Causation and Models of Disease in Epidemiology

Alex Broadbent

Department of History and Philosophy of Science, University of Cambridge

This is a post-peer-review but pre-copy-edited version of a paper that has been accepted for publication in the journal Studies in the History and Philosophy of the Biological and Biomedical Sciences.

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, e.g. 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 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.

(Robert Koch, 1882, 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” (R. Koch, 1876; cited 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 $C_i$, 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 typhoid bacteria satisfy (i) and (ii), then no other cause can satisfy both (i) and (ii), for the reasons just given.

In fact it may not be possible to state explicitly the circumstances for sufficiency mentioned in (ii). The patient would have to be well supplied with oxygen; the Earth would need to continue a peaceful orbit; hungry lions would need to be absent from the patient’s vicinity. Strictly speaking, the statement would have to be extremely extensive, excluding every way that $D$ might be prevented when $C$ occurs. Difficulties with *ceteris paribus* clauses are not specific to the medical context, however: they pose a problem across the philosophy of the special sciences. But that does not mean that we can say nothing useful about the circumstances in which $C$ is causally sufficient for $D$. When Koch’s postulates insist that the causative agent should be able to induce the symptoms anew after being isolated and regrown in a pure culture, they are in effect operationalising the notion of causal sufficiency present in (ii), by specifying the circumstances in which $C$ must be causally sufficient for $D$. It is interesting that Koch’s actual specification of circumstances proved rather too restrictive even for infectious diseases; viruses cannot be grown in a lifeless culture,
and even some bacteria (e.g. those causing leprosy and syphilis) are difficult to grow (Evans, 1993, p. 41). And diseases of deficiency, such as scurvy, require an even more far-reaching revision.

The monocausal model of disease, as I have laid it out, states that every disease has a necessary and circumstantially sufficient cause, and by its logic ensures (or near enough for practical purposes) that it will have at most one such cause. This leads to an obvious question: it is true that every disease has exactly one cause meeting (i) and (ii)? Do all, or indeed any, diseases satisfy the monocausal model?

Not if diseases are classified according to their symptoms. There is no guarantee that symptoms or sets of symptoms will share common causal factors. And frequently diseases, as defined in the early nineteenth century, did not satisfy the monocausal model. As K. Codell Carter puts it:

If hydrophobia is an extreme inability to swallow, it really can be caused by blows to the throat, by psychological factors, or by the bites of rabid dogs… As long as diseases were defined in terms of symptoms, different episodes of any one disease simply did not share a common necessary cause. And no research, however brilliant, can find what isn’t there.

(Carter, 2003, p. 37)

The monocausal model of disease should not be interpreted as making an empirical claim about diseases. Interpreted as a descriptive claim about diseases as they are already defined, the monocausal model would have been simply false when it was initially proposed: the number of recognised diseases satisfying it would have been close to zero, as Carter points out. The monocausal model ought to be interpreted as a normative model. It tells us that we ought to identify diseases so that they satisfy the model. Even in the abstract, it is easy to see that this will probably require reclassification. Recognised conditions might have to be carved up; separate conditions might be amalgamated; and some conditions may remain largely undisturbed apart from the exclusion of a handful of cases. All this, in order to individuate a disease satisfying the monocausal model. And this is exactly what happened in the nineteenth century, according to Carter. As he puts it, “The only way of characterizing diseases so that each disease will have a universal necessary cause is by defining them in terms of their causes” (Carter, 2003, p. 38).
What is the point of this classificatory exercise? Why should we define disease as the monocausal model does, and then reclassify diseases so that they fit it? Carter makes a convincing case for the historical importance of this way of thinking about disease and classifying diseases, and interested readers should consult his work (Carter, 1994, 2003). Carter suggests that the “etiological standpoint”, that is, the reclassification of diseases in terms of causes satisfying certain conditions, was instrumental in dramatic medical advances. He emphasises the conceptual nature of the move, and argues that it predates – and may have even led to – the development of germ theory. He also shows how the stance is useful beyond germ theory: for example, in identifying diseases of deficiency such as scurvy.

If Carter’s analysis is correct, this should prompt us to find out why the monocausal model was so successful. The central theoretical advantage of the monocausal model of disease, is that it encourages general explanation. Semmelweis’s work on childbed fever is not only an exemplar in epidemiology, it is also a classic case study in the philosophy of science, used by several philosophers to develop theories of explanation (Hempel, 1966; Lipton, 2004). Semmelweis was faced with differing rates of childbed fever in two hospital wards, and he was sceptical of existing “miasm” explanations for the disease because they did not explain the difference between the two wards. To explain the difference, sought a causal difference between the two wards: a cause that was present to a greater extent in the ward with higher incidence, and to a lesser extent in the other. This is an particular instance of the monocausal model as I have described it. Semmelweis effectively sought a cause satisfying the monocausal criteria: a cause that was present in cases of disease and absent otherwise.

