Toward causality and improving external validity

Peter Bühlmann[a,1]

"Felix, qui potuit rerum cognoscere causas," from the Latin poet Virgil (1), literally translated as "Fortunate, who was able to know the causes of things," hints at the importance of causality since a very long time ago. In PNAS, Bates et al. (2) start their contribution with the sentence "The ultimate aim of genome-wide association studies (GWAS) is to identify regions of the genome containing variants that causally affect a phenotype of interest," and they provide a highly innovative and original statistical methodology to provide sound answers to this aim. As we will argue, the causal inference problem is ambitious, and one has to rely on assumptions. The assumptions in ref. 2 are easy to communicate; the ability to communicate underlying assumptions makes their approach transparent, and, in our own assessment, their assumptions are very plausible.

When we observe correlation or dependence between some variables of interest, a main question is about the directionality: whether one variable is the cause or the effect of another one. Of course, it may happen that neither is true, because of hidden confounding. See Fig. 1 for a schematic view where all observed variables are exhibiting association dependence between each other but these are, in part, arising due to unseen hidden factors. If we were able to gain knowledge of causal directionality, obviously, this would lead to much improvement in understanding and interpretability of an underlying system. In Fig. 1, this means to infer the directed causal relations between the observed variables.

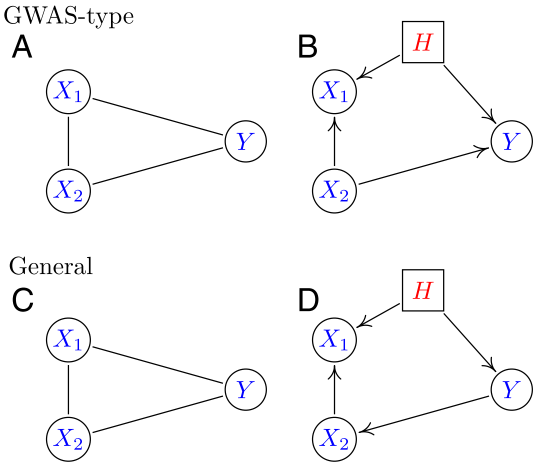

Fig. 1. **Observed and true system in two different settings (*A* and *B* setting and *C* and *D* setting). Response variable *Y* (phenotype) and covariates X_j $(j = 1,2)$ (for example, SNPs). (*A* and *C*) Observed variables X_1, X_2, Y in blue. An undirected edge represents association between the corresponding variables, for example, in terms of correlation or of (nonlinear) regression dependence (partial correlation) given all other observed variables. (*B* and *D*) True underlying systems, with observed variables in blue and hidden latent variable *H* in red. A directed edge represents a direct causal relation between the corresponding variables, with tail being the cause and head being the effect (i.e., the variable which is directly influenced by the causing variable). (*A* and *B*) Setting where all arrows between X_j to *Y* in *B* must point to *Y*, as in (most) GWAS. (*C* and *D*) The arrow direction in *D* between X_j and *Y* can go either way, as in general situations. The true underlying systems in *B* and *D* generate the association dependence in *A* and *C*, in terms of correlation or (nonlinear) regression dependence. Looking at such associations leads to spurious findings, that is, false positives with respect to causality.**

Association measures alone, like correlation or from (multivariate potentially nonlinear) regression, based on so-called observational data (data from the "steady state"), cannot provide answers to directionality and hence for causality in general; one needs additional assumptions or data from other experimental design settings. A randomized control trial (RCT) is a powerful gold standard for inferring causality, thanks to its very special experimental design (cf. ref. 3 and also *Perturbation Data as Input*). However, unfortunately, this gold standard method is often infeasible or unethical to do. In the absence of RCTs, other methodology has to be used, always relying crucially on some assumptions. Bates et al. (2) provide a highly interesting approach with plausible assumptions for causal inference in the particular field of GWAS; see

[a]Seminar for Statistics, ETH Zürich, CH-8092 Zürich, Switzerland

Author contributions: P.B. designed research, performed research, and wrote the paper.

The author declares no competing interest.

See companion article, "Causal inference in genetic trio studies," 10.1073/pnas.2007743117.

[1]Email: buehlmann@stat.math.ethz.ch.

First published October 12, 2020.

below. Before discussing this, we briefly elaborate more generally on the purpose of causality.

Main Scope of Causality

Besides having improved understanding of a mechanism, thanks to causal knowledge, we highlight two main (additional) goals of causal inference. They are often less ambitious and more realistic than inferring the entire network or graph with corresponding functional edge weights as in Fig. 1.

