Qualitative approximations to causality: non-randomizable factors in clinical psychology

Running title: Qualitative approximations to causality

Reviews on the new methodical toolbox for causal analysis of observational data usually focus on quantitative methods, although any decision on using such a formal method requires profound and context-dependent qualification. Such qualitative approaches may yet sometimes suffice, and the quantitative methods have hidden potential in qualitative use.

Michael Höfler[1], Sebastian Trautmann[2], Philipp Kanske[1,3]

[1]: Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, Dresden, Germany, [2]: Department of Psychology, Medical School, Hamburg, Germany; [3]: Max Planck Institute for Human Cognitive and Brain Sciences, Leipzig, Germany

Corresponding author: Michael Höfler, Chemnitzer Straße 46, Clinical Psychology and Behavioural Neuroscience, Institute of Clinical Psychology and Psychotherapy, Technische Universität Dresden, 01187 Dresden, Germany. michael.hoefler@tu-dresden.de, +49 351 463 36921

Submission to Clinical Psychology in Europe

Abstract

Background: Causal quests in non-randomized studies are unavoidable just because research questions are beyond doubt causal (e.g. aetiology). Large progress during the last decades has enriched the methodical toolbox.

Aims: Summary papers mainly focus on quantitative and highly formal methods. With examples from clinical psychology, we show how qualitative approaches can inform on the necessity and feasibility of quantitative analysis and may yet sometimes approximate causal answers.

Results: Qualitative use is hidden in some quantitative methods. For instance, it may yet suffice to know the direction of bias for a tentative causal conclusion. *Counterfactuals* clarify what causal effects of changeable factors are, unravel what is required for a causal answer, but do not cover immutable causes like gender. *Directed acyclic graphs (DAGs)* address causal effects in a broader sense, may give rise to quantitative estimation or indicate that this is premature.

Conclusion:

No method is generally sufficient or necessary. Any causal analysis must ground on qualification and should balance the harms of a false positive and a false negative conclusion in a specific context.

Keywords: Causality, causal considerations, counterfactuals, directed acyclic graphs

Highlights

- Causal inference outside randomized, controlled experiments and trials is rare in clinical psychology, regardless of the rich methodology that has evolved in the last decades.
- The attractiveness of these new formal tools distracts from their limits and expenditure, but considerable benefit is hidden in their qualitative use.
- Qualitative considerations may suffice to approximate causal answers.

Causal questions drive most scientific reasoning. This should entail plenty of causal analyses, but clinical psychology often avoids causality because the established gold standard, a randomized controlled experiment or trial (RCT), is in many cases infeasible. Although we cannot or should not manipulate variables such as gender, traumatic events, personality traits and other constructs, their effects on clinical outcomes must be investigated to inform prevention, intervention, policies, theories and further research.

The specific problem of causality in observational studies

The methodological toolbox has been greatly expanded. It now offers approaches to causal answers in non-randomized studies (Greenland, 2017). These new tools mainly address the *specific* problem of causality: Without randomization, a binary factor **X** (group comparison, e.g. with and without a bipolar disorder diagnosis) and outcome **Y** (e.g. amount of substance use) often have *shared causes*, **Z** (e.g. parental mental health), that are out of experimental control and cause bias in an estimate of the average effect of **X** on **Y.** In linear models and for just a single **Z**, this bias is the product of the effect of **Z** on **X** and Y, meaning that it equals α_1

$* \alpha_2$, where α_1 denotes the effect of **Z** on **X**, and α_2 the effect of **Z** on **Y** (e.g. Gelman & Hill, 2007, chapter 9). This simple formula implies that

a. bias occurs only if $\alpha_1 \neq 0$ *and* $\alpha_2 \neq 0$
b. the direction of bias just depends on the *signs* of α_1 and α_2. If they are equal, bias is upward, otherwise downward.
c. bias is small if *either* is small

These properties generalize to non-linear relations and any distributions of **Y** and **Z** and to multiple **Z** that are independent or positively inter-related (Groenwold et al., 2018; Pearl's "adjustment formula" is the most general expression; Pearl, 2009). We refer to the above as the *basic confounding relation*.

Experimental control and randomization together disconnect all *confounders* **Z** from **X** and thus eliminate *confounding bias*. Otherwise, **X** is just *observed*, and in life-sciences like clinical psychology the number of natural causes of an **X** might be vast. The new methodical tools try to unravel the **X-Y** relation in an imaginary world in which **X** (or **Y**) was independent of **Z** and thus simulate what *changing* (rather than observing) **X** would do with **Y** ("do(**X**)", Pearl, 2009). The new methods mimic what might be observed if **X** were changed, but unlike real-world change experiments where **X** is isolated, their use requires an explicit understanding of the relationships between variables **Z** and **X**. Likewise, during their elaboration it has been stressed that one must consider *how* an **X** is to be changed because this may make a large difference (Greenland, 2005a). For example, just stopping drug use might even worsen an outcome if intervention does not address factors like stress coping, a putative cause of drug use. In this sense, the new methods complement randomized experiments and RCTs through the more explicit need to go beyond a single **X**,

thus to move from “causal description” to “causal explanation” (Johnson et al., 2019). For other (non-specific) sources of bias like selection and measurement error that also effect the results of randomized studies, see the supplemental material [see below].

Instead of making use of the new methodological toolbox to approach causal answers in observational studies, clinical psychology was dominated by the “mantra“ that “correlation is not causation“ (Pearl & MacKenzie, 2018, back of the book). For a historical account on how this stance has emerged through the statistical pioneer Karl Pearson, who had considered causality to equal perfect (deterministic) correlation, see Pearl and MacKenzie (2018).