Identifying a cause of this sort, when it is common to a broad class of cases of ill health, will give us a general explanation of why those cases occur – general in that it is true for all of them, and independent of the historical details of each particular case. This sort of explanation, and its value, are vividly illustrated by some famous interventions: Semmelweis’s introduction of post-autopsy hand washing in the Vienna hospital; and Snow’s suggestion to remove the handle from the Broad Street pump in London. In each case, a single intervention prevented a large class of cases. This sort of episode is strong evidence for the pragmatic value of this way of thinking about disease.
3. Problems for the monocausal model

Even if successes can be claimed for it, the monocausal model as I have sketched it suffers from a number of objections. Starting with the most obvious, the monocausal model provides no justification for its restriction on the number of causes by which a disease may be defined. Why should there not be states of ill health where two, or a hundred and one, kinds of cause are used for classification? To take a simple example, swine influenza as it occurs naturally is caused by the synergistic action of a bacterium and a virus (Evans, 1993, p. 41). Neither will meet both (i) and (ii), if instances of the virus may exist in healthy animals without the bacteria, and vice versa. To fit the monocausal model, swine fever must be classified as infection by either the virus or the bacteria, and then the presence of the other must be specified in the circumstances for causal sufficiency mentioned in (ii). So either swine fever as it is usually classified is not a disease, or else the monocausal model is wrong, imparting advice which sound classificatory practice is right to ignore.

This theoretical problem has become a pressing practical concern in modern epidemiology. Many modern epidemiologists feel that it is unhelpful to presume that they are seeking causes of disease that are in any sense necessary or sufficient. The monocausal model does not seem to be fruitful any more, even though it may have been methodologically fruitful at first. It fits infectious diseases, parasitic infestations and nutritional deficiencies, but it does not seem well-suited to CNCDs, whether they are to be explained by environmental or genetic factors, or some combination. It is hard to see how insistence that, for example, bowel cancer must be defined by reference to a single causal factor could further our understanding of that condition, when so many factors seem to be relevant. It is surely this more than any philosophical objection that has led many working epidemiologists to adopt multifactorial models of at least some diseases, and hence to reject the monocausal model as a fully general model of disease.

Nor is it clear that the monocausal model can restrict its attention to amenable sorts of diseases (e.g. providing a model of infectious diseases only). There may be links between these and less amenable diseases, especially CNCDs. For example, hookworm infection accounts for a significant percentage of the attributable risk for
anaemia, and similar links are notable for many of the neglected tropical diseases (Hotez & Daar, 2008). The existence of links of this sort suggests that, even if some diseases fit the monocausal model, their medical consequences may include conditions which do not. This casts doubt on the medical value of defending the model by restricting it to a subclass of diseases, even if that is a philosophical option.

Another question for the monocausal model concerns the role played by definition. That role is central; unless we are extraordinarily lucky, it is not likely that symptomatic groupings of disease will yield causes satisfying the model. So diseases will have to be reclassified. The monocausal model does not tell us how to do this, beyond the satisfaction of (i) and (ii). There will generally be more than one reclassification of symptomatically grouped illnesses that will satisfy these conditions.

The point can be brought out with an example. Suppose infection by tubercle bacilli is, by definition, necessary for a case of tuberculosis. The bacilli are not the only causally necessary kind of event for tuberculosis. A lack of inherited immunity is also necessary. But the disease is usually defined by reference to the tubercle bacilli, and not to the lack of immunity; instead, that lack of immunity is a part of the background against which the bacilli are sufficient for TB. Compare phenylketonuria, which is usually characterised as a genetic disorder; sufferers lack the enzyme necessary to metabolise phenylalanine (an amino acid). To suffer any ill effects of this disorder, the ingestion of phenylalanine is causally necessary; but again, this necessary causal condition is not used to define the disease. What, we might wonder, is to stop us asserting “Ingesting phenylalanine causes phenylketonuria,” thus defining phenylketonuria as an illness due to an environmental toxin – not a genetic condition, but rather a kind of poisoning? We could then add that many – indeed, most – people have a hereditary resistance to the disease. Likewise, what is to stop us saying “Tuberculosis is inherited,” and defining tuberculosis as a hereditary vulnerability to an organism that occurs in our environment?

