Journal of Data Mining in Genomics & Proteomics
ISSN: 2153-0602

Yazdani and Boerwinkle, J Data Mining Genomics Proteomics 2014, 6:1
DOI: 10.4172/2153-0602.1000163

Research article Open Access

Causal Inference in the Age of Decision Medicine

Yazdani A* and Boerwinkle E

Human Genetics Center, University of Texas Health Science Center at Houston, Houston, TX, USA

Abstract

Causal analyses and causal inference is a growing area of biostatics. In parallel, there is increasing focus on using genomic information to guide medical practice, i.e. personalized medicine or decision medicine. This perspective discusses causal inference in the context of personalized or decision medicine, including the assumptions and the concept that the task is different depending on whether the primary goal is the average response of treatment in the population or the ability to characterize the response for an individual or a subgroup. This perspective provides a tutorial of modern causal inference and then provides suggestions how application of specific kinds of causal inference would promote advances in translational sciences. The concept of the subpopulation causal effect is one path toward improved decision medicine. A dataset containing cardiovascular disease risk factor levels and genomic information is analyzed and different causal effects are estimated.

Keywords: Structural equation modeling; Coronary heart disease; Assignment mechanism

Introduction

Students are taught the perils of inferring causality from observational studies, and the shortcomings of nonrandomized clinical trials. In his seminal text book on modern epidemiology, Rothman et al. [1] dedicate considerable discussion to causal inference and even goes so far as to try to present and critique criteria necessary to consider or establish when concluding a causal relationship between two variables. Although formal conceptualization of causal inference began early in the last century, there remains disagreement concerning the ability to discovery novel causal effects from all but the most rigorous of controlled clinical trials and mechanistic experiments. Causality is connected to probability by some experts (e.g. [2], and [3]), whereby an attempt is made to quantitate the probability that A causes B, with assumptions about the mechanism by which individuals were assigned to levels of A. If that probability exceeds some threshold, a causal relationship is claimed. However, interpreting probabilities as causal quantities in the absence of clear knowledge about the assumptions underlying, this interpretation can lead to confusion. To avoid such confusion, Pearl promoted a deterministic interpretation of causal inference at the population level using structural equation modeling [4-6]. Rubin also defines the causal effect deterministically, but at the individual level [7]; for discussions on causal effect definition see also [8-11] and for causality in genetics effects see [12-13].

The difference between the "individual" and "population" causal effect has meaning that transcends esoteric or theoretical considerations. As an example, let's consider a drug, a desired outcome and an adverse event. The policy arm of health care wants to know whether prescribing the drug to the population of patients will increase the frequency of the desired outcome (and presumably reduce disease incidence) without undo increase in the frequency of the adverse event. The physician, on the other hand, wants to know whether prescribing the drug to the patient in his/her office at that time will elicit the desired outcome without leading to the adverse event in that patient. Typically, analyses and inference are done on a large sample from the population and then the results are used to make inference about whether the next individual sampled from the same population will respond or not. In its simplest form, inference about the response of the next individual sampled from the population is the average response in the population. Personalized medicine connotes the idea that treatment has been tailored to specific characteristics of the individual. In practice, treatment is not tailored to each individual, but rather is tailored to groups of individuals based on the results of specific diagnostic information, such as the level of a biomarker or genetic information. The term "Decision Medicine" has recently been suggested, which indicates a more immediate translational perspective [14].

The question we ask is whether we should approach causal inference including the assumptions and data analysis task differently depending on whether our primary interest is the average response of treatment in the population or the ability to characterize the response for an individual or a subgroup. Regardless of the term, the field of personalized medicine has much to benefit from advances in causal inference. This perspective provides a tutorial of modern causal inference and then provides suggestions how application of specific kinds of causal inference would promote advances in translational applications of personalized or decision medicine. The example application is carried out pragmatically using a graphical approach followed by Structural Equation Modeling (SEM).

A short tutorial on causal inference

Because causal inference is a challenging concept for many clinicians and researchers, even those trained in biostatistics, it is informative to reconsider the relationships among prediction, causation and association. Knowing the menu of causes of an outcome improves prediction above what may be done with a set of variables that are only associated with the response variable. Clearly, causation facilitates prediction, but the ability to predict does not imply causation because of the ubiquitous presence of association. The general dogma is that causal variables are better predictors of response compared to variables that are associated with response only because of their correlation with the causative variables. In some cases, these relationships have been

***Corresponding author:** Azam 'Mandana' Yazdani, School of Public Health, University of Texas Health Science Center-Houston, USA, Tel: 713-500-9808, E-mail: azam.yazdani@uth.tmc.edu

Received September 10, 2014; **Accepted** October 07, 2014; **Published** October 09, 2014

Citation: Yazdani A, Boerwinkle E (2014) Causal Inference in the Age of Decision Medicine. J Data Mining Genomics Proteomics 6: 163. doi:10.4172/2153-0602.1000163

medicine, where the groups may be defined by genotype or other relevant characteristics (e.g. gender). The groups may be defined based on the results of a statistical test of interaction between the treatment and a covariate or based on *a priori* biologic or clinical knowledge. The subpopulation causal effect is a practical compromise between the population and individual causal effect, which either do not take into account the unique characteristic of each individual or do show a practical path forward, respectively.