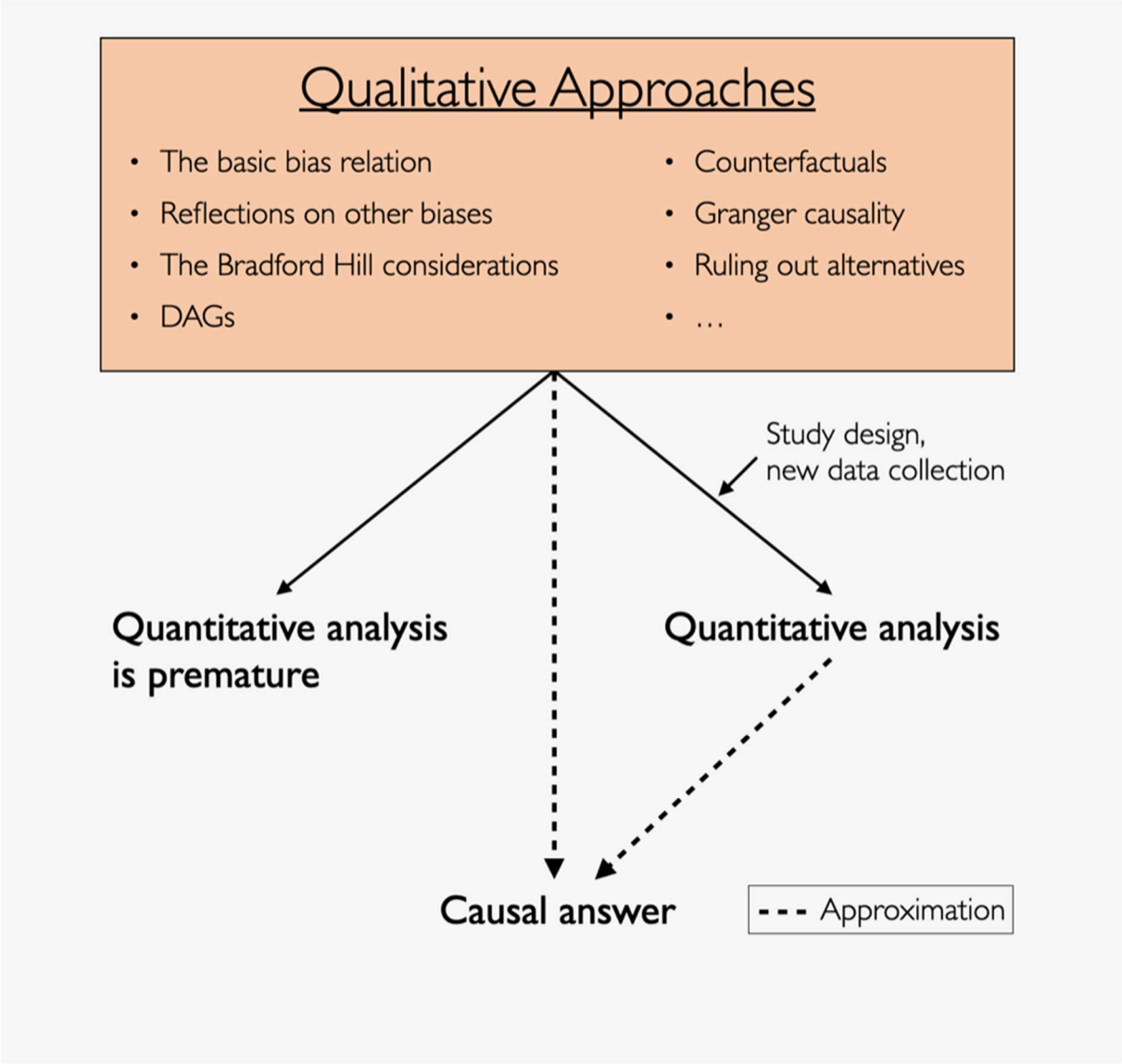
Aim of this paper

Some papers have already introduced tools from the new methodical box in (clinical) psychology and summarized the meanwhile vast literature on them (Dablander, 2020; Marinescu et al., 2018). However, these have mainly focussed on quantitative approaches in a discipline where methodical causal thinking is new and, thus, requires qualitative guidance beforehand. One such instance is that psychology needs not only to overcome “retreating into the associational haven” (Hernán, 2005), but also immunization against overconfidence (Greenland, 2012) in novel methods. Overconfidence mainly concerns the quantitative and highly formal methods, because the mathematical sophistication in these easily obstructs the sight for hidden assumptions and over-simplification through translation into mathematics (Greenland, 2012, 2017; VanderWeele, 2016). Costs of using these methods also include learning and conducting them (which is error-prone) and the further degrees of freedom in analysis through their use which promotes p-hacking. We argue that qualitative approaches as exemplified in this article are easier to access and invite more debate and refinement on them

and should at least inform the decision of using a particular quantitative method. We focus on a few causal conceptions that we believe are most illustrative for causal quests: the above basic confounding relation (1), counterfactuals (2), popular qualitative considerations (3) and directed acyclic graphs (4).

The following figure illustrates the scheme by which we describe how qualitative approaches may guide a causal quest.

---------- INSERT FIGURE 1 ABOUT HERE ----------



Qualitative approaches

Gender effects and the basic bias relation

The effects of gender (biological sex) may play an important role for the development and maintenance of mental disorders. If they exist to considerable extent, they contribute to explaining the different aetiology of disorders that are more prevalent in females (e.g. internalizing disorders such as depression) and males (e.g. externalizing disorders such as substance use disorders). This is because gender may also affect many putative aetiological factors (e.g. response styles such as rumination (Johnson & Whisman, 2013), which, in turn, may influence the onset of disorders (Emsley & Dunn, 2012)).

But is the causal wording "effect" warranted here? With the basic bias relation, we are equipped to ask: Are there shared causes of gender and a disorder **Y**? If it holds true that gender is largely random in the sense that it depends only on factors that do not also affect the disorder (Scarpa, 2016, and references therein), then no confounding bias is expected. If such factors exist (e.g. environmental pollution; Astolfi & Zonta, 1999) but affect **Y** only weakly, they may be neglected since the bias through them should be small. If bias from other sources is also negligible like selection and measurement, a causal conclusion seems informed.

Upward bias through confounders that affect X and Y with the same sign

In the presence of reliable associational results, the basic bias relation can be applied well beyond gender effects. If there is at most a *weak* association between an **X** and a **Y**, and assuming that the common causes of **X** and **Y** affect both positively or both negatively (and

are unrelated or positively inter-related), bias should be *upward*. Hence, the effect of **X** on **Y** should be smaller than the association and, thus, be absolutely small (and probably negligible). For example, the relatively weak and often inconsistently reported association between anxiety and alcohol use might be explained by genetic and personality factors increasing the risk for both (Schmidt et al., 2007). Such risk increasing may frequently apply: psychopathology in parents, genetic factors, stable personality traits, stressful life events and prior mental disorders are factors that might all affect disorders positively and be positively inter-related (Uher & Zwicker, 2017). However, with a larger number of shared factors, the probability rises that some have negative relations, but if these are few and unlikely to dominate bias (because their effects on **X** *and* **Y** are not very large as compared to those of the other factors), a researcher may still use the consideration.

Counterfactuals and a defendable assumption on them

The above gender example brings up an important limitation yet in the standard "counterfactual" definition of a causal effect. Biological sex cannot be entirely changed (beyond transsexual transformation) or imagined to be changed, but social aspects of gender can (Glymour & Glymour, 2014).

Imagining a person under an alternative **X** condition is called *counterfactual* and defines an effect as the amount of change in **Y** if **X** is changed from one value to another (if this equals zero, there is no effect). Consider the putative effect of childhood trauma (CT) on depression (DE). Yet the idea of counterfactuals points out that "the effect" is imprecise since there are actually two counterfactuals and associated effects: a) trauma *experience* in individuals who actually do *not* experience trauma and b) trauma *recovery* in those who actually had experienced a trauma (but do not recover). Just referring to "the effect" denotes the *total*

effect, which means that we imagine *both* changes at once (Pearl, 2009). Such a summary appears pointless in clinical psychology, at least if one aims to keep aetiology and persistence/maintenance apart which seems important since in many cases, different factors seem to be involved in the onset versus the persistence of mental disorders (McLaughlin et al., 2011).

The effect of *experiencing* a CT is, in principle, subject to a prevention RCT, but such studies would be highly ineffective. This is because CT prevention will never succeed among all individuals and is unethical if the control group is deliberately exposed to CT although exposure (and associated harm) could have been prevented. The effect of *recovery* from a trauma on the other hand; i.e., of successful intervention, can in principle be investigated in an RCT, but only with regard to specific *consequences* of CT. This not only heavily depends on what is meant with "consequences" (e.g. distress, symptom onset, incidence of a diagnosis) and the mode of intervention, it is confounded with the aim of investigating the recovery effect (Greenland, 2005a).

