Some interpretational issues connected with observational studies

D.R. Cox
Nuffield College, Oxford, UK
and
Nanny Wermuth
Chalmers/Gothenburg University, Gothenburg, Sweden

ABSTRACT

After some general comments about observational studies and experiments, three examples of observational studies and their graphical representation are outlined. A distinction between direct and indirect distortion of effects by unobserved variables is drawn. Indirect distortion is illustrated by an example on bladder cancer.

1 Introduction

The classification of studies into experiments, in which in principle the investigator has full control of the situation under investigation, and observational studies in which some key aspects are imposed externally, is of central importance. Observational investigations may themselves be subdivided into prospective (cohort) studies, retrospective studies (often case-control studies) and cross-sectional studies. We assume a common aim, namely that of gaining understanding of a data-generating process in which one or more

outcome variables depend on explanatory variables. While on the whole this task gets progressively more difficult as we move through the sequence of study types listed above, all may face interpretational difficulties. Further, combinations of study types are common; case-control studies may be embedded in randomized experiments, for instance. On the other hand some observational studies may have a reasonably clear-cut interpretation. There are issues of data quality and completeness and adherence to study protocol by no means excluded in randomized experiments.

In this paper we consider a primary way in which the conclusions from observational studies may be distorted, namely by the presence of unobserved variables which may seriously amend the apparent effect of those explanatory variables that are present.

2 Formulation

We suppose that a number of variables are measured on study individuals, at least one of the variables being an outcome or response variable and the others being explanatory. For any two variables considered in isolation either one is explanatory to the other considered provisionally as a response or the two are on equal footing such as symptoms of a disease or the items in a multiple choice test. Variables are represented by the nodes of a graph. If a response variable is considered to be statistically dependent on a given explanatory variable then a directed edge is drawn, with an arrow starting at the explanatory variable and pointing to the response. If the two variables are on equal footing any edge joining them is undirected and represents an association with the direction of dependence unspecified; they may be joint response variables such as systolic and diastolic blood pressure.

Explanatory variables may be classified in various ways for instance into

primary risk factors or quasi-treatments. These are variables that would be the treatments in a randomized experiment were such an experiment feasible. Another kind of explanatory variables are background or intrinsic variables that essentially define the study individuals but are not conceivable treatments in the context in question, i.e. are not subject to real or conceptual intervention. The role of the intrinsic variables is partly to ensure that the levels of the explanatory variables of primary interest are compared under similar conditions and partly to assess possible interactions. For further discussion of these issues and the relation with causality, see Cox and Wermuth (2004).

3 Three examples

We outline three examples which illustrate various aspects of observational studies.

Example 1. Fig. 1 represents a tentative ordering of the variables in a cross-sectional observational study of the relation between knowledge about the disease and success at controlling the glucose level in diabetic patients at University Hospital, Mainz.

A key issue here is that knowledge and success were measured essentially simultaneously and in the first place could be considered on an equal footing. The not directly testable hypothesis that knowledge was explanatory to success, plus the isolation of an important interaction led to a provisional interpretation, confirmed by a second study. Fig. 2 summarizes statistically dependent variables, relevant for the two response variables of main interest. For more details, see Cox and Wermuth (1996, Section 6.2).

Example 2. The cause of bovine tuberculosis, a mycobacterium (*M. bovis*) lives both in cattle and in wildlife, in the UK in the badger (*meles meles*).

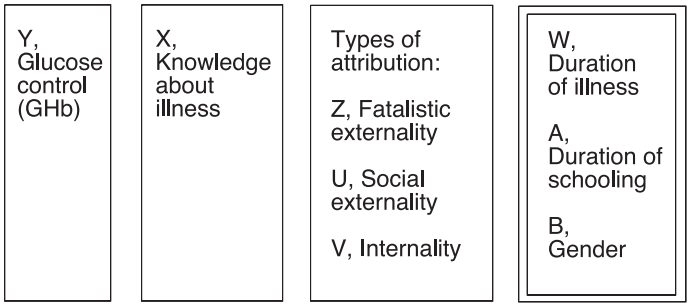

Figure 1: Assumed dependence chain for variables of the diabetes study. Primary response: glucose control, Y; secondary response: knowledge about illness, X; three intermediate variables on equal footing: Z, U, V, different ways of attributing success of disease control; context variables: duration of illness, W, duration of schooling, A, and gender, B.

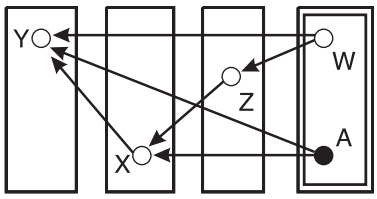

Figure 2: Directly explanatory variables for response Y of the diabetes study are X, W, and A; indirectly explanatory for Y are Z, W, and A.

