Preventive Veterinary Medicine 113 (2014) 281–297

Contents lists available at ScienceDirect

Preventive Veterinary Medicine

journal homepage: www.elsevier.com/locate/prevetmed

Making valid causal inferences from observational data

Wayne Martin*

Professor Emeritus, University of Guelph, Guelph, Ontario, Canada N1G 2W1

ARTICLE INFO

Article history:
Received 19 January 2013
Received in revised form 29 August 2013
Accepted 13 September 2013

Keywords:
Causal inference
Cause
Component cause
Causal diagram
Counterfactual
Multivariable model
Propensity score
Instrument variable
Boosted regression
Marginal structural model
Critical appraisal
Forward projection
Causal guidelines

ABSTRACT

The ability to make strong causal inferences, based on data derived from outside of the laboratory, is largely restricted to data arising from well-designed randomized control trials. Nonetheless, a number of methods have been developed to improve our ability to make valid causal inferences from data arising from observational studies.

In this paper, I review concepts of causation as a background to counterfactual causal ideas; the latter ideas are central to much of current causal theory. Confounding greatly constrains causal inferences in all observational studies. Confounding is a biased measure of effect that results when one or more variables, that are both antecedent to the exposure and associated with the outcome, are differentially distributed between the exposed and non-exposed groups. Historically, the most common approach to control confounding has been multivariable modeling; however, the limitations of this approach are discussed. My suggestions for improving causal inferences include asking better questions (relates to counterfactual ideas and "thought" trials); improving study design through the use of forward projection; and using propensity scores to identify potential confounders and enhance exchangeability, prior to seeing the outcome data. If time-dependent confounders are present (as they are in many longitudinal studies), more-advanced methods such as marginal structural models need to be implemented. Tutorials and examples are cited where possible.

1. Introduction

Humans have made causal inferences based on their observations for thousands of years. As one example, over 2000 years ago, a philosopher, Lucretius Caras, showed incredible perception into how nature works and he recorded his insights in a poem entitled "On the Nature of Things". Based on his observations, Caras thought that the basic building blocks of everything, living or dead, were eternal invisible particles that were infinite in number, but limited in size and shape. He posited that everything was formed of these "seeds" and on death or dissolution, everything returns to them (Greenblatt, 2011). As a second example, all epidemiologists know of the work of John Snow and colleagues in the 1800s (Bingham et al., 2004; Koch, 2008), and how with astute observations, and what today we might call a cohort study—with exposure based on level of salt in the home water supply—they concluded that cholera was caused by invisible (at the time) microorganisms that entered the water supply via human fecal material contamination. Today, "The goal of most, if not all, scientific investigation is to uncover causal relationships." (Aiello and Larson, 2002). As De Vreese (2009) states "the goal of epidemiologic research is, ultimately, disease prevention" which requires the identification of causal factors. And, according to Constantine (2012) to help achieve this goal requires a "deep understanding of the research topic, respect for the assumptions and limitations of the analytical tools employed, and perhaps most importantly, a strong theoretical foundation."

Despite the good intentions of most epidemiologists, there has been concern about the large number of

* Tel.: +1 519 824 9637.
E-mail address: swmartin@uoguelph.ca

0167-5877/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.prevetmed.2013.09.006

apparently false rejections of the null hypothesis in observational studies (de Jonge et al., 2011). In the mid-1990s, Taubes (1995) opined that epidemiology had reached (or had passed) its limits as a science. And, in 2001, editors of the *International Journal of Epidemiology* wondered if it was time to "call it a day?" (Davey Smith and Ebrahim, 2001). Thus, in 2012, it is comforting to see that over the past decade epidemiology has flourished as a science. Furthermore, credible epidemiologists are holding to the view that observational studies, "warts and all", are our best scientific approach for improving the health of humans (Hernan, 2011). Nonetheless, we must recognize that making valid causal inferences from observational data is a challenging, "risky" process (Hernan and Robins, 2006a). Indeed, Rothman and Greenland (2005) reminded us "... all of the fruits of scientific work, in epidemiology or other disciplines, are at best only tentative formulations of a description of nature ... the tentativeness of our knowledge does not prevent practical applications, but it should keep us sceptical and critical." Hence, because most epidemiologists will continue to rely on data from observational studies, to identify causal associations between exposures and outcomes, new approaches to support our making valid causal inferences are needed.