At least for onset, "target trials" (here prevention trials) may be an effective further tool to clarify what a counterfactual specifically means (VanderWeele, 2016). A *target trial* is an ideal trial (or experiment) the data of which would provide the desired causal answer. It clarifies qualitatively what we *would* require, what we cannot do, but what we can anyway *imagine* (Lewis, 1973; Pearl, 2013), including the target population to infer on.

For a conclusion on the *existence* of either effect crude estimates of counterfactual depression rates (generally mean outcomes) among those with and without CT, respectively, are necessary. If we know empirically that, say, 5% of those without CT have depression later in life, and we assume that the experience of CT in all the observed individuals would have increased this rate (i.e. the counterfactual rate is > 5%; probably few clinical psychologists

would doubt this), the conclusion that *CT experience increases the risk for depression* is valid. Likewise if, say, 10% of those with CT have depression later on, we may conclude that an intervention *decreases* the rate provided that we are willing to assume that the intervention would achieve a rate below 10%.

This line of qualitative argument determines the „target quantity" (Petersen & Van der Laan, 2014) one wishes to estimate. It may also trigger other considerations like *substituting* unknown counterfactual depression rates from other, „analogous" (Hill, 1965) studies. For trauma experience, a sample of children traumatized by war may be used and for recovery, a sample of traumatized, untreated but resilient children.

Granger causality

Imagining counterfactual states of brains in Neuroscience and Neuroimaging research seems meaningful, but in associated longitudinal studies there is a shortcut to the specific causal problem of common causes hidden in the term „Granger causality" (Friston et al., 2013). Originally, the term states that, given „all the information in the universe up to time t" (Eichler & Didelez, 2010), and provided that the prediction of **Y** at time t+1 is worse if an **X** at any time up to t is disregarded, then this prior **X** is a cause of **Y** (Granger, 1969). Although equivalent with the counterfactual definition, Granger causality has been frequently mistaken as only referring to *observed* **X** variables (Eichler, 2012; Eichler & Didelez, 2010) or even just a time-series of a single **X** (Marinescu et al., 2018). This downgrades the conception into a heuristic for practical use with the easily wrong qualitative suggestion that adjustment for common causes has been sufficient. Researchers who use it must be aware of the basic bias relation indicating that they play into their own hands if they ignore unobserved common causes that effect **X** and **Y** with the same sign. These may include variables that have occurred

before study onset. Generally, collecting *big data* like thousands of voxels in a brain scan is no substitute for thoughtful reflections on the processes beyond the data that any defendable causal analysis relies on (Pearl & MacKenzie, 2018).

In the supplement we briefly discuss other popular and, mostly long-used approaches: multimethod evidence, mixed methods research and ruling out alternatives [see below].

Directed acyclic graphs (DAGs)

So far, we have only addressed direction of bias but not when and how bias can be removed. In the supplement, we revisit the example of the effect of CT on DE to outline the qualitative answers that the qualitative method of DAGs provides, including the subsequent study design and analysis that a particular DAG model may give rise to. The example uses a model with four common causes and causal relations among them. It reveals that adjustment for them is possible in subsequent quantitative analysis (whereby one shared cause does not require adjustment [see below]).

Importantly, DAGs may include effects of unchangeable factors like “socio-economical family status” in the example where the counterfactual conception of an effect does not apply. The conception, however, may be extended to include other actors than humans who could change an **X** (Bollen & Pearl, 2013). Sometimes such an actor is difficult to name let alone to translate into a mathematical model, wherefore instances like “socio-economical family status” are more suited “to describe something as a cause“ than to “reasonably define a quantitative causal effect estimand“ (VanderWeele, 2016).

Qualitative assumptions may make quantitative approaches seem premature

In contrary to the above instance, a DAG might reveal that bias can *not* be fully eliminated, or leave open whether an adjustment decreases or increases bias (Morgan & Winship, 2014, chapter 3). The practical utility of DAGs for quantitative analysis rises with fewer variables in them and the number of causal relations that can be assumed not to exist (Greenland, 2017). However, setting up a DAG model should reveal this. Per se, a DAG renders all associated assumptions transparent and invites for debate and refinement on them (the reader might ask herself if this happens with the figure in the supplement).

Anyway, controversy on a model might be so large that grounding a study and quantitative analysis on it appears unwarranted (Petersen & Van der Laan, 2014). Also, if the number of potential common causes is large and there is no way to prioritize them for reducing bias, quantitative analysis seems premature. Instead, more research is required beforehand to set up a defendable DAG. An example is the effect of internalizing symptoms on substance use where common causes may include a variety of genetic, parental, childhood, personality and environmental factors, as well as all sorts of individual variables related to neurobiological, cognitive and emotional processes (Pasche, 2012).

Conclusions

No method can fully cover all aspects of causality across research fields and specific applications, especially in a life science as complex as clinical psychology (Greenland, 2017), and “there is no universal method of scientific inference” (Gigerenzer & Marewski, 2014). Likewise, a causal query can never be fully objective, because it always involves assumptions beyond the data (Greenland, 2005b). In sharp contrast, researchers tend to “mechanizing

scientists' inferences" (Gigerenzer & Marewski, 2014) and downgrade methods from tools for thoughtful cooperation between methodologists and substantive experts (Höfler et al., 2018) into empty rituals (Gigerenzer, 2018).

In this article, we have outlined some qualitative approaches through which one may approach a crude causal answer on an average effect, plan a quantitative analysis or unravel that any analysis is currently infeasible. In fact, any causal quest must start with qualification because otherwise it would be just a mechanical exercise. The qualitative conceptions outlined here are meant as provisory heuristics that must not be ritualized but should be taken as invitations for refinement and adjustment to any particular application.

Above all, the two possible errors in causal conclusions should guide causal quests and the decision on whether the use of a highly formal method pays off (Greenland, 2012): false positive and false negative. Statistical decision theory provides the framework to formalize the balance between false positive and false negative causal conclusions. It states that the better decision is the one with the lower expected costs (Dawid, 2012).

Thoughtful causal quests are essential for explaining why phenomena occur the way they do and in providing levers through which things could be changed, for instance, in preventing disorders and improving life. Assessing causality is complex, demanding and ambivalent, but so is science. However, it makes use of the natural capacity of causal modelling which is deeply grounded in us human beings and structures how we view the world (Pearl & MacKenzie, 2018).

Legend to figure 1: Scheme of qualitative approaches guiding causal quests. These might be sufficient for overall causal answers, give rise to designing a new study and/or quantitative

analysis, or suggest that such analysis is premature. The basic bias relation, counterfactuals and DAGs belong to the new toolbox of causal methods.

Funding: The authors have no funding to report.

Competing Interests: The authors have declared that no competing interests exist.

Acknowledgments: We wish to thank Konrad Lehmann for the layout of the figure.