Predicting Specific Interventions: Treatment Effect

A classical goal of causality is prediction of an intervention or manipulation which has not been observed before. Causality gives quantitative answers to questions like: What would happen if we treat a patient with a certain drug (and the treatment intervention has not been done yet)? What would happen if we knock out a certain gene (and the gene intervention has not been performed yet)? Thus, causality gives an answer to a "what if I do" question (4, 5). In many applications, having accurate predictions to such questions is highly desirable.

Robustness against Unspecific Perturbations: External Validity

The problem tackled in ref. 2 is perhaps not so directly related to specific interventions, since it deals with single-nucleotide polymorphisms (SNPs) in GWAS where interventions on SNPs cannot be done. As a thought experiment, however, one can still think about what would happen to a disease status if a certain SNP were intervened on. Our message is that, even in absence of the possibility of doing direct interventions, causal inference is highly interesting (besides the interpretation issue mentioned above). The main reason is that the causal structure leads to certain invariances and robustness, as we briefly explain next.

Most scientific studies come with the claim that findings and results generalize to other individuals or populations and aim for external validity. In other words, the goal is replicability of findings: We want to infer results which are stable across different subpopulations, where each of the latter may be a perturbed version of a reference. Interestingly, such stability across different subpopulations or different perturbations has a very intrinsic relation to causality: Regression on the causal variables, the causal solution, exhibits (some) robustness or stability against perturbations arising from different subpopulations (6–8), and hence, a causal solution with its robustness leads to improved replicability and better external validity (in new studies, for new patients, etc.). In our view, this is a major advantage of the approach and findings from ref. 2: Their methodology, due to targeting causal relations, improves external validity!

Causal Inference Methods

Inferring causality from data is an ambitious task and crucially relies on the design of experiments or additional, often nontestable, assumptions.

Perturbation Data as Input

Learning causal structure and effects is easier with access to data from different perturbations of the system of interest. As mentioned already, the gold standard is a perturbation in the form of an RCT. There, the experimenter has the ability to do an intervention at a variable (being a candidate to be causal) or to assign a treatment: The randomization breaks all dependencies between the intervened variable and any possible hidden confounder. The powerful conclusion is that, after randomization, if there is an effect left between the intervened or treatment variable and a response of interest, it must be a (total) causal effect. An RCT leads to stability and external validity of (regression or group comparison) effects for a large class of perturbations. This is exactly the aim in, say, development of robust pharmacotherapy: The medication or active treatment effects should be "always" externally valid. If an RCT is infeasible, perturbation data from (nonrandomized) specific interventions or from unspecific changes of environment still are much more informative than having only access to observational data. Information from perturbation data leads to invariances and stability of (regression) effects which are induced by the different environments but where one has not really control over the "nature" of the perturbations which are either harmless or harmful for inferring (regression) effects. However, roughly speaking, when observing more perturbations, one can identify more invariance, stability, and robustness, and, eventually, the causal structure and effects (8). Thus, the most challenging setting for inferring causal effects happens when only observational data from the "steady state" are available.

The Approach by Bates et al. (2) Using Observational Data Only

The method in ref. 2 uses only observational data as input. However, two main assumptions are exploited. First, the directionality is naturally postulated pointing from genetic SNPs to the phenotype; that is, if there is unconfounded regression association between a phenotype Y and an SNP variable X_j, it must be directed $X_j \rightarrow Y$. This is the situation in Fig. 1 *A* and *B*. The same directionality is assumed from parental haplotypes to offspring SNPs. Second, for inferring unconfounded regression association, that is, the regression strength which is left after having adjusted for potential hidden confounding, a special so-called trio design study leads, in an elegant way, to such unconfounded regression effects. The assumption is that the stochastic mechanism of SNPs conditional on the parental haplotypes, that is, the corresponding conditional distribution, is independent of other potential hidden confounders, and this, in turn, allows the conclusion that a (potentially nonlinear) regression association between an SNP and a phenotype, given all other SNPs and the parental haplotypes, must imply a causal dependence. This is in exact analogy to an RCT: Conditioning on the haplotypes serves as a substitute for randomization! Bates et al. (2) refer to this as "variation in inheritance as a randomized experiment." Both assumptions can be clearly communicated and are very plausible, and this makes the claimed causal findings very convincing. Of course, there can still be violations of assumptions, and the authors mention unmeasured SNPs or selection bias, to name two prominent examples. Nevertheless, overall, the methodology in ref. 2 is a huge step forward to come closer to "true underlying causality."

