Studies in History and Philosophy of Biological and Biomedical Sciences 40 (2009) 302–311

Contents lists available at ScienceDirect

Studies in History and Philosophy of Biological and Biomedical Sciences

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

Causation and models of disease in epidemiology

Alex Broadbent

Department of History and Philosophy of Science, University of Cambridge, Free School Lane, Cambridge CB2 3RH, UK

ARTICLE INFO

Article history:
Received 14 January 2009
Received in revised form 11 May 2009

Keywords:
Epidemiology
Monocausal
Multifactorial
Disease
Risk factor
Cause

ABSTRACT

Nineteenth-century medical advances were entwined with a conceptual innovation: the idea that many cases of disease which were previously thought to have diverse causes could be explained by the action of a single kind of cause, for example a certain bacterial or parasitic infestation. The focus of modern epidemiology, however, is on chronic non-communicable diseases, which frequently do not seem to be attributable to any single causal factor. This paper is an effort to resolve the resulting tension. The paper criticises the monocausal model of disease, so successful in the nineteenth century. It also argues that a multifactorial model of disease can only be satisfactory if it amounts to more than a mere rejection of the monocausal model. A third alternative, the contrastive model, is proposed and defended on the grounds that it links the notions of disease and of general explanation, while avoiding the philosophical naiveties and practical difficulties of the monocausal model.

1. Introduction

Two conceptual questions currently face epidemiology, both relating to causation. First, how should it handle certain *diseases*, which appear to be etiologically more complex than the infections and deficiencies by which epidemiology made its name? In particular, chronic non-communicable diseases (CNCDs) account for a larger proportion of deaths, at least in the industrialised world, than they did in 1900 (Rockett, 1999, p. 8), and attract more epidemiological attention. Yet they often do not seem susceptible to definition in terms of any one causative agent: their etiology is typically complex.

Second, how should epidemiology respond to newly identified *causes* of disease? Epidemiology has moved beyond obvious environmental causes of illness (such as prolonged extreme cold) and uncovered increasingly complex and sometimes surprising environmental causes of disease. And in place of the old notion of a 'constitution', the discipline has had to grapple with a newly discovered category of cause: genetics. The increased depth and complexity of our knowledge of both genetic and environmental determinants of health places pressure on aspects of the conceptual framework of epidemiology: in particular, on the way it thinks about disease causation.

Devising a conceptual framework for thinking about disease causation has proved astonishingly difficult. On the one hand, the early history of epidemiology appears to attest to the power of insisting that every disease has one cause that is necessary and, in limited circumstances, sufficient for the disease. I call this way of thinking about disease etiology the *monocausal model* of disease. This model suits infectious diseases such as TB and cholera well, along with parasitic infestations and diseases of deficiency. On the other hand, the monocausal model is a terrible fit for CNCDs such as lung cancer or diabetes. It is theoretically possible that a condition like diabetes has a single necessary and, in some circumstances, sufficient cause, which we have not yet discovered. But surely, it is also a theoretical possibility that there is no cause for diabetes satisfying that description. And even if there is, it is not clear how insisting that there *must* be such a cause helps us achieve public health or clinical goals, if we don't know what it is. The causes that we are able to identify are causal risk factors: neither necessary nor sufficient. These are all we have to work with. Accordingly, a view of disease as *multifactorial* now dominates epidemiology. But this is not an entirely happy situation, because it fails to mark what looks like a real etiological difference between diseases like cholera and conditions like lung cancer. The monocausal model

E-mail address: abb24@cam.ac.uk

1369-8486/$ - see front matter © 2009 Elsevier Ltd. All rights reserved.
doi:10.1016/j.shpsc.2009.09.006

has had some striking successes in the history of epidemiology, and these successes are left unexplained by the mere assertion that disease causation is multifactorial. Unless we can explain the successes of the monocausal model in terms of modern multifactorial thinking, there is a risk of throwing the baby out with the bathwater.

In this paper I want to address the tension that arises between monocausal and multifactorial models of disease. Both are, to some extent, rational reconstructions of positions that are implicit in the epidemiological literature. There have been very few (if any) attempts to lay out these two ways of thinking about disease causation, in a fully explicit and philosophically rigorous manner. Accordingly there is a danger of attacking straw men. It should be understood that these models of disease, as I state them, are attempts to make explicit ways of thinking that are implicit: so I refrain from attributing the result of this exercise as an opinion of any historical or contemporary figure. Nevertheless, I do think—and will argue—that these ways of thinking are present in various extant efforts to conceptualise disease causation. Once I have assessed the strengths and weaknesses of these two models of disease, I will propose a 'contrastive' model, which attempts to preserve the strengths of the monocausal model within a multifactorial framework.