References

Astolfi, P., & Zonta, L. A. (1999). Reduced male births in major Italian cities. *Human Reproduction, 14*(12), 3116-3119. S

Bollen, K. A., & Pearl, J. (2013). Eight Myths about Causality and Structural Equation Models. In: S. L. Morgan (ed). Handbook of Causal Analysis for Social Research (pp. 301–28). New York, NY: Springer

Dablander, F. (2020). An Introduction to Causal Inference. *PsyArXiv*. February 13. d https://doi:10.31234/osf.io/b3fkw

Dawid, P. (2012). The Decision Theoretic Approach to Causal Inference. In: C. Berzuini, P. Dawid, L. Bernardinelli (eds). *Causality: Statistical Perspectives and Applications* (pp. 25-42). New York, NY, USA: Hoboken.: Wiley.

Eichler, M., & Didelez, V. (2010). On Granger causality and the effect of interventions in time series. *Lifetime Data Analysis, 16*(1), 3–32. https://doi: 10.1007/s10985-009-9143-3

Eichler, M. (2012). Causal inference in time series analysis. In: C. Berzuini, P. Dawid, L. Bernardinelli (eds). *Causality: Statistical Perspectives and Applications* (pp. 327-354). New York, NY, USA: Hoboken.: Wiley.

Emsley, R., & Dunn, G. (2012). Evaluation of potential mediators in randomized trials of complex interventions (psychotherapies). In: C. Berzuini, P. Dawid, L. Bernardinelli (eds). *Causality: Statistical Perspectives and Applications* (pp. 290-309). New York, NY, USA: Hoboken.: Wiley.

Friston, K., Moran, R., & Seth, A. K. (2013). Analysing connectivity with Granger causality and dynamic causal modelling. *Current Opinion in Neurobiology, 23*(2), 172–178. https://doi.org/10.1016/j.conb.2012.11.010

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

Glymour, C., & Glymour, M. (2014). Commentary: Race and Sex Are Causes. *Epidemiology*, 25(4), 488–490. https://doi.org/10.1177/1077559506289524

Gigerenzer, G. (2018). Statistical Rituals: The Replication Delusion and How We Got There. *Advances in Methods and Practices in Psychological Science, 1*(2), 198–218. https://doi.org/10.1177/2515245918771329

Gigerenzer, G., & Marewski, J. N. (2014). Surrogate Science: The Idol of a Universal Method for Scientific Inference. *Journal of Management, 41*(2), 421-440. https://doi.org/10.1177/0149206314547522

Granger, C. W. J. (1969). Investigating Causal Relations by Econometric Models and Cross-spectral Methods. *Econometrica, 37*(3), 424–438. http://links.jstor.org/sici?sici=0012-9682%28196908%2937%3A3%3C424%3AICRBEM%3E2.0.CO%3B2-

Lhttp://links.jstor.org/sici?sici=0012-9682%28196908%2937%3A3%3C424%3AICRBEM%3E2.0.CO%3B2-L

Greenland, S. (2005a). Epidemiologic measures and policy formulation: lessons from potential outcomes. *Emerging themes in Epidemiology*, *2*, 5. https://doi.org/10.1186/1742-7622-2-5

Greenland, S. (2005b). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society A, 168*(2), 267–291. https://doi.org/10.1111/j.1467-985X.2004.00349.x

Greenland, S. (2012). Causal inference as a prediction problem: Assumptions, identification and evidence synthesis. In: C. Berzuini, P. Dawid, L. Bernardinelli (eds). *Causality: Statistical Perspectives and Applications* (pp. 43-58). New York, NY, USA: Hoboken.: Wiley.

Greenland, S. (2017). For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. *European Journal of Epidemiology, 32*(1), 3-20. http://doi:10.1007/s10654-017-0230-6

Groenwold, R. H. H., Shofty, I., Miočević, M., et al. (2018). Adjustment for unmeasured confounding through informative priors for the confounder-outcome relation. *BMC Medical Research Methodology, 8*(1). https://doi:10.1186/s12874-018-0634-3

Hernán, M. A. (2005). Invited commentary: hypothetical interventions to define causal effects – afterthought or prerequisite? *American Journal of Epidemiology, 162*(7), 618–20. https://doi:10.1093/aje/kwi255

Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. https://doi.org/10.1177/0141076814562718

Höfler, M., Venz, J., Trautmann, S., & Miller, R. (2018). Writing a discussion section: how to integrate substantive and statistical expertise. *BMC Medical Research Methodology, 18*, 34. https://doi:10.1186/s12874-018-0490-1

Johnson, D. P., & Whisman, M.A. (2013). Gender differences in rumination: A meta-analysis. *Personality and Individual Differences*, 55(4), 367–374. https://doi:10.1016/j.paid.2013.03.019

Johnson, R.B., Russo, F., Schoonenboom, J. Causation in Mixed Methods Research: The Meeting of Philosophy, Science, and Practice. *Journal of Mixed Methods Research*. 2019;13(2):143-162.

Lewis, D. (1973). Counterfactuals and comparative probability. *Journal of Philosophical Logic, 2*(*4*), 418–446. *(Reprinted (1981) in W. L.* Harper, R. Stalnaker, *& G. Pearce (Eds.), Ifs (pp. 57–85). Dordrecht, The Netherlands: D. Reidel.)*

McLaughlin, K.A., Breslau J., Green J.G., et al. (2011). Childhood socio-economic status and the onset, persistence, and severity of DSM-IV mental disorders in a US national sample. *Social Science & Medicine, 73*(7), 1088-1096. https://doi:10.1016/j.socscimed.2011.06.011

Marinescu, I. E., Lawlor P. N., & Kording K. P. (2018). Quasi-experimental causality in neuroscience and behavioural research. *Nature Human Behaviour 2*(12), 891-898. https://doi:10.1038/s41562-018-0466-5

Morgan, S. L., & Winship, C. H. (2014). *Counterfactuals and Causal Inference. Methods and Principles for Social Research.* 2nd edition. Cambridge, UK: Cambridge University Press.

Pasche, S. (2012). Exploring the Comorbidity of Anxiety and Substance Use Disorders. *Current Psychiatry Report, 14*(3):176-81. https://doi:10.1007/s11920-012-0264-0

Pearl, J. (2009). *Causality, models, reasoning and infere*nce. 2nd edition. Cambridge, UK: Cambridge University Press.

Pearl, J. (2013). Structural Counterfactuals: A Brief Introduction. *Cognitive Science 37*(6), 977–985. https://doi:10.1111/cogs.12065

Pearl, J., & MacKenzie, D. (2018). *The Book of Why: the new Science of Cause and Effect*. New York, NY: Basic Books.

