

# Against Popperized Epidemiology

M. JACOBSEN<sup>1</sup>

Jacobsen, M. (Institute of Occupational Medicine, Roxburgh Place, Edinburgh, EH8 9SU, Scotland). Against Popperized Epidemiology. *International Journal of Epidemiology*. 1976, 5: 9–11.

The recommendation that Popper's philosophy of science should be adopted by epidemiologists is disputed. Reference is made to other authors who have shown that the most constructive elements in Popper's ideas have been advocated by earlier philosophers and have been used in epidemiology without abandoning inductive reasoning. It is argued that Popper's denigration of inductive methods is particularly harmful to epidemiology. Inductive reasoning and statistical inference play a key role in the science; it is suggested that unfamiliarity with these ideas contributes to widespread misunderstanding of the function of epidemiology. Attention is drawn to a common fallacy involving correlations between three random variables. The prevalence of the fallacy may be related to confusion between deductive and inductive logic.

In a recent paper (1) Dr. Carol Buck protests that epidemiologists should not be regarded, by themselves or by others, as data-gathering adjuncts to the medical-scientific community. She argues that creative epidemiology requires the formulation of refutable hypotheses and that epidemiologists should involve themselves more in the "exciting process of scientific deduction". Her view is that the achievement of these laudable objectives would be promoted if epidemiologists were to adopt the ideas of Karl Popper. Yet Dr. Buck herself and Davies (2) refer to some of Popper's predecessors who understood and emphasized the importance of attempting to refute scientific hypotheses. Dr. Davies asks 'why pick on Popper?'

A major distinguishing feature in Popper's philosophy is his insistence that his ideas are 'directly opposed to all attempts to operate with the ideas of inductive logic' (3). Dr. Buck accepts this view enthusiastically. I agree with Dr. Davies. This is not a helpful way to free epidemiologists from the misunderstanding that their task is merely to 'gather data from which other scientists will build hypotheses' (1). On the contrary, I have argued elsewhere (4) that because good medical practice relies heavily on the deductive art of diagnosis, many physicians appear to have great difficulty in appreciating the rules and discipline of epidemiological research, since here the inductive method of reasoning is paramount. Unfamiliarity with inductive reasoning is one of the barriers obstructing an understanding of the role of statistical

inference in epidemiology; and statistical inference lies at the heart of epidemiology. Why? Because epidemiology is concerned largely with the study of observations which are subject to random fluctuations.

Dr. Buck warns against 'an excessive recourse to statistical devices for tidying up epidemiological data'. It is useful to draw the attention of epidemiologists to Cox's objection to equating statistical inference with data summarization (5). Cox reminds us that in statistical inference 'an essential element is the uncertainty involved in passing from the observations to the underlying population'. In Cox's formulation it is the element of uncertainty which makes the inference statistical; the *direction* of the argument, *from* the observations *to* the (statistical) population sampled, is characteristic of the inductive method. Note that if this inductive step is taken with a proper regard to the logic underlying (for instance) a test of statistical significance (the logic includes its restricted relevance to the population sampled), then the induction leads naturally to what Dr. Buck calls the 'deductively more powerful' reason for replication in epidemiology: to identify circumstances under which a previously observed association does not exist.

In common with some other epidemiologists Dr. Buck appears to regard applied statistics essentially as a set of numerical 'devices'. However, in epidemiology as elsewhere, the application of probability theory to real data embraces more than significance testing, standardization and data summarization. For instance, an increasingly important part of modern epidemiology is directed to determining the relationship between the dose of a suspected or

<sup>1</sup> Institute of Occupational Medicine, Roxburgh Place, Edinburgh EH8 9SU, Scotland.