In response, we could insist that tuberculosis is essentially an infection, while phenylketonuria is essentially a genetic disorder. Or we could allow that the definition is entirely a matter of stipulation. Neither response is terribly satisfactory on its own: further arguments would be needed to support each. The monocausal model of disease does not suggest any such arguments. That is a weakness, because in effect the
monocausal model raises a question about disease classification which it cannot answer by itself.

Another difficulty for the monocausal model is that it does not readily account for cases where a cause of disease is present in a healthy individual. Yet this is nearly ubiquitous. Almost every disease is such that its cause may be present without causing the disease: there are carriers, there are asymptomatic infections, there are immunities. The tubercle bacilli can live in the body for years without causing illness, until the host’s immune system is weakened through malnutrition or some other influence. Typhoid Mary lived for many years with the typhoid bacteria but without typhoid fever. Carriers like her pose a difficulty for the specification of the circumstances in which the defining cause is sufficient for the disease. Either the circumstances must be specified, or we must reclassify typhoid fever. The former course is attractive; indeed, Alfred Evans argues the search for the “third factor” ought to be a central concern of modern epidemiology (Evans, 1993 Ch 11). But it is a major undertaking, and we will normally have to classify diseases before that task is complete. Until the circumstances in which a disease cause is sufficient are specified, we have no principled grounds on which to ignore cases Typhoid Mary.6

Finally, among social epidemiologists, the monocausal model might be thought to represent a sort of biological chauvinism: a refusal to countenance causes of ill health that are not biological. There are certainly prominent epidemiologists who seem to fear that their colleagues are guilty of something of this sort (e.g. Krieger, 1994; Marmot, 2006). Social, economic and psychological causes are, prima facie, very unlikely to display the sort of universality and uniformity which would allow them to qualify causes for diseases on the monocausal model (Angel, 2008 Ch 2). So a social epidemiologist, interested in the links between socioeconomic factors and ill health, could be forgiven for feeling that her subject of interest is being slighted somehow. She might feel that socioeconomic factors are centrally important in the etiology of disease, but that the monocausal model focuses attention on biological causes. There are all sorts of reasons why she might regard this as undesirable: for example, Nancy Krieger argues that overlooking social causes of ill health leads to overlooking life-saving public health interventions and to a deficit in scientific understanding (Krieger, 2007). This line of objection is reminiscent of the nineteenth century hygienists’ worry that the bacterial theory would let the authorities off the hook with regard to
improving living conditions for the urban poor (Vandenbroucke, 1988). If a model of
disease focuses attention on biological causes, this is bound to cause concern among
those whose scientific or political interests concern the social, economic,
psychological, or otherwise non-biological causes of disease. One line of response
would be to insist that biological causes of disease are somehow categorically
different to these others, and that epidemiology ought to restrict its scope to biological
causes. However, this response is not directly motivated by the moncausal model.
Moreover, the interaction between the biological and social may mean that it is
impossible to fully understand public health phenomena if attention is restricted to
biological causes (Venkatapuram & Marmot, 2009). For example, it is hard to
understand the current obesity epidemic if one ignores economic factors.

To defend the moncausal model against all these objections we would have to start
by making its rationale clearer; and if we are to substitute a better model, that too
must have a clearer rationale if it is to count as an improvement.

4. Multifactorial approaches

Until the latter part of the nineteenth century, medical text books contained lengthy
lists of causes for any given condition. Effort was sometimes made to distinguish
immediate causes from remote, or to make some other distinction of this sort; but it
was nevertheless standard to find a great many different causes listed in the etiology
of a single condition, and also to find a number of repeat offenders among the causes
of a great many different diseases (Carter, 2003 Ch 1). Various kinds of woe and
morally disreputable behaviour featured heavily, as did weakness of constitution. It is
hard not to draw parallels with modern “lifestyle” diseases, studies of which
frequently reveal either their relation to a familiar range of environmental factors (e.g.
eating excessively or badly, not exercising enough, certain kinds of sex, stress), or a
genetic marker that is widespread in the population, and “predisposes” people towards
the condition in question without being either necessary or sufficient for it.

Nevertheless, the attractions of a suitably modernised multifactorial model cannot
be dismissed. The fact is that the conditions which are the focus of much modern
epidemiology (e.g. cancer, diabetes, heart disease) naturally suggest a multifactorial
model, because of their complexity. Moreover, modern epidemiology also concerns
itself with public health problems that are hard to see as diseases in the traditional sense of the term at all: obesity, suicide, and the ratio of male to female births, for example. Epidemiology surely has techniques that can usefully be applied to the study of such topics, and they all have a bearing on public health, the improvement of which is surely central to the purpose of the discipline. Why, then, should epidemiology concern itself with an abstract and potentially rigid notion of what counts as a disease?