Petersen, M.L., & Van der Laan, M.J. (2014). Causal Models and Learning from Data: Integrating Causal Modeling and Statistical Estimation. *Epidemiology 25*(3), 418–426. https://doi: 10.1097/EDE.0000000000000078

Scarpa, B. (2016). Bayesian inference on predictors of sex of the baby. *Frontiers in Public Health, 4*, 102. https://doi:10.3389/fpubh.2016.00102

Schmidt, N. B., Buckner, J. D., & Keough, M. E. (2007). Anxiety sensitivity as a prospective predictor of alcohol use disorders. *Behavior Modification, 31*(2), 202-19. https://doi:10.1177/0145445506297019

Uher, R., & Zwicker A. (2017). Etiology in psychiatry: embracing the reality of poly-gene-environmental causation of mental illness. *World Psychiatry*, 16(2), 121–129. https://doi:10.1002/wps.20436

VanderWeele, T. J. (2016). Commentary: On Causes, Causal Inference, and Potential Outcomes. *International Journal of Epidemiology*, *45*(6), 1809-1816. https://doi:org/10.1093/ije/dyw230

Supplementary material to the article "Qualitative approximations to causality: non-randomizable factors in clinical psychology"

Michael Höfler, Sebastian Trautmann, Philipp Kanske

This online supplement provides some additions to the paper, namely other sources of bias than confounding, and other popular approaches to causality besides those from the new roolbox and Granger causality. We also elaborate on the example of the effect of childhood trauma (factor **X** = CT) on depression (outcome **Y** = DE) using a DAG (directed acyclic graph) model on common causes of this effect and subsequent study design and data analysis the model gives rise to.

1. Other sources of bias that also affect randomized studies

Randomized experiments and RCTs are also prone to other data-generating mechanisms that might distort causal analysis (Hernán et al., 2008). Such mechanisms, however, are *not specific* for causal quests but also bias estimates of associations. They include selection, measurement error in **Y** and non-compliance with the compared **X** conditions. Observational studies may yield similar results especially if assignment to **X** is largely independent of **Y** (Rosenbaum, 2010; Shadish et al., 2002), or if common causes are adjusted for (Anglemyer et al., 2014). Total bias might be even larger in randomized studies; for instance, if selection bias (in randomized clinical studies) dominates confounding bias (in non-randomized general population studies, Greenland, 2005, 2012; Lash et al., 2009; Mansournia et al., 2017).

Consider *selection bias* in clinical studies. This occurs, for instance, if treatment seeking is related with the treatment effect of interest. It generally happens if whatever **X** and **Y** are both related with being selected into a study (through helpseeking). This mechanism became famous as "Berkson's bias" for the potential to create fully spurious comorbidity and is nowadays (within the framework of the below described DAGs) referred to as "conditioning on a collider" (Elwert & Winship, 2014; Höfler & Trautmann, 2019). For example, an association between years of education and severity of depression is expected in a clinical sample just because both higher education and disorder severity are associated with seeking treatment (Magaard et al., 2017).

2. Other qualitative methods

Multi-method evidence

A good advice is not to draw causal conclusions from just a single study, especially if studies with different methods (e.g. cell and animal studies, neuroimaging, self-reports) are available (Greenland, 2017) with *differently* biased estimates, because the study designs open the door for sources of bias to stream in differently (confounding, selection, measurement, non-compliance, etc.; Greenland, 2012). Then *convergent* evidence for an effect can hardly be explained by *common* bias (bias that equally occurs in all kind of studies; Campbell & Fiske, 1959). However, different biases could distort the estimates in the same direction. For instance, placebo treatment effects have been argued to be upward biased both due to selection into a clinical sample and reporting **Y** (Hróbjartsson et al., 2011). Thus, studies in both clinical (biased through both) and general populations (biased only through reporting) might be inflated, in which case the common bias argument fails.

On the other hand, the argument can be strengthened through adding some of the nine Bradford Hill considerations that Hill had once formulated inspired by the historical controversy on the effect of smoking on lung cancer (Hill, 1965). Such are: "strength of association" (as above) and "plausibility": there is a (biological) model that explains an effect. The latter evaluates a (maybe quantitative) finding or putative conclusion through the ability to link it to substantive knowledge. However, which finding and which knowledge weighs how much is strongly context-dependent (Höfler, 2005), and different integrations of "ragged evidence" might appear plausible but yield different conclusions (Greenland, 2012).

Mixed methods research

Recently, similar proposals have been made under the term of "mixed methods research". "Mixed methods" call for study type "pluralism" and urges researchers to reflect on different levels (persons, populations, time, situations, factors and outcomes) where causal effects may occur equally or differently (Johnson et al., 2019, and references therein).

Ruling out alternatives

Another popular instance of considering more background knowledge than addressed in an analysis is ruling out alternatives to **X** causing **Y** (Greenhouse, 2009). Statistically, an **X-Y** association may be explained, among others, by shared causes, shared measurement error, selection bias, and inverse causation (**Y** causing **X**) (Maclure & Schneeweiss, 2001; Pearl, 2009). (See the appendix for an example on selection bias in clinical studies [insert link]).

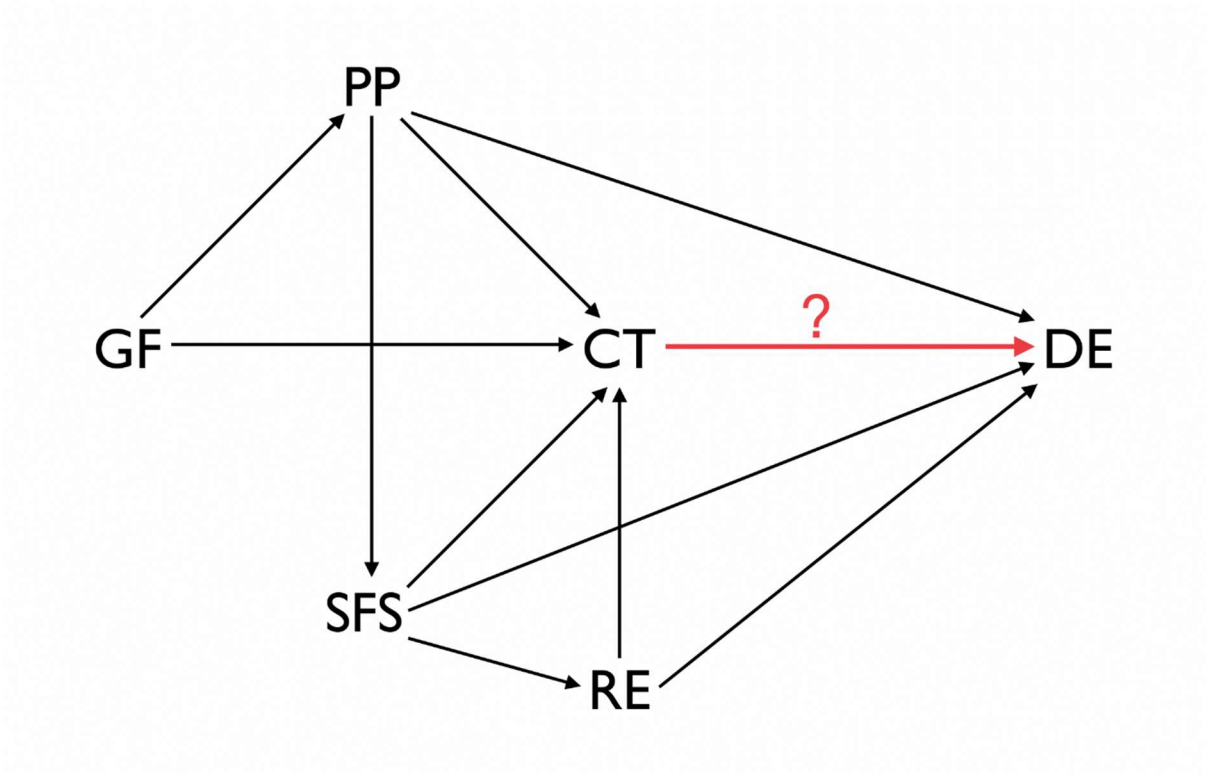
3. DAGs and the effect of childhood trauma (CD) on depression (DE)