known environmental pollutant and the degree of response in human populations. The solution to this quantitative problem requires, among other things, statistical point estimation (of the regression coefficient for instance) and interval estimation. The latter may involve controversial inferences about the distribution of the unknown regression coefficient being estimated or the interval containing the estimate itself, at a fixed probability level. The fact that significance testing may be relatively unimportant in such situations, and that it may be difficult to force the pertinent research questions into a formal hypothesis implying a refutable prediction, does not convince me that this activity therefore lies outside the boundary of science. Of course, once the dose-response relationship has been estimated then this provides the basis for subsequent predictions which are refutable in principle. Indeed I have suggested two ways whereby such epidemiologically derived predictions can be tested in practice (6). The important point to note however is that the initial estimation of the dose-response relationship requires inductive arguments of statistical inference, from the sample to the population parameters of interest. Deductions and hypotheses follow.

The discussion in the two preceding paragraphs illustrates Fisher's point that 'deductive arguments are, in fact, often only stages in an inductive process' (7). Popper (3) would be unmoved by these considerations. He distinguishes between inductive processes, which he relegates to the realm 'psychology of knowledge', and the deductive method, which he considers as part of the 'logic of knowledge'. He then asserts that he is concerned only with the latter. ('There is no need even to mention induction.') Even if this view is acceptable to one who is dedicated specifically to the search for a wholly abstract theory of knowledge, it is difficult to see why Dr. Buck commends it to epidemiologists who are involved directly in the practice of an empirical science which leans heavily on inductive methods.

For her 'exercise in Popperian epidemiology' Dr. Buck (1) reformulates the 'infectious hypothesis' concerning cancer of the uterine cervix. Her hypothesis asserts, *inter alia*, a negative correlation between the risk of cervical cancer and progesterone levels. Dr. Buck then proceeds to discuss testable predictions from the hypothesis. One of them states 'A history of repeated spontaneous abortion should be associated with cervical cancer, because susceptibility to abortion is to some extent related to low levels of progesterone.'

This prediction does not follow from the hypothesis (quite apart from the careful caveats made by Dr. Buck herself). Let  $X$ ,  $Y$  and  $Z$  represent the random variables of cervical cancer incidence, progesterone level and spontaneous abortion incidence respectively; let  $r_{XY}$ ,  $r_{ZY}$  and  $r_{XZ}$  be the correlation coefficients between  $XY$ ,  $ZY$  and  $XZ$ . Dr. Buck's prediction can be written as

$$\left. \begin{array}{l} \text{'if } r_{XY} < 0 \\ \text{and } r_{ZY} < 0 \\ \text{then } r_{XZ} > 0'. \end{array} \right\} \quad [1]$$

The assertion [2] is a *non-sequitur* not uncommon among epidemiologists. The difficulty seems to be connected with confusion between deductive and inductive methods of argument. Only the simplest algebra or common (deductive) sense suffices to show that

$$\left. \begin{array}{l} \text{if } x = a_0 + a_1 y \\ \text{and } z = b_0 + b_1 y, \end{array} \right\} \quad [3]$$

where  $x$ ,  $y$  and  $z$  are real variables;  $a$ ,  $b$  are constants; and  $a_1$ ,  $b_1$  are non-zero;

$$\text{then } x = c_0 + c_1 z, \quad [4]$$

where  $c_1 = a_1/b_1 \neq 0$ . Moreover, if  $a_1$  and  $b_1$  have the same algebraic sign, then  $c_1 > 0$ . Now it is easy to write down linear regression models [5] embodying the left-hand sides of the inequalities [1] and [2], and they look deceptively similar to equations [3] and [4].

$$\left. \begin{array}{l} E(X|y) = \alpha_0 + \alpha_1 y \\ E(Z|y) = \beta_0 + \beta_1 y \\ E(X|z) = \gamma_0 + \gamma_1 z \end{array} \right\} \quad [5]$$

If  $\alpha_1$  and  $\beta_1$  are known to be negative it does not follow that  $\gamma_1$  is positive.