At this stage I need to distinguish two ways to understand the multifactorial approach. A multifactorial approach could simply reject the strictures of monocausal model, asserting that diseases have many causes, which may sometimes be present without the disease or absent when the disease is present. This is the bare multifactorial model: it consists of no positive assertions about disease causation, and asserts no restrictions on what causal structures are specific to disease. On the other hand, we might devise a contentful version of the multifactorial approach, one which does make some positive claims about the causal structures appropriate for disease.

Modern epidemiology seems to endorse the bare multifactorial approach, for the simple reason that this is the logical default position if one rejects the monocausal model and says nothing else. Although causation receives considerable treatment from theoretically inclined epidemiologists, none of the proposals in the literature say anything specific about disease: they are designed to help epidemiologists think clearly about causation more generally, not to provide criteria – as the monocausal model does – for disease individuation. Thus, for example, the “web of causation” (MacMahon & Pugh, 1970) seems designed primarily to wean epidemiologists away from monocausal thinking, and the “causal pie” model (Rothman, 1976) is explicitly drawn from a philosophical theory of causation in general (Mackie, 1974). Neither identify special causal structures associated with disease.

The lack of specific claims about disease causation leads to a number of difficulties for the multifactorial model. The most obvious problem is that on conventional philosophical views of causation, everything is multifactorial: almost every actual caused event has very many causes. This view is orthodox among metaphysicians (e.g. Mill, 1887; Lewis, 1973, 1986; Beebee, 2004; Hall, 2004). If it is accepted, then saying that diseases are multifactorial – have many causes – is trivially true, because everything that is caused is multifactorial. Since it is trivially true of any caused event,
the multifactorial approach cannot possibly identify any distinctive feature of disease, and so cannot be methodologically helpful in this way. On the other hand, we might take an unorthodox, selective view of causation is taken, on which it is not the case that every event has an enormous number of causes (e.g. Ducasse, 1926; Hart & Honoré, 1985; Lipton, 1992; Schaffer, 2005; Broadbent, 2008). Then the multifactorial model seems to be false, unless it can make good on the implausible claim that diseases are multifactorial while everything else is not.

For this reason, the bare multifactorial approach does not give us any hints about how to classify diseases – which features to use to individuate diseases and distinguish one disease from another. This is in contrast to the monocausal model which, as we have seen, does constrain (or at least is intended to constrain) how diseases are individuated. As I have already mentioned, Carter makes a convincing case that the classification of diseases by causes played an important role in the development of modern medicine (Carter, 2003 Ch 3). The multifactorial model suggests no particular scheme of classification; we are left to pick our way towards groupings of symptoms and causal factors. There may be all sorts of reasons to prefer some groupings to others, but they are extraneous to the bare multifactorial approach.

To my mind, the most serious objection to the bare multifactorial model is that it makes no connection between the concept of disease and any notions of explanation or understanding. This point is wonderfully expressed by Jacob Henle, who was a teacher of Koch’s:

Only in medicine are there causes that have hundreds of consequences or that can, on arbitrary occasions, remain entirely without effect. Only in medicine can the same effect flow from the most varied possible sources. One need only glance at the chapters on etiology in handbooks or monographs. For almost every disease, after a specific cause or the admission that such a cause is not yet known, one finds the same horde of harmful influences – poor housing and clothing, liquor and sex, hunger and anxiety. This is just as scientific as if a physicist were to teach that bodies fall because boards or beams are removed, because ropes or cables break, or because of openings, and so forth.

(Henle, 1844; cited in Carter, 2003, p. 24)
Henle’s complaint here concerns, not the ability of medical science to identify causes of disease, but its inability to forge a link between the identified causal factors, and our understanding. Henle is not saying that the medicine of his day has wrongly identified causes of disease. He is saying that it is focusing on the wrong causes: causes that are not “scientific”. The primary objection is not that poor housing and liquor do not cause disease. No doubt some scepticism is implied on that point. But that cannot be the main point, or the analogy with physics would be ludicrous: who can deny that objects fall because boards or beams are removed? The point is not that such explanations are false; it is that they are not what you would expect from a physicist. Likewise, the point is not that sex or bad clothes are not causes of disease: it is that, even if the are, they are not the sort of causes you would expect to hear about from a medic.

Two questions arise. First: why not? Second: what bearing does this objection, directed as it is against a certain sort of explanation, have on models of disease?