A directed acyclic graph (also "causal diagram") displays arrows that encode whether a variable is assumed to affect another (arrow) or not (no arrow) (Pearl, 1995, 2009). An entire graph expresses *qualitative* assumptions on common causes and causal dependencies among them. After setting up a DAG, it is evaluated with regard to study design and quantitative analysis it may suggest (see below). A DAG is *non-parametric*; that is, all mathematical theorems apply independent of how the variables are scaled and distributed. A diagram must be *complete* with regard to the shared causes of **X** and **Y**. The arrows in the graph code assumptions on *direct effects*; for instance, we besides assume that "socio-economical family status" affects depression both directly and indirectly via "risk environment". Then DAGs are *non-parametric*; i.e., they do not make assumptions on how the variables are scaled and distributed and according to what mathematical function they affect one another.

Different and very much recommended introductory accounts for the field are provided by Dablander (2020) and Rohrer (2018). Here, we illustrate how a DAG accounts for *bias due to confounding*. Note that also the the bias sources of measurement (Hernán & Cole, 2009), non-compliance (Morgan & Winship, 2014; ch. 9), selection (Bareinboim & Pearl, 2016; Elwert & Winship, 2014) and missing data (Thoemmes & Mohan, 2015) can be addressed with DAGs, while revealing whether and how an effect can be estimated without bias.

For the effect of CT on DE we suppose just *four common causes* to demonstrate the use of the method rather than to provide an exhaustive account of all variables that are theoretically plausible and in line with the evidence: parental psychopathology (PP; Hankin, 2015; Lizardi & Klein, 2000), socio-economical family status (SFS; Freeman et al., 2016; Freisthler et al., 2006), risk environment (RE, e.g. living in a dangerous neighbourhood; Coulton et al., 2007)

and genetic factors (GF; Dunn et al, 2015; Peyrot et al., 2014). Figure 1 shows the model including the supposed effects of these factors on one another:



The factor of interest, CT is supposed to be influenced by parental psychopathology (PP), socio-economical family status (SFS) and risk environment (RE). "Non-parametric" also means that CT if affected by these variables through whatever function, f_{CT}, CT = f_{CT}(PP, SFS, RE, ϵ_{CT}). ϵ_{CT} is a summary of all other variables that influence CT. These are not visualized in the graph for parsimony since they have no effect on DE and, thus, on how adjustment should be done. Note that genetic factors (GF) also influence CT, but only indirectly through the path GF → PP → CT. DAGs encode *direct* effects in terms of the other variables that the graph contains. PP, for instance, may actually affect CT through other variables, but if these do not have causal connections with DE, these can be omitted as well.

Now, the "backdoor criterion" (Pearl, 1995) provides an algorithm to identify paths in the model that bring about non-causal associations between CT and DE and are to be eliminated.

These are called “backdoor paths“, a backward path must contain an arrow into CT and a backward connection into DE.

In the given example, there are seven backdoor paths:

(1) CT ← PP → DE

(2) CT ← PP ← GF → DE

(3) CT ← SFS → DE

(4) CT ← RE → DE

(5) CT ← SFS → RE → DE

(6) CT ← PP → SFS → DE

(7) CT ← PP → SFS → RE → DE

Furthermore, a backdoor path must also be "collider-free" to bring about non-causal association. This is because a common consequence (“collider”) of two factors does not cause an association between these factors — unless the common consequence is wrongly adjusted for (Elwert & Winship, 2014). In the example, a variable outside the model, school grades (SG), could be a common consequence of CT and SFS: CT → SG ← DE. Adjusting for SG would yield an otherwise not present CT-SFS association and, in turn, new backdoor paths and additional bias.

Now, the backdoor-criterion states that we have to identify a subset of confounders, that, if adjusted for, “blocks“ paths 1-7 such that, given the subset, the association between CT and DE is independent of the omitted confounders. This means that each backdoor path must lead through at least one variable of the subset. Here, only the sets (PP, SFS, RE) and (PP, SFS, RE, GF) fulfil this criterion. Thus, we don’t have to consider GF if we take PP, SFS and RE into account.

Importantly, the results of effect quantification may yet vary strongly across different specified DAGs. Different DAGs may appear similarly plausible, in which case researchers are well advised to collect *all variables* that are necessary for the different adjustments that the different DAGs call for. If one is lucky, however, sensitivity on this level is small because the different DAGs address the same key features of bias, although each model may just be a crude map of reality (VanderWeele, 2016).

4. Study design

The model tells us that we have to design a study that collects data not only on CT and DE (while establishing temporal sequence, the prerequisite for causation, ideally in a prospective study), but also on PP, SFS and RE. This is the essence of designing an observational study for causal inference (Rosenbaum, 2010; Shadish et al., 2002). In the example, PP, SFS and RE are summary variables. For adjustment to be sufficient, these constructs should be completely covered through a collection of variables that addresses all of a construct's aspects, and these must share the same causal relations (Ramsahai, 2012). For example, RE should include information on local crime rates, regional conflicts, air pollution and lack of infrastructure. To address PP sufficiently, parents should be comprehensively diagnosed. Moreover, to succeed completely in adjustment, the confounders have to be measured without error. Otherwise "residual bias" is expected (Morgan & Winship, 2018). For example, PPP assessment should use the best available instrument. Lastly, to keep bias due to selection small, a sample should be drawn from a source population that resembles the target population in the parameter of interest, here the magnitude of the effect of CT on DE (Elwert

& Winship, 2014). Again, this requires assumptions beyond the data, and these can be expressed with a DAG (Bareinboim & Pearl, 2016).

5. Adjustment after data have been collected

For the present purpose, we provide a brief summary with principal guidance on the adjustment methods because their details are vast and better explained with data. A comprehensive account with rich citations of original work is provided by Morgan and Winship (2014, chapters 5-7).

The simplest adjustment method to estimate the average treatment effect, **ATE**, is jointly regressing **Y** on **X** and the chosen **Z** variables in order to remove the **Z-Y** relations when comparing the **X**-groups. The method is very crude and may be pretty ineffective in balancing, in the example, individuals with and without CT with regard to the distributions of PP, SFS and RE. It performs increasingly poorly the more the effect of CT on depression varies across individuals. Heterogeneity in effects appears at least plausible for many if not most mental disorders and their aetiological factors. Regression may also work not well if the model does not fit the data or makes wrong assumptions on the distribution of errors. However, better fitting models (like generalized linear models) can be used. Besides, adding interactions between CT and effect-modifyers (which may differ from PP, SFS and RE) could capture at least some important features of effect-heterogeneity. Such a model also serves estimating potential outcomes: Given whatever value of whatever **Z**, the regression equation predicts the average outcome **Y** under **X** = 1 and **X** = 0 and, thus, the group difference in **Y** given **Z** (i.e., in individuals with such a **Z** value). This works for every combination of **Z** values, and averaging over these yields an estimate of **ATE** like the average effect of CT on depression. Alternatively, the individuals may be differently weighted (using the propensity

score, see below), such that **ATT** or **ATC**, respectively, are estimated: the effect of CT removal or CT experience, respectively.

