ELSEVIER

Journal of Clinical Epidemiology 67 (2014) 190−198

Journal of Clinical Epidemiology

Causal inference algorithms can be useful in life course epidemiology

Sacha la Bastide-van Gemert*, Ronald P. Stolk, Edwin R. van den Heuvel, Václav Fidler

Department of Epidemiology, University Medical Center Groningen, University of Groningen, Hanzeplein 1, 9713 GZ, PO Box 30.001, 9700 RB Groningen, The Netherlands

Accepted 31 July 2013; Published online 22 November 2013

Abstract

Objectives: Life course epidemiology attempts to unravel causal relationships between variables observed over time. Causal relationships can be represented as directed acyclic graphs. This article explains the theoretical concepts of the search algorithms used for finding such representations, discusses various types of such algorithms, and exemplifies their use in the context of obesity and insulin resistance.

Study Design and Setting: We investigated possible causal relations between gender, birth weight, waist circumference, and blood glucose level of 4,081 adult participants of the Prevention of REnal and Vascular ENd-stage Disease study. The latter two variables were measured at three time points at intervals of about 3 years.

Results: We present the resulting causal graphs, estimate parameters of the corresponding structural equation models, and discuss usefulness and limitations of this methodology.

Conclusion: As an exploratory method, causal graphs and the associated theory can help construct possible causal models underlying observational data. In this way, the causal search algorithms provide a valuable statistical tool for life course epidemiological research.

Keywords: Causality; Causal graphs; Search algorithms; Life course epidemiology; Metabolic syndrome; Cohort studies

1. Introduction

Chronic multifactorial diseases develop over time, sometimes over the course of decades. In different individuals, they are influenced by different genes, life events, and environmental factors at different points during the life course [1]. The challenge of life course epidemiology lies in unraveling possible causal roles of different variables and estimating the magnitude of their effects. In statistics, causal inference has been a cause for debate for a long time [2−5]. Although the results of statistical analysis are often causally interpreted, statistical theory in itself is rarely concerned with causal inference [6].

The use of randomized experiments is the most commonly accepted method for inference on causality [7]. However, when studying risk factors for chronic diseases that change over time, this approach is usually not applicable. In these cases, observational data can be used to infer on the plausibility and consistency of causal models.

The development of relevant statistical theory has accelerated since the late 1980s. Robins [8] developed a formal theory of counterfactual causal inference, providing a way to deal with direct and indirect effects and time-varying confounders in longitudinal studies. Pearl [4] presented a general theory of causation, thereby bridging the gap between causal connections and statistical associations. Now, testable constraints such as (conditional) independence constraints could be deduced from hypothetical causal models [3−6,9]. The characterization of statistical indistinguishability of causal models (see Section 4) led to the development of search algorithms for equivalence classes of models, among others by Spirtes et al. [6].

In spite of these ongoing developments, only few applications of this theory found their way to the literature yet. An application of one particular causal inference search algorithm (the inductive causation algorithm) was recently used in the context of quantitative genetics [10]. Here, we apply various search algorithms in a life course epidemiological context.

1.1. An example from the PREVEND study

We use an example from the research field that studies temporal relationships between certain medical disorders that increase the risk of developing cardiovascular disease and diabetes [11]. We tried to unravel the mechanisms underlying the temporal associations between body fat and

Conflict of interest: The authors declare that they have no conflict of interest.

* Corresponding author. Tel.: +31-50-361-1148; fax: +31-50-361-4493.

E-mail address: slabastide@xs4all.nl (S. la Bastide-van Gemert).

0895-4356/$ - see front matter
http://dx.doi.org/10.1016/j.jclinepi.2013.07.019

What is new?

This article exemplifies a statistical tool for exploring causal inference with observational data, which can be of great importance in the context of looking for "the mechanism" in life course epidemiology.

Key findings

- Causal inference search algorithms (based on the use of causal diagrams) appear to be a useful tool for exploration of underlying causal mechanisms in life course epidemiology.