The first question is hard to answer without a fully general account of scientific explanation, and no such account is available. Nevertheless, Henle’s example suggests a useful distinction, which I shall cast as between general explanation and particular explanation. A general explanation is one which holds for all the events in a given class, while a particular explanation says why some particular event or events occurred. Henle’s example from physics illustrates the distinction: the reason that it would sound strange if a physicist were to tell us about boards and beams is that physics is concerned with general explanations, that is, explanations that hold good for a certain class of events. Falling is a class events; the removal of boards and beams, the cutting of cables, and so forth do not explain all the members of that class, even if they explain some of its members. Of course, science gives both general and particular explanations: we can have a scientific explanation for the dropping of a single penny as well as for the falling of objects everywhere. I take it, though, that Henle’s point is not supposed to exclude particular scientific explanations. It is rather that the medicine of his day lacked generality of explanation, and that this lack of general explanation is a shortcoming.

This distinction between particular and general explanation helps show how Henle’s objection bears on the multifactorial model of disease. The bare multifactorial model fails to push us towards general explanations of disease. It merely
acknowledges that diseases have many causes; accordingly, it does not by itself suggest that there might be anything more to understanding disease than cataloguing more of its causes. Unless a positive, non-bare version is advanced, this charge will hold good for the modern multifactorial model of disease as well. If it adopts bare multifactorialism, the only difference in this regard between modern epidemiology and the medicine criticised by Henle is that modern epidemiology is able to supply much more precise lists of causal factors, since it has more advanced statistical and experimental techniques. But, as I argued, Henle’s primary objection was not that medicine got the causes wrong; it was that it focussed on the wrong causes.

Worse, unless we are extremely lucky, it may be that general explanations of diseases are not possible without reclassifying diseases. I have already argued that the multifactorial model does not force reclassification of disease; and I also argued, in the previous section, that there is no reason to expect pre-scientific groupings of symptoms to share a common causal factor (that is not also a cause of some cases absence of that disease). Unless we are extraordinarily lucky – and we already know that we are not – it follows that general causal explanations of disease will rarely be possible on the multifactorial model of disease. In this sense, the multifactorial approach makes the notion of disease unscientific; because on this model diseases are not susceptible to general explanation.

Arguably, Henle’s objection is not only true of some modern epidemiology, but can be demonstrated by some of its shortcomings. Most notably, the identification of “risk factors” is directly associated with the rise of multifactorial thinking. Risk factors are mere correlations; a study which shows that a risk factor is unlikely to be due to chance and which eliminates potential confounding variables may have a good case that the risk factor is causal. Identifying risk factors can be useful for devising public health interventions, and it is a necessary step along the road to general explanation. However, the exercise of identifying causal risk factors must not be confused with the activity of seeking general explanations. Cataloguing the risk factors for falling is not equivalent to describing a law of gravity. Even if the catalogue may be useful for identifying that law, the law of gravity will not follow from the catalogue. There is no escaping this point: it is an instance of the problem of induction. Likewise, no catalogue of the risk factors for disease will entail a general explanation of that disease, no matter how thorough the catalogue is.
The trouble with the bare multifactorial approach is that it threatens to obscure this fundamental difference between knowing about many of the causes of a certain class of events, and having a general explanation for a whole class, independent of the details of each case. When a risk factor for, say diabetes is identified, the next question ought to be, “Does this cause explain why some people get diabetes, and others don’t?” The answer for known risk factors is “No”, because risk factors are typically absent from some of the diseased cases to be explained, and also present in some healthy people. That is not to deny the usefulness of identifying risk factors. But they are a stepping stone towards general understanding. On its own, knowledge of the risk factors for diabetes cannot provide such understanding, any more than identifying the risk factors for falling can provide us with a theory of gravity.

J.P. Vandenbroucke makes a related point against “black box epidemiology” (Vandenbroucke, 1988), an approach advocated by Doll and Peto in their famous report on smoking and lung cancer (Doll & Peto, 1981). Vandenbroucke criticises the suggestion that epidemiologists should treat diseases as “black boxes”, which he calls a “miasm theory for the twentieth century” (1988, p. 708). He suggests that epidemiologists ought to try to illuminate underlying causal mechanisms.

Vandenbroucke’s argument is evidence that the theoretical tension I have identified is felt by some practitioners. However, Vandenbroucke’s focus on mechanisms is not quite right. Mechanisms have attracted some attention in the philosophy of science, but it proves rather hard to say exactly what a mechanism is. Moreover, mechanistic explanation is neither necessary nor sufficient for the sort of general understanding I am advocating. Mechanisms are absent from many general explanations: the law of gravity, for example. Even in medicine, general explanation sufficient for dramatic intervention may be achieved before mechanisms are known: Semmelweis is an example of just this. (In fact, both Semmelweis and John Snow were criticised precisely because they could not identify mechanisms to underpin their respective explanations: specifying a mechanism was a point in favour of miasm theory.) Nor does identifying a mechanism guarantee us a helpful explanation. The causal pathway of a risk factor may be discovered, yet we might still lack a fully general explanation of the disease.

