) P(X = x) \mathrm{d}x$$

or

$$E(XZ \mid F) = \int_x x \cdot \text{logistic}(\alpha_Z + \beta x + \phi_F) \cdot f(x) \mathrm{d}x,$$

where $f(x)$ is the pdf of $X \sim \text{Normal}(0, 1)$. Unfortunately, this does not seem to have an analytical solution, so we numerically integrate this.

The total causal effect of *Rigour* on *Reproducibility* is very close to the direct causal effect of *Rigour* on *Reproducibility* (which is 1), because the indirect effect via *Rigour* $\rightarrow$ *Open data* is small.

Appendix C: Theoretical effect of Open data on citations

There are two causal paths of the effect of *Open data* on *Citations*. The first causal path is mediated by *Data reuse* and the second is mediated by *Published*. Let X be *Open data*, Y be *Citations*, D be *Data reuse* and P be *Published*. Since we use a normal distribution for *Citations* we can simply write

$$E(Y) = \alpha + \beta_{DY} D + \beta_{PY} P + \beta_{\text{novelty,Y}} \cdot \textit{Novelty} + \beta_{\text{rigour,Y}} \cdot \textit{Rigour} + \beta_{\text{field,Y}} \cdot \textit{Field},$$

where we can consider *Field* a dummy variable, representing the effect of field 2 relative to field 1 (i.e. field 1 is the reference category).

The change in Y, i.e. ΔY, relative to changing X, i.e. ΔX, from 0 to 1 is then

$$\frac{\Delta Y(X)}{\Delta X} = \beta_{DY} \frac{\Delta D(X)}{\Delta X} + \beta_{PY} \frac{\Delta P(X)}{\Delta X}$$

The first part is simple, since D is a normal distribution, yielding $\frac{\Delta D(X)}{\Delta X} = \beta_{XD}$. The second part is more convoluted, since P is a logistic distribution of a normal variable. For that reason, we calculate $\frac{\Delta P(X)}{\Delta X}$ numerically using `logitnorm::momentsLogitnorm` (version 0.8.38) in R.

Appendix D: Validation of argument against stepwise regression

In Section III B, we claimed that stepwise regression to identify the effect of *Open data* on *Citations* would not work. In particular, it suggests to include the mediating variables *Data reuse* and *Published* and to remove *Open data* from the regression model. The output below demonstrates this behaviour.

We first start with a full model that includes all variables.

```
full_model <- lm(citations ~ ., data = df)
```

Next, we let R select variables in a stepwise fashion, considering both directions (including or excluding variables) at each step.

```
step_model <- MASS::stepAIC(full_model, direction = "both", trace = TRUE)
```

```
Start:  AIC=14.39
citations ~ rigour + novelty + field + open_data + published +
    data_reuse + reproducibility

                  Df Sum of Sq     RSS     AIC
- reproducibility  1       0.0   998.4   12.42
- open_data        1       0.2   998.6   12.59
<none>                         998.4   14.39
- published        1     407.8  1406.2  354.90
- rigour           1    1683.9  2682.2 1000.65
- novelty          1    1946.4  2944.7 1094.02
- data_reuse       1    3620.3  4618.7 1544.10
- field            1   10158.0 11156.4 2426.01

Step:  AIC=12.42
citations ~ rigour + novelty + field + open_data + published +
    data_reuse

                  Df Sum of Sq     RSS     AIC
- open_data        1       0.2   998.6   10.61
<none>                         998.4   12.42
+ reproducibility  1       0.0   998.4   14.39
- published        1     407.8  1406.3  352.94
- novelty          1    1946.4  2944.9 1092.06
- rigour           1    3199.3  4197.7 1446.55
- data_reuse       1    3623.4  4621.8 1542.79
- field            1   10171.5 11170.0 2425.23

Step:  AIC=10.61
citations ~ rigour + novelty + field + published + data_reuse

                  Df Sum of Sq     RSS     AIC
<none>                         998.6   10.61
+ open_data        1       0.2   998.4   12.42
+ reproducibility  1       0.0   998.6   12.59
- published        1     505.6  1504.2  418.26
- novelty          1    2394.6  3393.2 1231.78
- rigour           1    3281.9  4280.5 1464.06
- data_reuse       1    4743.2  5741.9 1757.78
- field            1   14922.0 15920.6 2777.61
```

We can see that the algorithm first removes *Open data*, and then *Reproducibility*. The final model is then as follows:

```
summary(step_model)
```

```
Call:
lm(formula = citations ~ rigour + novelty + field + published +
    data_reuse, data = df)

Residuals:
     Min       1Q   Median       3Q      Max
-3.08249 -0.68542 -0.01525  0.70217  3.02677

Coefficients:
             Estimate Std. Error t value Pr(>|t|)
(Intercept)   -1.26727    0.11955  -10.60   <2e-16 ***
rigour         1.92751    0.03372   57.16   <2e-16 ***
novelty        2.01798    0.04133   48.82   <2e-16 ***
field         10.11475    0.08299  121.87   <2e-16 ***
publishedTRUE  2.06206    0.09192   22.43   <2e-16 ***
data_reuse     1.95682    0.02848   68.71   <2e-16 ***
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 1.002 on 994 degrees of freedom
Multiple R-squared:  0.9878,    Adjusted R-squared:  0.9878
F-statistic: 1.616e+04 on 5 and 994 DF,  p-value: < 2.2e-16
```

Appendix E: The case against causal salad

Table III illustrates the result of the 'causal salad' approach of including all variables. Because this model controls for mediators, the effect of *Open data* on *Citations* appears to be zero. The researcher could thus be led to conclude that *Open data* has no effect on *Citations*, which is incorrect.

Table III. Example of "causal salad". The "Correct model" to estimate the causal effect of *Open data* on *Citations* identifies the effect to be 5.29. If the researcher were to include all variables, it might seem as if there was no effect of *Open data* on *Citations*. Values in brackets show p-values.

	Correct model	'Causal salad' model
Intercept	−2.519	−1.295
	(<0.001)	(<0.001)
Open data	5.294	−0.061
	(<0.001)	(0.655)
Field	10.213	10.140
	(<0.001)	(<0.001)
Rigour	2.355	1.919
	(<0.001)	(<0.001)
Novelty		2.010
		(<0.001)
Data reuse		1.964
		(<0.001)
Published		2.082
		(<0.001)
Reproducibility		0.006
		(0.849)
Num.Obs.	1000	1000
R2	0.726	0.988