The organism can be genetically typed in particular into so-called spologitypes, there being eight such types relatively common in one part or another of the UK. In ten regions of south west and central England, infected cattle and badgers were spologityped. There is an extremely high association between the observations in the two species. Typically in any one region, one spologitype accounted for 80 to 90 per cent of the cases in both species. Up to a point, a strong interpretation can be drawn even though this is a purely observational study. The main possibilities are first that there is cross-infection between the two species, see Fig. 3(a), secondly that there is a common unobserved direct confounder, the ecological character of the different regions, see Fig. 3(b). Thirdly, the two species do not cross-infect but that there is a

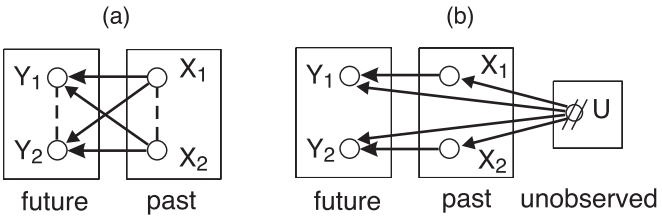

Figure 3: (a) Pair (Y_1, Y_2): infection rates in two species; (X_1, X_2): the infection rates at earlier time. Dependence between species explained by cross-infection. (b) Dependence between species explained by unobserved background variable U.

further common source of infection, represented also by Fig. 3(b) but with a different interpretation of U. There are strong external reasons for rejecting the last two possibilities. This leaves open, however, the crucial question of whether the cross-infection is in one or in both directions. To address this, the data must be divided temporally and a suitable probability model fitted, the interpretation of which is, however, rather more tentative because direct observation at the level of an individual animal is not possible.

Example 3. An important aspect of successful surgical treatment is the change in quality of life as perceived by the patient before and after treatment. A stepwise data-generating process shown in Fig. 4 is for determinants of quality of life after removal of the bladder because of a tumour. There are four quantitative, directly and indirectly explanatory variables for physical quality of life after surgery, Y, and one binary variable, A, which captures whether the bladder substitute leads to continent or incontinent urine diversion. The key question is here, how Y depends on A and whether this dependence can be estimated without distortions when both U and V are unobserved.

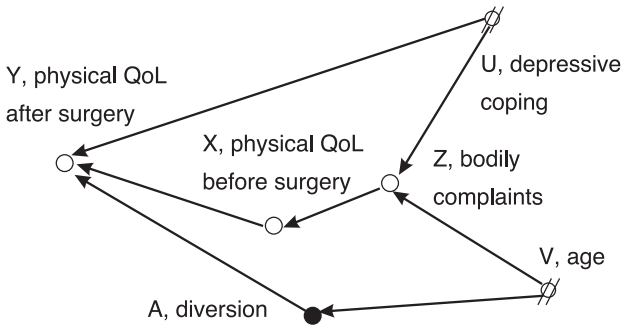

Figure 4: Potential generating process for physical quality of life after surgical removal of the bladder, in male patients with a bladder tumour. With both U and V unobserved, there is no direct confounding, no over- or under- conditioning but the generating dependence of physical quality of life after surgery, Y, on the type of diversion, A given X, U is distorted in the conditional dependence of Y on A given X or given X, Z.

4 Types of distortion

We distinguish two broad kinds of distortion that arise from the absence of important variables in the analysis, typically because they are unobserved. The simplest form, called direct confounding, arises when there is a background variable that is directly explanatory both to the main explanatory variable of concern and to the response. The second listed interpretation of Example 2 illustrates this, ecological character of the region being the unobserved confounder. A more general form of direct confounding arises when one or both pathways of dependence from the background variable are via other variables present in the data generating process but themselves unobserved. Occasionally, in special more complicated cases, the distortion induced this way can be corrected by the use of instrumental variables. While randomization in an experiment protects against direct confounding, the possibility of important interactions between the treatment and unobserved variables remains.

The second type of confounding, we call indirect, is also an issue in intervention studies, even when individuals are randomly allocated to treatments. It is illustrated for an observational study with Example 3, and discussed in the next section.

There are further situations in which dependencies appear distorted when conditional dependencies are estimated from a set of data with a reduced number of variables compared to the generating process. We do not discuss them in detail here since they are absent in the given three examples. One is under-conditioning. It arises by omitting from an analysis those variables that are intermediate between the explanatory variable and an outcome of primary interest. The other is over-conditioning. Its simplest form arises by using as an explanatory variable to the outcome of primary interest a variable which is, in fact, itself a response of this outcome.

An important aspect of indirect confounding is that it can be present if there is no direct confounding, no over- and no under-conditioning. It arises by a combination of omitting some relevant context variables and of conditioning on the remaining important explanatory variables. The possible presence of indirect confounding can be detected with the help of graphical criteria. If detected, distortions due only to indirect confounding can be corrected if all associations are of linear form. For other systems it is still unknown how to correct.

5 Example of indirect confounding

Omission of the two variables U and V in Fig. 4 does not lead to direct confounding of the dependence of Y on A, the estimation of which is a primary objective of the study. This is because neither is a directly explanatory variable to both outcome Y and the directly explanatory variable A. The

omission does however distort the dependence of Y assessed by regression analysis on the remaining observed three variables. This is because absence of U induces an association between Y and Z, whereas the omission of V induces an association between Z and A. There is now a distorting pathway of association via Z between the explanatory variable A and outcome, Y. It can be shown, however, remarkably, that for linear systems an appropriate analysis of least squares regressions and of covariance matrices of residuals allows correction for this distortion, provided its presence is detected. This is discussed in detail by Wermuth and Cox (2007) using a mixture of graph theory and matrix arguments, see also Wermuth, Wiedenbeck and Cox (2006).

REFERENCES

Cox, D.R. & Wermuth, N. (1996). *Multivariate Dependencies – Models, Analysis and Interpretation.* London: Chapman & Hall.

Cox, D.R. & Wermuth, N. (2004). Causality: a statistical view. *International Statistical Review*, **72**, 285-305.

Wermuth, N., Wiedenbeck, M. & Cox, D.R. (2006). Partial inversion for linear systems and partial closure of independence graphs. *BIT, Numerical Mathematics*, **46**, 883-901.

Wermuth, N. & Cox, D.R. (2007). Distortion of effects. Submitted.