Recently, philosophical discussions of causal inferences have included approaches to identifying and understanding causal factors in complex systems (Campaner, 2011; De Vreese, 2009; Rickles, 2009; Ward, 2009a). More complete reviews on the philosophy of causal inference are available also (Aiello and Larson, 2002; Weed, 2002; White, 2001; Robins, 2001). A number of important papers on the quantitative aspects of causal modeling are published in two special issues of the *International Journal of Biostatistics* (Moodie and Stephens, 2010a,b; Moodie et al., 2012).

My objectives here are to review the literature on making causal inferences from non-experimental data and to make recommendations on how we might improve our ability to make valid causal inferences from observational study-derived data.

2. Defining a cause

Rothman (1976) reviewed the concepts of "cause" from an epidemiological perspective. For practical purposes, like Susser (1991), I define a cause of disease as any factor that produces a change in the nature or frequency of the health outcome. Often, epidemiologists have separated biological causes (those operating within individuals) from population causes (those operating at or beyond the level of the individual). For example, infection with a specific microorganism often is viewed as a biological cause of disease within individuals. In contrast, lifestyle, nutrition, or other factors that act at the group level or beyond (e.g. weather) and affect whether or not individuals are exposed to the microorganism (or alternatively, affect the individual's susceptibility to the effects of exposure), would be deemed population causes. Epidemiologists recognize that whereas disease occurs in individuals, "epidemiology deals with groups of individuals because the methods for determining causality require it" (De Vreese, 2009). Further, it is vital that we include social as well as biological factors in our study of health and disease, especially in humans (Kaplan, 2004; Harper and Strumpf, 2012). Because most causes act in concert with other causes, we recognize that a single cause need not invariably produce the outcome, and a cause need not be directly causal of the outcome. Given this complexity, our challenge is how to develop a standardized approach to identify when an exposure should be deemed to be a cause of the effect (more on this later) and to estimate the magnitude of its' effect.

In searching for causes, although we stress a holistic approach to health and disease we cannot consider every potential causal factor in a single study. Rather, we need to place limits on the portion of the "real world" we study and, within this, we constrain the list of factors we identify for investigation. Being pragmatists, we seek to identify causal factors that we can manipulate to prevent disease, while recognizing that some non-manipulatable causal factors (e.g. age, sex, race) might be crucial to our understanding of disease patterns in populations. Usually, extant knowledge and current beliefs form the basis for selecting potential causal factors for study. Thus, I will begin my discussion with a brief review of some concepts of how causal factors might act, and interact, to alter the health status of individuals.

3. Conceptual mechanistic models of causation

The biological details of causation often are unknown, and the statistical measures of association epidemiologists use reflect—but do not explain—the number of ways in which the exposure might cause disease (Hernan, 2004; Hernan and Robins, 2006a). Nevertheless, mechanistic models of causation have been helpful in guiding our research efforts. Because our inferences about causation typically are based on the observed differences in outcome frequency, or severity, between exposed and unexposed subjects (Campaner, 2011), we will examine the relationship between a postulated causal model and the resultant, observed, outcome frequencies. We begin with a description of a simple mechanistic model known as the component-cause model.

3.1. Component-cause model

The component-cause model is based on the concepts of sufficient causes (Rothman, 1976). In this model, a sufficient cause always produces the disease (i.e. if the factor is present, the disease invariably follows). Both experience and formal research have indicated that very few exposures (potential causal factors) are sufficient in and of themselves; rather, different groupings of factors combine and become a sufficient cause. In this context, a component cause is one of a number of factors that, in combination, constitute a sufficient cause. Within each sufficient cause, the factors might be present concomitantly—or they might follow one another in a temporal chain of events (Rothman and Greenland, 2005).

In Table 1, I portray some potential causal relationships between four risk factors (potential causes) and childhood