International Journal of Research in Medical Sciences
Yazdani A et al. Int J Res Med Sci. 2014 Nov;2(4):1368-1370
www.msjonline.org
pISSN 2320-6071 | eISSN 2320-6012

DOI: 10.5455/2320-6012.ijrms20141123

Research Article

Causal inference at the population level

Azam Yazdani*, Eric Boerwinkle

School of Public Health, University of Texas Health Science Center, Houston, USA

Received: 22 July 2014
Accepted: 9 August 2014

***Correspondence:**
Azam "Mandana" Yazdani,
E-mail: azam.yazdani@uth.tmc.edu

ABSTRACT

Three elements are needed to formalize a causal quantity at the population level: response, treatment, and the causal element, which are introduced here by notation. Inclusion of two essential causal assumptions, the monitoring and illumination assumptions, in a function distinguishes causal from association analyses. The discussion provides insight into causal inference.

Keywords: Assignment mechanism, Causal inference, Observational study

INTRODUCTION

Traditional statistical inference focuses on response and treatment variable, and their relationships. With growing interest in causal inference, there needs to be greater attention paid to the third leg of a "causal stool": the causal element. In this paper, a causal element at the population level is introduced in order to formalize a causal quantity. The formalization not only comprises randomization and intervention but also leads to an approach to observational studies.

Structural equation modeling is often the applied statistician's first foray into causal inference. Structural equation models are used by many, but causal interpretations are generally questioned or avoided, even by leading practitioners. In applications of structural equation modeling, one needs to postulate the structure behind probability distributions or mechanism of assigning treatment. This lesson was gradually forgotten, and often conditional probabilities have been mistakenly considered as a causal quantity. To cope with this frequent mistake, scientists began to formalize the assignment mechanism and represent it by a rigorous notation. Rubin[1-3] formalized the assignment mechanism and rendered a framework for causal inference; see also Imbens.[4] Pearl[5-7] rendered and simplified causal framework through application of causal graphs, which satisfy Markov conditions. Despite efforts by these scientists and others such as Dawid,[8] Pearl[5] commented that causality is still seeking to find an applied language.

Since causality is a challenging novel concept for many, notations and explanations are presented in order to help the reader better understand the concept and, as a result, promote appropriate applications. The outline of this perspective is as follows. In section 1, the underlying question is explained in order to motivate the problem. In section 2, the necessary elements for a causal quantity are introduced. In section 3, the causal quantity is explained; and a conclusion is represented in section 4. The notation and formalization presented here will help the interested reader gain a better understanding of causal concepts.

WHAT IS THE UNDERLYING QUESTION?

To understand the effect of treatment on a response variable in a population, we compare treated and untreated units and make inference about the effect of treatment. We would like any difference between the

treated and untreated units to be ascribed to the effect of treatment. This inference is accurate if we can assume that other inputs on the response variable are balanced between treated and untreated groups. In many cases, this assumption is best satisfied through randomization of the units to the treated and untreated groups. However, randomization is often expensive, impractical or unethical. In this case, a clear understanding of the assignment mechanism is critical.

Assume we have N individuals in the population and that all are under study. Consider a treatment variable (*T*) which, for simplicity, has two levels active treatment "*t*" and control treatment "*c*". Each unit can be assigned to either treatment or control, and in any single study only one of the 2^N different assignment possibilities will be realized. To have causal inference, we need to understand why one of the 2^N possibilities was actually realized and not the others. In other words, we need to understand the assignment mechanism. If knowledge related to a unit's reaction to treatment is used, either knowingly or unknowingly, then the assignment mechanism itself is considered confounded. In the presence of such confounding, the ability to make causal inference is seriously compromised, and the ability to adequately adjust for such confounding should always be questioned.

NOTATION AND ASSUMPTIONS

Let $Y_i(t)$ and $Y_i(c)$ be the i^{th} unit's reactions assigned to the treatment and control, respectively called "potential outcomes".[2,9] Although unrealistic in most simple practical applications, if all N individuals are assigned to the treatment group, we observe $Y(t) = (Y_1(t),...,Y_i(t),...,Y_N(t))$; and if all N individuals are assigned to the control group, we observe $Y(c) = (Y_1(c),...,Y_i(c),...,Y_N(c))$. In this special case, two values are defined for each individual- one under treatment and the other under control. For the entire population, we can define two arrays *Y(t)* and *Y(c)*. In most applications, treatment and control cannot be assigned to each individual. Instead, we observe some values of array *Y(t)* and some values of array *Y(c)* regarding the actual treatment assignment. Let variable $Y_{obs}(t)$ stands for the observed components of *Y(t)* and variable $Y_{obs}(c)$ for the observed components of *Y(c)*.

In causal inference, the element which plays a role is not only the value of revealed responses but also the reason why one of the 2^N possible assignments has been realized and not the others. Clarification of the Assignment Mechanism (AM) is essential in causal inference because the practitioner makes inference using the observed responses which are determined by the one realized assignment. The notation $AM\ (K_R)$ is introduced as the third element and is called causal element. Identification of causal element implies that we fully understand the mechanism used to assign treatment to individuals. K_R comprises any knowledge related to response, so the practitioner must be diligent to gather as much information regarding the assignment mechanism as possible and cautious not to let a priori biases misinform knowledge about the assignment mechanism.

$AM\ (K_R)$ is called the causal parameter or causal element. Through a better understanding of $AM\ (K_R)$, we can identify a realization in which the responses are proper for causal inference. In probability language, the assignment mechanism of units in the population with a particular causal element is represented as following:

$$P(T_i = t \mid AM(K_R) = f) = P(T_j = t \mid AM(K_R) = f),$$

for $\forall i, j \in$ the (sub)population identified with $AM\ (K_R) = f$. *f* simply indicates that the causal element $AM\ (K_R)$, is well-identified.