To discover an interaction effect, the illumination of the AM is not enough because there might be an interaction between T and a covariate Z on response, while Z does not vary with treatment. In such a case, we need to consider Z as an effect modifier. However, the illumination of the AM does not provide this information because Z is not a confounder of the treatment AM. All of this leads to fundamental challenges for the applied practitioner of causal inference; for each covariate we must consider the possibility that it is modifying the effect of treatment (i.e. interaction) [10].

To make the difference between the population and subpopulation effects more tangible, we consider a simple linear model. To find the population causal effect of T on Y, only the potential confounders of the treatment AM denoted by X are modeled, $y = \alpha + \beta t + \gamma x + u$. In this equation, u is a realization of U which includes variables independent of T, and stands for population causal effect of treatment on response Y. Now consider that there might be an interaction between T and a covariate P on response Y: $y = \alpha + \beta' t + \gamma x + \eta p + \lambda t.p + u'$, where U' is independent of T. The subpopulation causal effect is a combination of the effect of T alone and the effect of an interaction between T and a covariate, P. In the equation below, we see the difference between the coefficients of T in the two above equations:

$$\beta = \beta' + \lambda \cdot P(p = 1) \cdot$$

The effect of treatment in the class of P=1 is $\beta' + \lambda$ and the effect of treatment in the class of P=0 is β' . If λ is positive and we do not classify individuals regarding P, we will overestimate the effect of treatment for individuals in the group P=0 and underestimate it for individuals in the group P=1.

Example application

In practice and in real data applications, the concept of causal inference is best visualized as a Bayesian Network. In addition, the Bayesian network framework facilitates both estimation and hypothesis testing (i.e. statistical inference) in a real data analysis setting. A causal graph (Bayesian Network) is an illustration of the causal relationship among covariates, treatment, and response variable, as well as representation of assumptions. The existence of a directed edge $X \rightarrow Y$ means that X may have a direct causal effect on Y. Assume a DAG $D = (v, \varepsilon)$ where v is a set of random variables represented by nodes (or simple by letters) in DAG D and ε is a set of edges which connect the variables. The concept of a causal graph $D = (v, \varepsilon)$ depends on the variables v and edges ε and any inference depends on the set (v, ε). Assume P is a joint probability distribution on v . D and P must satisfy the basic Markov condition that every variable, $X_i \in v$, is independent of any subset of its predecessors conditioned on the set of its direct or immediate causes (parents), [4]. The two primary underlying assumptions are that there are no latent variables and no loops in the graph. With this brief review, we now embark on a real data example.

The aim of this example application is to identify causal relationships among 5 cardiovascular disease risk factors: body mass index (BMI), glucose, triglycerides, HDL-cholesterol and total cholesterol. We apply graphical models to visualize the AM and use SEM to estimate the causal effect, since they are the most pragmatic approaches to causal analysis. The data were collected on 14,749 individuals (10,753 European-Americans and 3,996 African-Americans) from the Atherosclerosis Risk in Communities study [32]. GWAS array genotype data were also available and principal components over these genotype data were calculated and used in the analysis to account for population structure [33]. There are multiple algorithms to identify causal structure which can be categorized in constraint-based, score-based, or hybrid learnings [34-36]. Here, the Peter and Clark (PC) algorithm, which is a constraint-based algorithm, was used. The PC algorithm is available in the pcalg package implemented on CRAN [37] and was extended to consider both genotype and phenotype data. The causal graph across the entire sample set is shown in Figure 1, but the principle components from the genome are not depicted to highlight the causal relationship among the phenotypes.

The central role of BMI on plasma glucose and lipid levels is evident from the graph. BMI influences TRG levels both directly and via HDL levels. Based on this topology, the structural equation for triglyceride levels is:

$$TRG = \alpha \cdot BMI + u \,, \quad (1)$$

where BMI is the treatment and TRG is the response variable, and there is no confounder of the AM depicted in figure 1. The estimated coefficient of BMI is $\hat{\alpha} = 0.18$. Because of the central role of BMI in the above graph and because BMI levels differ markedly between race groups, we hypothesize that the analyses should be repeated stratified by race. The race stratified causal graphs are shown in figure 2, a and b.