Propensity-score methods instead focus on removing the **X-Z** relations. They firstly require a model on group assignment; that is, on the probability that **X** equals 1 given **Z**. This probability is called propensity score, ps. If ps is known and adequately adjusted for, ps is sufficient for unbiased effect estimates, the distinct **Z** variables do not need to be taken into account anymore (Rosenbaum & Rubin, 1983).

In the example, ps is the probability that an individual is traumatized in childhood, given his or her values in PP, SFS and RE (irrespective of whether an individual has truly experienced CT or not). ps is estimated for each individual with a model, often through logistic regression of P(**X** = 1) on **Z** yielding model-predicted probabilities of **X** = 1 given **Z**. Importantly, the model must be true in describing the actual assignment process. For instance, if PP and SFS interact in affecting the CT risk on the logistic scale, this interaction must be included in the regression equation. Otherwise **X** is not fully disconnected from **Z** and, again, residual confounding occurs. Uncertainty in model selection can be addressed by averaging an individual's ps across models that similarly fit the data. This is prevalent if the number of **Z** variables is large and in small samples and can be handled with "random forests" if the sample proportion of X = 1 ranges, say, between 30 and 70 percent. (A random forest is a collection of possible models, each of having the shape of a "tree" that predicts the average **Y** in each "terminal node" of the tree. Each such node results from a series of binary splits according to values of **Z** variables that bring about the largest difference in **Y** (Strobl et al., 2009).

After a model has been fitted, the range and distribution of ps values in both **X** = 0 and **X** = 1 must be inspected. It may turn out that the groups have different ps ranges, which may

indicate that, for example, some **X** = 1 individuals have no "twin" in **X** = 0 with regard to ps (and, thus, one or more **Z** variables). Such a finding may suggest that, for these individuals, **X** = 0 is not a meaningful counterfactual. This should, however, rarely happen if the model in the underlying DAG is sound. Otherwise, individuals that lack twins in the other group should be omitted from analysis and the analysis be restricted to the "region of common support". In consequence, the inferential population is limited to individuals with ps in this region. In the example, the data might contain individuals with low RE, low PP and high SFS only in the non-CT group. Omitting these would not allow including them in a causal conclusion. Another practical issue is ensuring whether ps sufficiently summarizes the **Z** variables; that is, whether **X** and each **Z** are independent given **Z**.

Once ps is computed and the points discussed above are addressed, there are several approaches that use ps to balance **X** = 0 and **X** = 1. Unlike parametric regression methods statistical matching methods draw each individual from one group with its observed **Y** value, calculate the difference with each observation in the other group that is similar in ps, then average across these "twins" and finally average across individuals. These procedures are non-parametric because they do not rely on assumptions on the distributions of **Y** in **X** = 0 and **X** = 1.

Matching can be done in individuals with **X** = 0, e.g. adjusting individuals with CT to those without CT (**ATC** estimation), or within **X** = 1 individuals, adjusting individuals with CT to those without CT (**ATT** estimation). Averaging **ATC** and **ATT** (weighted by the number of individuals with CT and without CT, resp.) yields **ATE**.

Specific matching algorithms differ in how they operationalize "similar" and weight the observations according to the magnitude of similarity (e.g. "kernel matching", "genetic

matching" and "optimal matching"). Unfortunately, the literature is inconsistent in what method works best. Therefore, the recommendation is using a range of methods to ensure that a particular conclusion is not due to the specific method used. Of course, all results then have to be reported and the exact analytic plan should be registered beforehand to prevent p-hacking.

Finally, *doubly robust estimation* methods ("inverse-probability-weighted regression adjustment" and "augmented inverse-probability") combine the approaches of eliminating the **Y-Z** and **X-Z** relations. They do this by jointly a) regressing **Y** on **X*Z** (and the main effects of **X** and **Z**) *and* b) weighting the individuals with functions of the propensity score. They are called doubly-robust because their bias in estimating the **X-Y** effect is the product of the biases in a) and b): If one bias is zero (small) the total bias is zero (small). Thus, one has two attempts to succeed (and each attempt is robust against wrong assumptions in the other). Today, various software packages including R, Python, Stata and SPSS include a range of these methods. However, care in their use must be taken since their implementation might differ what may cause unwarranted variation in results.

Finally note that DAGs give rise to two *other quantitative methods* to address bias due to confounders: a) "Mechanism-based" unravels an **X-Y** effect into direct and indirect effects that may be less confounded than the total effect (Morgan & Winship, 2014, chapter 10). b) The "instrumental variables" approach is often used to model non-compliance in **X** in a treatment or experimental study. An *intended* randomized treatment may serve as an "instrument", **I**, for estimating the effect of treatment according to the protocol (= **X**, what is confounded) as if there was perfect compliance. Under certain assumptions, the **X-Y** effect can be calculated from the **I-X** and **I-Y** associations (which may be less confounded), but

possibly only in a restricted population (Morgan & Winship, 2014, chapters 10 and 9., resp.). Also note how a DAG relates to observation and thus be empirically tested: It predicts a set of associations through pairs of variables that are causally or non-causally linked (e.g. due to common causes). Each of these predicted associations might be found in the subsequent study or not and, thus, invite model modification.

Legend to Figure 1:

DAG model of common causes of childhood trauma (CT) and depression (DE) and the causal relations between them. Assumed common causes are parental psychopathology (PP), socio-economical family status (SFS), risk environment (RE) and genetic factors (GF).

References

Anglemyer, A., Horvath, H. T., & Bero, L. (2014). *Healthcare outcomes assessed with observational study designs compared with those assessed in randomized trials*. Cochrane Database Syst Rev *4*, p. Mr000034 https://doi:10.1002/14651858.MR000034.pub2

Bareinboim, E., & Pearl, J. (2016). **Causal inference and the data-fusion problem.** *Proceedings of the National Academy of Sciences*, *113*(27), 7345-7352. https://doi.org/10.1073/pnas.1510507113

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, *56*(2), 81–105. https://doi.org/10.1037/h0046016

Coulton, C.J., Crampton, D.S., Irwin, M., et al (2007). How Neighborhoods Influence Child Maltreatment: A Review of the Literature and Alternative Pathways. *Child Abuse & Neglect, 31*(11–12), 1117–42. https://doi:10.1016/j.chiabu.2007.03.023

Dablander, F. (2020). An Introduction to Causal Inference. *PsyArXiv*. February 13. d https://doi:10.31234/osf.io/b3fkw

Dunn, E. C., Brown, R. C., Dai, Y., Rosand, J. et al. (2015). Genetic Determinants of Depression: Recent Findings and Future Directions. *Harvard Review of Psychiatry, 23*(1), 1–18. https://doi:10.1097/HRP.0000000000000054