Why is the deductive algebraic syllogism of [3] and [4] not applicable to the models [5]? Because the models refer to *random* systems; they are not simple linear equations. The coefficient  $\gamma_1$  is an unknown population parameter. It may be estimated from appropriate data using inductive, statistical arguments. The degree of uncertainty in the estimate would be related to the variance of the random variable  $X$ . (In reality, we cannot 'know' that  $\alpha_1$  and  $\beta_1$  are both negative; but, given enough data, and using inductive arguments, we may have great confidence that this is true.)

The argument from [1] to [2] is false; but it can

be shown that there are conditions constraining  $r_{xz}$  to be positive when  $r_{xy}$  and  $r_{zy}$  have the same algebraic sign. The inequalities

$$1 > r_{xy}r_{zy} > 0 \tag{6}$$

$$r_{xy}r_{zy} - [(r_{xy}^2 - 1)(r_{zy}^2 - 1)]^{1/2} > 0 \tag{7}$$

are sufficient conditions to satisfy [2]. The necessary condition for  $r_{xz} \leq 0$ , given [6], is that the left-hand side of [7]  $\leq 0$ . Thus if  $r_{xy}$  and  $r_{zy}$  both exceed  $+\sqrt{0.5} \approx 0.71$  (or are both less than  $-0.71$ ) then  $r_{xz} > 0$ . However, there may be a wide range of paired values ( $r_{xy}, r_{zy}$ ) with the same algebraic sign and  $r_{xz} \leq 0$ . For example, if  $r_{xy} = -0.8$  and  $r_{zy} = -0.4$ , then non-positive values of  $r_{xz}$  may be as low as  $-0.2299$ ; certainly  $r_{xz}$  may be zero. Or, if  $r_{xy} = r_{zy} = 0.5$ , then  $r_{xz}$  may be as low as  $-0.5$ .

For those who prefer a non-theoretical refutation of the assertion hypothesising that 'the inequality [2] follows logically from [1]' it is sufficient to find any data set where the assertion is contradicted. The contrived data below meet the purpose.

x	y	z
0	4	6
1	5	1
2	0	4
4	6	2
5	1	5
6	2	0

$$r_{xy} = r_{zy} = -0.29; r_{xz} = -0.43.$$

Fisher (7) has remarked on a 'tendency to impose on inductive thought the conventions and preconceptions appropriate only to deductive

reasoning.' Popper states (3): 'Believers in probability logic may try to meet my criticism' (of inductive methods) 'by asserting that it springs from a mentality which is "tied to the framework of classical logic", and which is therefore incapable of following the methods of reasoning employed by probability logic. I freely admit that I am incapable of following these methods of reasoning.' The concluding admission is deplorable even if it was intended facetiously.

I agree with Dr. Buck that epidemiology would benefit if more of the data gathering were preceded by the formulation of hypotheses whose relevance to science will have been deduced from previously established knowledge. The element of deduction in the formulation of the hypotheses does not make induction irrelevant, any more than the essential inductive inferences from observed data negate the importance of prior and subsequent deductions.

REFERENCES

- (1) Buck, C.: Popper's philosophy for epidemiologists. *International Journal of Epidemiology* 4: 159, 1975.
- (2) Davies, A. M.: Comments on 'Popper's philosophy for epidemiologists' by Carol Buck. *International Journal of Epidemiology* 4: 169, 1975.
- (3) Popper, K. R.: *The Logic of Scientific Discovery*. Sixth impression (revised), pp. 30, 315, 265. London: Hutchinson, 1972.
- (4) Jacobsen, M.: A science which ought to be honourable. *Transactions of the Society of Occupational Medicine* 21: 38, 1971.
- (5) Cox, D. R.: Some problems connected with statistical inference. *Annals of Mathematical Statistics* 29: 357, 1958.
- (6) Jacobsen, M.: Evidence of dose-response relation in pneumoconiosis (2). *Transactions of the Society of Occupational Medicine* 22: 88, 1972.
- (7) Fisher, R. A.: *Statistical Methods and Scientific Inference*. Second edition, pp. 109, 110. Edinburgh: Oliver and Boyd, 1959.