The two popular approaches to identify the causal element are propensity score[11] and causal graphs.[5] In the latter, causal graphs are illustrations of AMs, we illustrate the causal structure behind observations and graphically find the confounders of AM; and then, individuals with the same likelihood to receive treatment. In the former, we apply mathematical method propensity score to find units with the same propensity to receive treatment. In both methods, the aim is to find treated and untreated individuals in observational study as if we have randomization. This is carried out through considering have confounder of AM, which is possible by (K_R).

Like any complex analytic, it is necessary for the practitioner to be mindful of the underlying assumptions. In the area of causal inference, the assumptions can be classified into two categories: Monitoring Assumptions, where the experiment must be monitored from the time of assignment to the time of observation; and Illumination Assumptions, where the assignment mechanism or data generating process themselves must be understood. These two assumptions are necessary to be considered in any causal inference. The monitoring assumption is satisfied if there is no interference either from the units or from external factors;[10] for instance, treatment is assigned blindly, units are independent, and during time no other event changes the conditions influential on units' responses. Violating the monitoring assumption means that the difference between treatment and control responses cannot be solely ascribed to the treatment; other factors are involved too. The monitoring assumptions are represented by notations $Y_{obs}(t)$ and $Y_{obs}(c)$ which means observations are units' reactions under only treatment or control, which is called intervention.

The illumination assumptions are represented by the causal element, $AM\ (K_R)$. Clarification of AM (i.e. illumination), which is done by K_R, represents a complete understanding of how the units have been assigned to

treatment or how the observations (data) have been generated.

CAUSAL QUANTITY

In the population or subpopulation in which the causal element *AM* (K_R) has been identified and the treated and control units have been observed, we are able to infer causal effects by comparing the following two quantities,

$$P(Y_{obs}(t) = y \mid AM(K_R) = f) \;\&\; P(Y_{obs}(c) = y \mid AM(K_R) = f). \quad (1)$$

Both quantities are conditioned on *AM* (K_R) = *f;* which means the assignment mechanism has been identified regarding knowledge related to response, and the assignment mechanism is the same over the population. In this case, we are able to make causal inference. Application of the causal element and the ability to make causal inference, in this case, is due to the illumination assumption. We conclude that a comparison of the quantities in (1) leads to causal effect over the population. The two quantities in (1) might not involve observations on all N units. In this case, to find the causal effect over the population of N units, we compute the weighted average of the causal effects measured in each subpopulation identified by the causal element. It is important to note the causal element *AM* (K_R) cannot be replaced by a design variable or an adjusting covariate. The notation *AM* (K_R) means the assignment mechanism has been fully identified by considering knowledge related to the response, thus conveying more information than conditioning on a covariate.

CONCLUSION

The quantity in (1) is comprehensive because it comprises both intervention and randomization, and it leads us to causal inference in observational studies. In the case of randomization, no knowledge related to the response is considered, and the causal element is the same in the treated and untreated groups. In the case of observational studies that cannot be randomized, we assume after determining causal element, the units are independent and we observe them under both the control and treatment. By introducing the causal element, the conditional probability is kept for the association component of the study, and an additional component with the causal element is present in order to facilitate causal inference. If one is given the conditional probabilities:

$$P(Y = y \mid T = t, X = x) \;\&\; P(Y = y \mid T = c, X = x),$$

the practitioner wishing to go beyond mere association and make causal inference, the first question he/she must address is: What is the causal element? In the case of the observational or non-randomized intervention, the challenge is to find the structure behind the observations, which is an accurate approach to find population causal effect.

Funding: This work was supported by Cancer Prevention Research Institute of Texas
Conflict of interest: None declared
Ethical approval: The study was approved by the institutional ethics committee

REFERENCES

1. Rubin DB. Teaching statistical inference for causal effects in experiments and observational studies. J Educ Behav Stat Fall. 2004;29(3):343-67.
2. Rubin DB. Causal inference using potential outcomes: design, modeling, decisions. J Am Stat Assoc. 2005;100:322-31.
3. Rubin DB. Reflections stimulated by the comments of Shadish (2010) and West and Thoemmes (2010). Psychol Methods. 2010;15(1):38-46.
4. Imbens, GW. Nonparametric estimation of average treatment effects under exogeneity: a review. Rev Econom Stat. 2004;86(1):4-29.
5. Pearl J. Introduction to probabilities, graphs, and causal models. Pearl J, eds. In: Causality: Models, Reasoning and Inference. 2nd ed. New York: Cambridge University Press; 2009: 1-64.
6. Pearl J. An introduction to causal inference. Int J Biostat. 2010 Feb 26;6(2):Article 7.
7. Hoyle RH. The causal foundations of structural equation modeling. In: Pearl J, eds. Handbook of Structural Equation Modeling. 3rd ed. New York: Guilford Press; 2011: Chapter 5.
8. Dawid AP. Fundamentals of statistical causality. In: Dawid AP, eds. Research Report 279. London: Department of Statistical Science, University College London; 2007: 94.
9. Rubin DB. Estimating causal effects of treatments in randomized and nonrandomized studies. J Educ Psychol. 1974;66:688-701.
10. Rubin DB. Discussion of "Randomization Analysis of experimental data in the fisher randomization test" by D. Basu. J Am Stat Assoc. 1980;75:591-3.
11. Rosenbaum PR. Causal inference in randomized experiments. In: Rosenbaum PR, eds. Design of Observational Studies. 2009th ed. New York: Springer; 2009: 21-64.

DOI: 10.5455/2320-6012.ijrms20141123
Cite this article as: Yazdani A, Boerwinkle E. Causal inference at the population level. Int J Res Med Sci 2014;2:1368-70.