What this adds to what was known?

- This article exemplifies a tool for exploration of causal mechanisms underlying observational data in life course epidemiological context and shows its application in the context of obesity and insulin resistance.

What is the implication and what should change now?

- As a rule, in searching for causal mechanism underlying observational data, search algorithms and causal diagrams should be used in addition to more traditional approaches.

glucose levels [12–14], details of which are not completely understood yet [15,16]. We used waist circumference as a marker for visceral adipose tissue and glucose level as an indicator of development of insulin resistance. The data come from the Prevention of REnal and Vascular ENd-stage Disease (PREVEND) study, a prospective, observational, cohort study, designed to evaluate the predictive value of albuminuria for renal and cardiovascular outcomes [17,18]. The participants of the study underwent three examinations with intervals of approximately 3 years. At these time points, denoted as $i = 1, 2, 3$, plasma glucose levels, gluc_i (in mmol/L; log-transformed because of skewness), and waist circumferences, waist_i (in centimeter), were recorded. We also included birth weight (in kilogram, recorded with 0.5 kg precision) and gender (0 = male, 1 = female) [19–22]. We used complete data of 4,081 participants in the analyses. The PREVEND study was approved by the local medical ethics committee and conducted in accordance with the Helsinki Declaration of Research Conduct in Humans. All participants gave written informed consent [22].

1.2. The structure of this article

Terminology and theory of causal graphs is introduced in Section 2. In Section 3, we describe various search algorithms. In Section 4, we apply such search algorithms to the introduced example. In Section 5, we discuss the relative merits of these algorithms. We conclude by explaining the advantages of applying this type of analysis for exploratory research in life course epidemiology.

2. Causal graphs: terminology and theory

Causal models consist of a statistical model and a causal graph. Causal graphs describe the causal relations between the variables in the model. Such a graph consists of vertices denoting variables, connected with edges, which can be oriented by an arrow, denoting a direct causal relationship. A graph is called a directed acyclic graph (DAG) when all edges are directed and the graph contains no feedback loops. The associated terminology and methodology is based on the work as found in the book by Spirtes et al. [6] and Pearl [3,4]. More easily accessible are the overviews on causal inference given by Spirtes [9] and Pearl elsewhere [5].

Fig. 1A shows an example of a causal graph, depicting (completely hypothetical) causal relationships among variables gender, birth weight, and plasma glucose levels (gluc_1, gluc_2, and gluc_3) and waist circumference (waist_1, waist_2, and waist_3), both measured at three time points.

When sampling from a certain population, the vertices in a DAG G represent a set V of random variables and thus are distributed according to some joint probability distribution P. In order for a graph and its associated probability distribution to be considered causal, some assumptions have to be obeyed, such as the causal Markov condition and faithfulness condition.

The causal Markov condition translates to the fact that once we know the parents of a variable in a causal model, knowledge of ancestors does not provide new information. The faithfulness condition implies that all and only the conditional independence relations that hold in P are entailed by the Markov condition in G [6]. If the graph—and therefore its probability distribution P—obeys the causal Markov condition, the joint probability distribution can be decomposed into conditional probabilities involving only the variables and their causal parents (assuming all involved probabilities are nonzero). For Fig. 1B, this yields the following factorization of the joint probability distribution P:

$$\begin{aligned} &P(\text{gender}, \text{birthweight}, \text{waist\_1}, \text{waist\_2}) \\ &\quad = P(\text{gender})P(\text{birth weight}|\ \text{gender}) \\ &\quad \times P(\text{waist\_1}|\ \text{gender}, \text{birth weight}) \\ &\quad \times P(\text{waist\_2}|\ \text{waist\_1}) \end{aligned}$$

A directed graph dictates conditional independence relationships in the joint probability distribution of observed variables under the causal Markov condition. In deriving these consequences, a property called *directed separation* (*d-separation*) is extremely useful. D-separation provides