Elwert, F., & Winship, C. (2014). Endogenous Selection Bias: The Problem of Conditioning on a Collider Variable. *Annual Review of Sociology, 40*, 31–53. https://doi: 10.1146/annurev-soc-071913-043455

Freeman, A., Tyrovolas, S., Koyanagi, A., et al. (2016). The Role of Socio-Economic Status in Depression: Results from the COURAGE (Aging Survey in Europe). *BMC Public Health 16*, 1098. https://doi:10.1186/s12889-016-3638-0

Freisthler, B., Merritt, D. H., & LaScala, E. A. (2006). Understanding the Ecology of Child Maltreatment: A Review of the Literature and Directions for Future Research. *Child Maltreatment, 11*(3), 263–80. https://doi.org/10.1177/1077559506289524

Greenhouse, J. B. (2009). Commentary: Cornfield, Epidemiology and Causality. *International Journal of Epidemiology 38*, 1199–1201. http://doi:10.1093/ije/dyp299

Greenland, S. (2005). Multiple-bias modelling for analysis of observational data. *Journal of the Royal Statistical Society A, 168*(2), 267–291. https://doi.org/10.1111/j.1467-985X.2004.00349.x

Greenland S. (2012). Causal inference as a prediction problem: Assumptions, identification and evidence synthesis. In: Berzuini C, Dawid P, Bernardinelli L (eds). *Causality: Statistical Perspectives and Applications*. Hoboken, NJ: Wiley:43–58.

Greenland, S. For and Against Methodologies: Some Perspectives on Recent Causal and Statistical Inference Debates. (2017). *European Journal of Epidemiology, 32*(1), 3-20.

Hankin, B. L. (2015). Depression from Childhood through Adolescence: Risk Mechanisms across Multiple Systems and Levels of Analysis. *Current Opinion in Psychology*, *4*,13–20. https://doi:10.1016/j.copsyc.2015.01.003

Hernán M. A., Alonso, A., Logan, R. et al. (2008). Observational studies analyzed like randomized experiments: an application to postmenopausal hormone therapy and coronary heart disease. *Epidemiology, 19*(6), 766–79. https://doi:10.1097/EDE.0b013e3181875e61

Hernán, M., A., & Cole, S.R. (2009). Invited Commentary: Causal Diagrams and Measurement Bias, *American Journal of Epidemiology*, *170*(8), 959–962, https://doi.org/10.1093/aje/kwp293

Hill, A. B. (1965). The Environment and Disease: Association or Causation? *Proceedings of the Royal Society of Medicine*, *58*(5), 295–300. https://doi.org/10.1177/0141076814562718

Höfler, M. (2005). The Bradford Hill considerations on causality: A counterfactual perspective. *Emerging Themes in Epidemiology, 2*, 11. https://doi:10.1186/1742-7622-2-11

Höfler, M., & Trautmann, S. (2019). Letter to the editor: When does selection generate bias in clinical samples? *Journal of Psychiatric Research, 116*, 189-190. https://doi:10.1016/j.jpsychires.2019.02.010

Hróbjartsson, A., Kaptchuk, T. J., & Miller, F. G. (2011). Placebo effect studies are susceptible to response bias and to other types of biases. *Journal of clinical epidemiology*, *64*(11), 1223–1229. https://doi.org/10.1016/j.jclinepi.2011.01.008

Johnson, R.B., Russo, F., Schoonenboom, J. Causation in Mixed Methods Research: The Meeting of Philosophy, Science, and Practice. *Journal of Mixed Methods Research*. 2019;13(2):143-162.

Lash, T.L., Fox, M.P., Fink, A. (2009). *Applying Quantitative Bias Analysis to Epidemiologic Data.* New York, NY: Springer.

Lizardi, H., & Klein D.N. (2000). Parental Psychopathology and Reports of the Childhood Home Environment in Adults with Early-Onset Dysthymic Disorder. *The Journal of Nervous and Mental Disease, 188*(2), 63–70. https://doi.org/10.1097/00005053-200002000-00

Maclure, M., & Schneeweiss, S. (2001) Causation of bias: the episcope. *Epidemiology*, *12*(1), 114–122.

Magaard, J. L., Seeralan, T., Schulz, H., & Brütt, A. L. (2017). Factors associated with help-seeking behaviour among individuals with major depression: a systematic review. *PLoS One.* 12(5):e0176730 https://doi:10.1371/journal.pone.0176730

Mansournia, M. A., Higgins, J. P. T., Sterne, J. A. C., & Hernán, M. A. (2017). Biases in randomized trials: a conversation between trialists and epidemiologists. *Epidemiology 28*(1): 54–59. https://doi:10.1097/EDE.0000000000000564

Morgan, S. L., & Winship, C. H. (2014). *Counterfactuals and Causal Inference. Methods and Principles for Social Research.* 2nd edition. Cambridge, UK: Cambridge University Press.

Pearl, J. *(1995). Causal diagrams* for empirical research. *Biometrika*, *82*(4), 669-688. https://doi.org/10.1093/biomet/82.4.669

Pearl, J. (2009). *Causality, models, reasoning and infere*nce. 2nd edition. Cambridge, UK: Cambridge University Press.

Peyrot, W. J., Milaneschi, Y., Abdellaoui, A. et al. (2014). Effect of Polygenic Risk Scores on Depression in Childhood Trauma. *The British Journal of Psychiatry: The Journal of Mental Science, 205*(2), 113–19. https://doi:10.1192/bjp.bp.113.143081

Ramsahai, R.R. (2012). Supplementary variables for causal estimation In: C. Berzuini, P. Dawid, L. Bernardinelli (eds). *Causality: Statistical Perspectives and Applications* (pp. 218-233). New York, NY, USA: Hoboken: Wiley.

Rohrer, J. (2018). Thinking Clearly About Correlations and Causation: Graphical Causal Models for Observational Data. *Advances in Methods and Practices in Psychological Science 2018, 1*(1), 27 –42. https://doi.org/10.1177/2515245917745629

Rosenbaum, P.R. (2010). *Design of Observational Studies*. New York, NY, USA: Springer.

Rosenbaum, P.R.; & Rubin D. B. (1983). The Central Role of the Propensity Score in Observational Studies for Causal Effects. *Biometrika,* 70(1), 41-55. https://doi.*org*/10.1093/biomet/70.1.41

Shadish, W.R., Cook, T.D., & Campbell, D.T. (2002). *Experimental and Quasi-Experimental Designs for Generalized Causal Inference.* Boston, MA, USA: Houghton Mifflin Company.

Strobl, C., Malley, J., & Tutz, G. (2009). An introduction to recursive partitioning: Rationale, application and characteristics of classification and regression trees, bagging and random forests. *Psychological Methods*, *14*(4), 323-348. https://doi.org/10.1037/a0016973

Thoemmes, F., & Mohan, K. (2015). Graphical Representation of Missing Data Problems *Structural Equation Modeling: A Multidisciplinary Journal 22*(4), 631–642. https://dx.doi.org/10.1007%2Fs10654-018-0447-z

VanderWeele, T. J. (2016). Commentary: On Causes, Causal Inference, and Potential Outcomes. *International Journal of Epidemiology*, *45*(6), 1809-1816. https://doi:org/10.1093/ije/dyw230