For this reason, the charge of black boxing is not the most pressing one: a more serious charge is that of failing to provide general explanations. This more serious
charge is by no means restricted to “social” epidemiology; it also applies to much modern “biological” epidemiology. Genetic risk factors have been identified for diabetes, for example. Do we understand diabetes better as a consequence? In some senses, maybe we do; but in an important way we don’t. Like the miasms postulated in the days of Snow and Semmelweis, these risk factors are inadequate to provide an explanation of the difference between diabetes sufferers and everyone else. If there is a miasm theory for our times, it is not black boxing: it is the undisciplined use of the risk factor. The bare multifactorial approach does not discipline the correlations epidemiologists discover; and merely establishing that they are causal is not enough. Bare multifactorialism does nothing to encourage the move from a catalogue of causes to a general explanatory hypothesis. For, if all we say about diseases is that they have many and diverse causes, it is not clear what more we can hope for than a catalogue. And the catalogue will never be complete, because in practice the causes of even the most ordinary event are beyond human counting.

By no means am I suggesting that epidemiologists themselves are uninterested in general explanations. On the contrary: many of them are. My claims is that the bare multifactorial model is a poor conceptual tool, given this interest. An example of recent concern over the explanatory value of risk factors is a recent discussion by Ian McDowell (2008). His central thesis is that more attention needs to be given to “moving from analysis towards explanation”; however, his “layer” theory of explanation is not especially clear, and his positive recommendations for the shift (e.g. parsimonious hypotheses, mathematical formulation, testability, scope), while agreeable, are disappointingly vague. Let me now offer some positive recommendations of my own.

5. **A contrastive model**

In Section 3 I listed some problems for the monocausal model of disease, and in Section 4 I argued that a multifactorial approach is also unsatisfactory if it consists in merely rejecting the monocausal model of disease. Accordingly, in this section I shall tentatively propose a more positive multifactorial approach.

The most obvious shortcoming of the monocausal model of disease concerned its restriction of the number of causes (of a certain sort) that a disease may have. Often it
seems helpful to classify diseases according to the action of more than one cause, even for relatively simple infectious diseases such as swine fever. Accordingly, I shall lift this restriction; my model allows that diseases may have more than one cause. The other kind of shortcoming concerned the causal structure specified by the monocausal model, which gave rise to the numerical restriction on classificatory causes. The principal difficulty here concerned the circumstances in which a specified cause was sufficient for a disease. Carriers, immunity, asymptomatic infections and similar seem to be ubiquitous. Either our current classificatory system is radically inadequate, or the monocausal model is too restrictive.

I want to preserve the central idea of the monocausal model – that a disease is a kind of ill health that is caused in a certain way. But I want to analyse both “ill health” and “caused in a certain way” using contrasts. A similar idea is already implicit to some degree in the monocausal model. By being explicit about the contrastive structure of the concept of disease, I hope to devise a more sophisticated condition that does not have the numerically restrictive consequences of the monocausal model, and that can cope with the ubiquity of carriers.

For person \( p \) to have disease \( D \), it is necessary that:

**SYMPTOMS** \( p \) suffers from some of a set of symptoms of ill health \( S \), which are differences between \( p \) and a contrast class \( X \)

**CAUSES** Among the causes of \( p \)’s symptoms are events of kinds \( C_1, \ldots, C_n \), at least some of which are not causes of the absence of the symptoms \( S \) from each member of \( X \)

More plainly, the idea is as follows.

First, to have a disease, you must have some symptoms of ill health. These symptoms are part of the definition of the disease. You need not have all the symptoms associated with that particular disease, but you must have at least some. A symptom of ill health is a difference between you and a contrast class, which is a just a certain set of people, some of whom may be merely hypothetical. I have not yet said who should be in this contrast class – more on that shortly. But for now, note that the contrast class need not be the same for everyone. Thus the contrast class for a 45 year old man might include some bald members, while the relevant contrast class for
a 7 year old boy might not. This allows the analysis to cover diseases that are specific
to age, gender, and indeed any other characteristic (including having another disease).