Within each race group, the topology or structures among phenotypes are similar, but not the same; there is no direct effect of BMI on triglyceride levels in the sample of African-Americans. To find the population causal effect of BMI on TRG, we must identify the causal effect within each race group:

$$TRG_{AA} = \alpha_{AA} \cdot BMI_{AA} + u_{AA} \quad TRG_{EA} = \alpha_{EA} \cdot BMI_{EA} + u_{EA} \,,$$

where the subscripts AA and EA indicate African-Americans and European-Americans, respectively. There is no confounder for this causal identification based on graphical back-door and front-door criteria, [4]. The coefficients $\alpha_{AA} = 0.09$ and $\alpha_{EA} = 0.26$ are interpreted causally because we assume $U \perp BMI$ in each subpopulation regarding race, which means by changing BMI, the

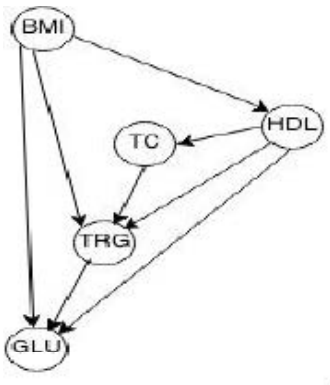

Figure 1: The estimated causal graph across the entire sample set.

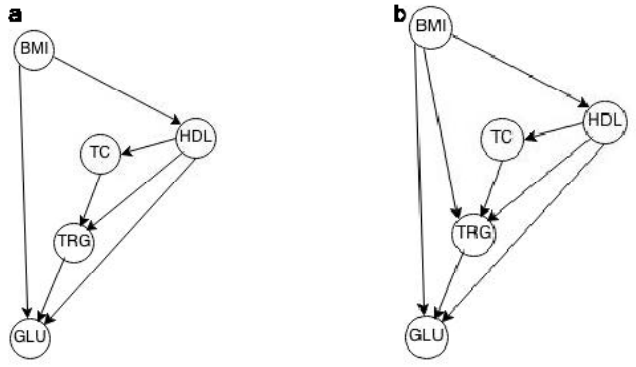

Figure 2: a. Estimated causal structure African-American (3856 individuals) b. Estimated causal structure European-American (10753 individuals).

rest of the model remains intact. The population causal effect of BMI on TRG is 0.22, which is a weighted sum of race-specific causal effects, $\alpha_{EA} \cdot r_{EA} + \alpha_{AA} \cdot r_{AA}$, where r_{EA} and r_{AA} are, respectively, the proportion of European-American and African-American in the population. We can see that the population causal effect of BMI on TRG, 0.22, is absolutely higher than the effect for AA due to the high effect for EA as well as the bigger portion of EA, the number of EA is nearly three times larger than AA. If we apply the population causal effect for every individual, we would overestimate the effect for AAs and underestimate it for EAs.

Note the above computations estimate the total causal effect of BMI on TRG which comprises direct effect as well as indirect effect through mediators, here through TC and HDL. In the analysis given below, other mediators are included in the structural equation to estimate the direct effect of BMI on TRG. The structural equation for the European-Americans is

$$TRG_{EA} = \alpha_{EA} \cdot BMI_{EA} + \lambda_{EA} \cdot HDL_{EA} + \theta_{EA} \cdot TC_{EA} + u_{EA}, \qquad (2)$$

and the direct effect of BMI on TRG is $\hat{\alpha}_{EA} = 0.11$.

To better reflect the personalized effects of BMI on the other phenotypes, we next consider genotypes which influence only the variables of interest. There are 11 genotype-derived principle components which influence TRG, denoted by vector G in the equation below. By entering G into the model, we are able to account for more of the variation of TRG and increase the coefficient of determination of the model:

$$TRG_{EA} = \alpha_{EA} \cdot BMI_{EA} + \lambda_{EA} \cdot HDL_{EA} + \theta_{EA} \cdot TC_{EA} + \beta_{EA}^{T} G_{EA} + u_{EA}, \qquad (3)$$

where G_{EA} and β_{EA} are 11×1 column matrixes. Since G comprises only variables influential on TRG and is independent of BMI, the coefficient of BMI does not change by entering G into the model. By taking into account genotype, equation 3 moves one step closer to estimate TRG for a new individual at the genotype level.