Besides the way the methodology deals with fundamental assumptions for causality, it provides finite-sample statistical guarantees on the false discovery or the family-wise error rate. The main assumption here is that the model by Haldane (9) is assumed to be "true" (i.e., a very good approximation), and the inference techniques build on earlier beautiful work on simulating synthetic false features which then serve for counting false positives (10, 11).

Particularly fascinating is the possibility to include external (nontrio design) GWAS data to improve power; trio design studies are rare and of much lower sample size than standard GWAS studies, which may come at large scale. As illustrated in ref. 2, one can use any machine learning algorithm on external GWAS data to potentially improve power while the finite-sample guarantee on false positive detection is still valid.

Additional Thoughts

Bates et al. (2) nicely demonstrate the use of external data to potentially enhance power for detecting causal SNPs in trio design studies. Reversing the role of using external data, one could, and perhaps should, also use part of them to validate the results (and not use them in the discovery phase); see also ref. 12. As mentioned in *Robustness against Unspecific Perturbations: External Validity*, if the inferred structure is causal, it should exhibit some external validity on new data, ideally, across a few datasets from different environments or subpopulations. As a proposal, one could inspect the stability of the conditional distribution of the phenotye given the found causal SNPs, for example, by testing conditional independence of the phenotype and the environments given the causal phenotypes (13, 14). In particular, this could be done with standard, nontrio design GWAS external datasets which are available through various platforms.

In the absence of having trio design studies and in the absence of postulating directionality (as in GWAS from SNPs to phenotypes), the causal inference problem is much harder. Fig. 1 *C* and *D* indicates this setting, which includes, for example, transcriptomics or proteomics in biology where postulating directionality is often difficult or error prone. Perturbation data will play a crucial role for reliably making progress toward inferring causal structures and effects. Even when it is not possible to have randomized experiments, nonrandomized perturbations help massively. For fields like molecular biology and many others, prioritizing good candidates with respect to being causal is very valuable, even when strict statistical confidence statements seem out of scope (15). Clearly, such causal prioritization should be performed by causal inference methods rather than pure association techniques, where the latter range from simple correlation to advanced nonlinear regression or classification machine learning.

Acknowledgments

The research was supported by the European Research Council under Grant Agreement 786461 (CausalStats - ERC-2017-ADG).

1 Virgil, *Georgica* (vers 490, Book II, 29 BC).
2 S. Bates, M. Sesia, C. Sabatti, E. Candès, Causal inference in genetic trio studies. *Proc. Nat. Acad. Sci. U.S.A.* **117**, 24117–24126 (2020).
3 G. Imbens, D. Rubin, *Causal Inference for Statistics, Social, and Biomedical Sciences* (Cambridge University Press, 2015).
4 J. Pearl, *Causality: Models, Reasoning and Inference* (Cambridge University Press, ed. 2, 2009).
5 J. Pearl, D. Mackenzie, *The Book of Why: The New Science of Cause and Effect* (Basic, 2018).
6 T. Haavelmo, The statistical implications of a system of simultaneous equations. *Econometrica* **11**, 1–12 (1943).
7 A. P. Dawid, V. Didelez, Identifying the consequences of dynamic treatment strategies: A decision-theoretic overview. *Stat. Surv.* **4**, 184–231 (2010).
8 J. Peters, P. Bühlmann, N. Meinshausen, Causal inference using invariant prediction: Identification and confidence interval (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **78**, 947–1012 (2016).
9 J. B. S. Haldane, The combination of linkage values and the calculation of distances between the loci of linked factors. *J. Genet.* **8**, 299–309 (1919).
10 R. F. Barber, E. Candès, Controlling the false discovery rate via knockoffs. *Ann. Stat.* **43**, 2055–2085 (2015).
11 E. Candès, Y. Fan, L. Janson, J. Lv, Panning for gold: Model-X knockoffs for high dimensional controlled variable selection. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **80**, 551–577 (2018).
12 B. Yu, K. Kumbier, Veridical data science. *Proc. Natl. Acad. Sci. U.S.A.* **117**, 3920–3929 (2020).
13 R. Shah, J. Peters, The hardness of conditional independence testing and the generalised covariance measure. *Ann. Stat.* **48**, 1514–1538 (2020).
14 M. Azadkia, S. Chatterjee, A simple measure of conditional dependence. arXiv:1910.12327 (27 October 2019).
15 N. Meinshausen *et al.*, Methods for causal inference from gene perturbation experiments and validation. *Proc. Nat. Acad. Sci. U.S.A.* **113**, 7361–7368 (2016).