Second, to have a disease requires that your symptoms be caused by a certain cause
or causes, which must not be causes of the absence of symptoms from the contrast
class. These causes are also part of the definition of the disease. For example, to have
cholera, you must exhibit some symptoms of ill health which a certain contrast class
does not have, e.g. diarrhoea; and these symptoms must be caused by a certain
specified cause, namely the activity of vibrio cholerae in your small intestine.

Note that CAUSES is not an epistemological requirement: we do not have to know
about vibrio cholerae in order to count cholera as a disease. Rather, by counting
cholera as a disease, we *commit* to the existence of something satisfying CAUSES.
The obvious next step is to find out what that is. This is one reason the model is
methodologically so useful.

We could additionally define a notion of *illness*, or mere ill health short that fails to
qualify as disease. This would consist in the satisfaction of SYMPTOMS but not of
CAUSES; the difference between illness and disease is then that, for a disease, a
cause or causes of certain symptoms (absent from the contrast class) are specified,
whereas for an illness they are not. I have no linguistic argument to prove that we
should use the words this way, of course; if different words are preferred to mark this
distinction, I have no principled objection. I do, however, think that the distinction is
worth marking, for reasons I will explain shortly.

The contrastive model as I have proposed it avoids the most obvious problem with
the monocausal model. My contrastive proposal does not restrict the number of causes
to one, or to any number; it merely requires that some cause or causes be specified for
each disease. Thus, for example, there is no bar to admitting that swine influenza is
caused by both a bacterium and a virus. The wording of CAUSES is meant to allow
that any or indeed all of $C_1, \ldots C_n$ may be present in the contrast class $X$, provided no
*member* of $X$ has them all at the same time. Thus, in the contrast class for swine
influenza, there may be pigs with either the bacteria or virus in them, but not both.
Otherwise, swine influenza would not, on this account, be a disease.

The contrastive model also avoids the difficulties with the monocausal model
concerning the circumstances in which a specified cause is sufficient for a disease.
The problem was that, faced with a case where symptoms are absent but the putative cause of disease is present, we have a choice. We can use the case to deny that the putative cause satisfies the special conditions in question ((iii) for the monocausal model and CAUSES for my model, respectively). Or we can insist that the putative cause is indeed sufficient in certain circumstances for the symptoms, and but deny that the circumstances are correct. The contrastive model tells us more than the monocausal model about how to solve this problem. The answer will depend on the contrast class. For example, Typhoid Mary displayed no symptoms of typhoid fever, but carried the infection. Should she be excluded from the contrast class for typhoid fever, or should another general explanation for typhoid fever symptoms be sought? The former approach still leaves us with an attractive general explanation of a large class of cases, while the latter would not. If only on these grounds, the former approach seems the more obvious of the two. If, however, there was no very satisfactory general explanation of typhoid fever, maybe we would take the case of Typhoid Mary differently.

Moreover, on the contrastive approach, the exclusion of cases from a contrast class can itself be analysed contrastively. Why did Mary show no symptoms of Typhoid fever? To find out, we look for further causal differences between Mary, with the infection but no symptoms, and less lucky people with the infection and the symptoms. Peter Lipton argues at length that seeking to understand the causes of contrast plays a central role in scientific discovery. His chief illustration is Semmelweis’s work in nineteenth century Vienna, which focussed on the difference between two hospital wards (Lipton, 2004 esp. Chs 3 & 5). My suggestion is that the notion of disease also has a contrastive structure, which links it to the notion of contrastive causal explanation.

My main criticism of the bare multifactorial approach was that it made no links with the notion of general explanation. The contrastive model establishes a clear link with contrastive causal explanation, but I have still not said how it relates to general explanation. After all, contrastive causal explanations can be particular.

Worse, you might object that there is nothing to stop the contrastive model from being trivially satisfied for any ill person. For example, it may be that among the causes of Mary Jones’s cold is a recent spree in the January sales; maybe she was exposed to the virus in a shop floor scrummage, or perhaps her immunity was lowered
by her being underdressed against the bitter chill. Mary Jones’s shopping spree is a causal difference between her and a contrast class, a selection of females of her age, living where she lives and similar in various other ways, who did not catch a cold. Suppose that none of these went shopping, whether by chance or by our devious selection of the contrast class. What is there to prevent us defining a disease, Mary-Jones-Shopping-Disease (MJSD), on the strength of this single case? The contrastive model would be satisfied, but we would have surely achieved little general, medical understanding. Nor can we respond by requiring that a disease definition should cover every symptomatically similar case: then we would lose all hope of etiological disease classification, because different causes of ill health can have overlapping symptoms.