Consider the following scenario and question related to i^{th} individual with a TRG level equal to 54 and a BMI equal to 25.75. What would be the value of TRG_i if the value of BMI_i was lowered by 5 while the value of other mediators were held the same? For individual causal effect by SEM see [6]. In this example, assume *i* is a European-American. Therefore, the structural equation is

$$TRG_i = 0.11 \cdot BMI_i - 0.40 \cdot HDL_i + 0.29 \cdot TC_i + 0.23, \qquad (4)$$

where the data have been rescaled to standardized units. To answer the question, we keep the observed values of HDL and TC unchanged and set the value of BMI to calculate $TRG_i(BMI_i - 5)$. The causal effect of this change on TRG level of individual *i* is

$$\text{Causal effect} = TRG(BMI_i - 5) - TRG_i = 43.78 - 54 = -10.22$$

In other words, when individual *i* loses weight to lower BMI_i by 5, the triglyceride level is predicted to decrease by 10.22.

Conclusion

There are two barriers slowing the integration of genomic information for translational studies. The first is the small effect sizes of most genetic loci. The second is the associative nature of most genetic studies, with little information about the causative mutations. To place decision medicine in a causal framework, the causal effect must be defined precisely. Pearl defines the causal effect over the population and renders a framework to identify population causal effect, which is often operationalized graphically in a practical setting [4]. Rubin on the other hand, defines the causal effect for each individual, and applies the concept of the "potential outcome" [7]. Because of the fundamental problem in discovering and measuring an individual causal effect, we typically compare similar (i.e. not the same) treated and untreated individuals. The degree of similarity must be defined and careful consideration of covariates and potential confounders must be considered.

Personalized medicine is a theoretical ideal that has given way to decision medicine for using genomic or other biomarker information to guide treatment decisions. In a typical epidemiologic or clinical trials scenario, an interaction analysis is done (and replicated) and then subgroups or subpopulations are created based on the interaction results. In the context of graphical causal inference used here, the topologic structural relationships among the variables may be different between groups. For example, in the field of pharmacogenetics, subgroups of patients are defined by genotype or other genomic information (e.g. gene expression), and the causal effect of drug treatment is different between subgroups but assumed to be the same among individuals within a subgroup. As the rate of gene discovery and a role of genomic information in disease association increases, the frequency of causal analyses in a translational setting will also increase. The purpose of this perspective was to provide brief tutorial of causal inference and to discuss the application of specific kinds of causal inference in decision medicine.

Acknowledgement

The Atherosclerosis Risk in Communities Study is carried out as a collaborative study supported by National Heart, Lung, and Blood Institute contracts (HHSN268201100005C, HHSN268201100006C, HHSN268201100007C, HHSN268201100008C, HHSN268201100009C, HHSN268201100010C, HHSN268201100011C, and HHSN268201100012C), R01HL087641, R01HL59367 and R01HL086694; National Human Genome Research Institute contract U01HG004402; and National Institutes of Health contract HHSN268200625226C. The authors thank the staff and participants of the ARIC study for their important contributions. Infrastructure was partly supported by Grant Number UL1RR025005, a component of the National Institutes of Health and NIH Roadmap for Medical Research.

Funding

This work was supported by Cancer Prevention Research Institute of Texas.