2. The monocausal model

Perhaps the closest medicine has come to an explicit statement of the monocausal model of disease is Koch's postulates. However, contrary to the impression sometimes conveyed, even these postulates have no authoritative statement. Koch's own work does not define the postulates authoritatively (see Evans, 1993, Ch. 2, for several versions and discussion). Moreover, the postulates are shot through with practical concerns; they do not constitute a philosophical model of disease. This reflects the fact that, for much of his professional life, Koch was interested primarily in a particular *kind* of cause—microbial infection. Consider this statement of the postulates:

> In order to prove that tuberculosis is a parasitic disease caused by the invasion of the bacilli and primarily influenced by the growth and proliferation of the latter, the bacilli had to be isolated from the body and cultivated in pure culture until devoid of all adherent products of disease originating from the animal organism; and, finally, through transfer of the isolated bacilli to animals, the same clinical picture of tuberculosis as is obtained empirically by the injection of naturally developed tuberculosis material had to be produced. (Koch, 1938, p. 861)

These steps are obviously informed by various commitments, concerning the existence of bacilli, which cause disease by invading organisms, but which can be grown outside the organisms, and so on. However, I contend that once this is stripped away, we can discern an independent, *conceptual* commitment. We can identify this by asking: what is the *purpose* of these steps? What is so special about this procedure?

The answer lies in what Koch wanted to prove, 'that tuberculosis is a parasitic disease caused by the invasion of the bacilli and primarily influenced by the growth and proliferation of the latter'. In earlier work on anthrax he claims that 'each disease is caused by one particular microbe—and by one alone. Only an anthrax microbe causes anthrax; only a typhoid microbe can cause typhoid fever' (Koch, 1876; quoted in Evans, 1993, p. 20). Koch's postulates do not prove causation simpliciter (though this is sometimes how they are presented). In fact, they offer a practical heuristic for proving the existence of *a particular causal structure*. Once we strip away the details specific to microbial infections, Koch's postulates seek to establish two things. First, that the disease in question does not occur in the absence of the putative cause (implied by the requirement that the organism into which the cause is introduced is healthy beforehand). Second, that the disease in question does occur when the putative cause is present, under certain circumstances. (The postulates are in part an effort to specify these circumstances.) In other words, the postulates seek to establish that a certain cause is both causally necessary and causally sufficient, in specified circumstances, for the disease.[1]

We can summarise this requirement as follows. A kind of event C is the requisite kind of cause for disease D if, and only if:

(i) a C-event is a cause of every case of D;
(ii) given certain circumstances, a C-event is not a cause of any $\neg D$ event (i.e. other diseases or good health).

The monocausal model of disease is the view that every disease has a cause satisfying (i) and (ii). The first condition states that C is causally necessary for D. The second states that, in certain circumstances, C is sufficient (because under those circumstances, there is no situation in which C occurs, and causes $\neg D$—or less obscurely, fails to cause D).[2]

Why is this a *monocausal* model of disease? Whence the suggestion that there is *only* one cause satisfying (i) and (ii)?—That suggestion arises because, between them, (i) and (ii) entail (or near enough) that *at most* one cause will satisfy them. To see this, suppose that two kinds of cause, C_1 and C_2, are proposed with respect to disease D. If there is any case where C_1 is present and C_2 is absent, then either (i) or (ii) will be violated with respect to C_1, depending on whether D is present or absent in that case. (And vice versa for C_2.) And if C_1 and C_2 are universally present or absent together, then the chances are that this is no mere coincidence, and that they are related either as cause to effect or as effects of a common cause. If the former, we have grounds to consider them parts of the same cause; and if the latter, it will in principle be possible to bring about cases where C_1 occurs without C_2, or vice versa.[3] Then, again, either (i) or (ii) will be violated, depending on whether D is present or absent.

This is how a commitment to the existence of necessary and circumstantially sufficient causes leads to the otherwise puzzling commitment to the existence of just *one* such cause, from which the monocausal model derives its name. Events of kind C will not be the only causes of D, but they will be the only kind satisfying (i) and (ii). To reinterpret Koch's example, typhoid fever may have lots of different causes—the causal history of one case may feature events of kinds that do not feature in another. But there is only one kind of cause of typhoid fever that satisfies conditions (i) and (ii). If

[1] Koch's commitment to this model is not to be confused with his commitment to germ theory. Believing that diseases are caused by germs does not logically compel one to believe that each disease is caused by *just one kind* of germ. That is a logically independent commitment.

[2] K. Codell Carter seeks to analyse the distinctive stance of modern medicine in terms of a 'universal, necessary cause' (Carter, 2003, Ch. 1). However, this analysis is not made entirely precise. Moreover it is vulnerable to an irritating but persistent objection. The presence of oxygen is a universal, necessary cause of cholera; it is present in every case, and without it there would be no cholera, as the virus relies on its host being alive. Can we therefore classify cholera as oxygen-disease? The obvious response is that oxygen is also a cause of good health. Once we try to exclude causes of good health, as I do in (ii), we effectively introduce a limited sufficiency condition. As long as we assume that a dead person is not an ill person, (ii) excludes oxygen as the classificatory cause of cholera.

[3] As long as there is some event or condition that is causally necessary for C_1 but not C_2, or vice versa, it will be possible for one to occur without the other. To claim that there is no such event or condition would be to assert that no physically possible intervention could prevent C_1 without also preventing C_2; in which case the distinctness of C_1 and C_2 may be doubted, at least in a world like ours.