References

1. Rothman KJ, Greenland S, Lash TL (2008) Modern Epidemiology, Lippincott Williams and Wilkins.

2. Williamson J (2005) Bayesian Nets and Causality, Oxford University Press.
3. Dawid AP (2007) Fundamentals of statistical causality. Research Report 279, Department of Statistical Science. University College London.
4. Pearl J (2009) Causality: Models, Reasoning, and Inference. Cambridge University Press, New York.
5. Pearl J (2010) An introduction to causal inference. Int J Biostat 6: Article 7.
6. Pearl J (2011) The Causal Foundations of Structural Equation Modeling. Handbook of Structural Equation Modeling. New York: Guilford Press.
7. Rubin DB (2005) Causal inference Using Potential Outcomes: Design, Modeling. Decisions. J Am Statist Assoc. 100: 322-331.
8. Russo F, Williamson J (2011) Causality in the Sciences, Oxford University Press.
9. Shadish WR, Sullivan KJ (2012) Theories of Causation in Psychological Science. APA Handbook of Research Methods in Psychology Science 1: 23-52.
10. Berzuini C, Dawid AP, Zhang H, Parkes M (2012) Analysis of interaction for identifying causal mechanisms. In Causality: Statistical Perspectives and Applications. University of Cambridge, Cambridge, UK.
11. Gillies D, Sudbury A (2013) Should causal models always be Markovian? The case of multi-causal forks in medicine. Eur J Philos Sci 3:275-308.
12. Vansteelandt S, Lange C (2012) Causation and causal inference for genetic effects. Hum Genet 131: 1665-1676.
13. VanderWeele TJ, Hernan MA (2012) Causal effects and natural laws: towards a conceptualization of causal counterfactuals for nonmanipulable exposures with application to the effects of race and sex. In: Berzuini,C, Dawid P, and Bernardinelli L (eds) causal Inference: Statistical Perspectives and Applications Wiley, Canada.
14. Goddard KA, Knaus WA, Whitlock E, Lyman GH, Feigelson HS, et al. (2012) Building the evidence base for decision making in cancer genomic medicine using comparative effectiveness research. Genet Med 14: 633-642.
15. Chakraborty R, Boerwinkle E (1989) Effects of multiple markers on variation of a quantitative trait: power analysis with measured genotype information . The American Journal of Human Genetics 45: A236-0926.
16. Nissen SE, Tardif JC, Nicholls SJ, Revkin JH, Shear CL, et al. (2007) Effect of torcetrapib on the progression of coronary atherosclerosis. N Engl J Med 356: 1304-1316.
17. Voight BF, Peloso GM, Orho-Melander M, Frikke-Schmidt R, Barbalic M, et al. (2012) Plasma HDL cholesterol and risk of myocardial infarction: a mendelian randomisation study. Lancet 380: 572-580.
18. Rubin DB (1974) Estimating Causal Effects of Treatments in Randomized and Nonrandomized Studies. J Educ Psychol 66: 688-701.
19. Holland PW (1986) Statistics and Causal Inference. J Amer Statist Assoc 81:945-960.
20. Herrington DM, Howard TD (2003) From presumed benefit to potential harm--hormone therapy and heart disease. N Engl J Med 349: 519-521.
21. Kasy M (2013) Why experimenters should not randomize, and what they should do instead. European Economic Association & Econometric Society, Gothenburg, Sweden.
22. Rosenbaum PR (2010) Design of Observational Studies.
23. Neyman J (1990) On the Application of Probability Theory to Agricultural experiments. Essay on Principles. Section 9. Statistical Science 5: 465-480.
24. Yazdani A, Boerwinkle E, (2014) Formalizing a Causal Quantity at Population Level, International Journal of Research in Medical Sciences.
25. Basu D (1980) Randomization Analysis of Experimental Data in the Fisher Randomization Test. J Amer Statist Assoc 75: 575-582.
26. Fisher RA (1918) The Correlation Between Relatives on the Supposition of Mendelian Inheritance. Transactions of the Royal Society of Edinburgh, 52: 399-433.
27. Teslovich TM, Musunuru K, Smith AV, Edmondson AC, Stylianou IM, et al. (2010) Biological, clinical and population relevance of 95 loci for blood lipids. Nature 466: 707-713.
28. Lange LA, Hu Y, Zhang H, Xue C, Schmidt EM, et al. (2014) Whole-exome sequencing identifies rare and low-frequency coding variants associated with LDL cholesterol. Am J Hum Genet 94: 233-245.
29. Morrison AC, Voorman A, Johnson AD, Liu X, Yu J, et al. (2013) Whole-genome sequence-based analysis of high-density lipoprotein cholesterol. Nat Genet 45: 899-901.
30. Morrison AC, Bare LA, Chambless LE, Ellis SG, Malloy M, et al. (2007) Prediction of coronary heart disease risk using a genetic risk score: the Atherosclerosis Risk in Communities Study. Am J Epidemiol 166: 28-35.
31. Visscher PM, Hill WG, Wray NR (2008) Heritability in the genomics era--concepts and misconceptions. Nat Rev Genet 9: 255-266.
32. [No authors listed] (1989) The Atherosclerosis Risk in Communities (ARIC) Study: design and objectives. The ARIC investigators. Am J Epidemiol 129: 687-702.
33. Pritchard JK, Stephens M, Donnelly P (2000) Inference of population structure using multilocus genotype data. Genetics 155: 945-959.
34. Spirtes P, Glymour C, Scheines R (2000) Causation, Prediction, and Search.
35. Korb K, Nicholson AE (2010) Bayesian Artificial Intelligence. Chapman & Hall/CRC Press,UK.
36. Nagarajan R, Scutari M, Lebre S (2013) Bayesian Networks in R with Applications in Systems Biology. Springer.
37. Kalisch M, Machler M, Colombo D, Maathuis MH, Buhlmann P (2010) Causal Inference using Graphical Models with the R Package pcalg. J Statistical Software 47: 1-25.