Nothing in the contrastive model prevents us classifying MJSD as a disease. Nevertheless, the contrastive model of disease does require us to identify general explanations, *relative to the chosen contrast*. A general explanation is one that holds good for a class of events. Even in the case of MJSD, the contrastive model does yield a general explanation of difference between members of the class we are explaining – that is, Mary – and the chosen contrast class. What is wrong with MJSD is that we have chosen a ridiculous contrast class. The contrastive model of disease does not guarantee that we have general explanations *tout court*, it merely encourages us to seek general explanations *given certain contrasts*. I have said, and will say, very little about the source of these contrasts. Most obviously they come from symptomatic groupings that strike us as salient or obvious; a contrastive approach can force us to reorganise these, to achieve general causal differences between explanandum and contrast class, just as the monocausal model can. The pressure for generality is still present, relative to the contrasts we pick. If, for whatever reason, the contrast class described strikes us as a plausible one, the contrastive model will drive us towards a general explanation of Mary Jones’s ailment: MJSD *is* a general explanation, relative to that choice of contrasts.

When we are interested in public health, of course we do not usually pick such specific and arbitrary contrasts to explain. It is plausible that the dominant contrast we will seek to explain is between ill and healthy people. The notion of health might be relative to age, gender, and where you live, but arguably it does not involve staying away from the shops on one particular afternoon in January. That is why MJSD is not a real disease. The contrastive model of disease does not supply any guidance on this.
point, and nor should it; rather, guidance comes from our judgements about what constitutes good health and ill, whatever their source. I have provided no substantive comments on that point, apart from the remark that they are a good deal more general than the concerns that led us to MJSD. I am only arguing that, given a certain set of such judgements, the contrastive model of disease drives us towards general explanations of illness, by encouraging us to identify general explanations of the differences between illness and health.

A recent example where the contrastive model might have been useful is the discovery of the role of the bacterium Helicobacter pylori in duodenal ulcers. In a fascinating and insightful discussion, Katherine Angel shows how the episode of this discovery brought out tensions between monocausal and multifactorial ways of thinking (Angel, 2008 Ch 2). She argues that

…many prominent discussions of ulcer assume that, since the implication of \( H. \) pylori, both acid (an intervening or prior or co-mechanism undercutting the monocausal model) and stress or psychosomatic factors (possible prior or co-conditions) have been made irrelevant as objects of etiological attention.

(Angel, 2008, p. 80)

Angel contrasts several texts which “emphasise \( H \) pylori as ‘the cause of ulcers’” (2008, p. 81) with epidemiologically better informed treatments, such as a report by the National Institute of Health. She says that “the NIH’s emphasis on multiple factors in pathogenesis reflects the extent to which multicausality is a staple of biomedical and epidemiological discourse” (Angel, 2008, p. 77).

Moreover Angel suggests that the etiological reality clearly favours the multifactorial treatment. \( H. \) pylori is neither necessary nor sufficient for duodenal ulcer, and nor is its elimination from a patient either necessary or sufficient for the curing of an ulcer (Angel, 2008, pp. 72-3). She argues, in effect, that monocausal thinking is wishful thinking, a consequence of “a desire for treatments that work on each and every case of disease” (2008, p. 98). But, she argues, this desire is not a good guide to reality.

It may be true that a uniform treatment targeted at necessary causes of disease would constitute a practical advantage; but from the fact that this would make life simpler and tidier, it is not rational to insist that there are single, uniform,
necessary causes of a disease… The impoverished vocabulary surrounding the ulcer debate singles out one element as a necessary, sufficient, and sometimes single cause, in a way that obscures manifest and complex realities.

(Angel, 2008, p. 98)

While Angel is surely correct to warn against wishful thinking, that may not be the only motivation for giving H pylori special treatment. Granted, the realities may be complex; but we might nonetheless feel there is something special about H pylori in the etiology of ulcer. The trouble is that bare multifactorialism does not have the resources to express what that might be; while the monocausal model is simply incorrect.

The contrastive model might have help here. On the contrastive model, duodenal ulcers satisfy SYMPTOMS, and H Pylori can be made to satisfy CAUSES. Cases where duodenal ulcer is present without H Pylori can be handled in one of two ways. Either they are cases of a different disease with some of the same symptoms. Or, if this involves excluding and unattractively large proportion of cases of ulcer, we could define the disease in terms of H Pylori and another cause, such as excessive hydrochloric acid in the stomach. Angel points out that H Pylori occurs without SYMPTOMS in the majority of cases. On the contrastive model, these are an invitation to further investigation, as was Typhoid Mary. We can thus give a more precise sense to the claim that H Pylori causes stomach ulcers, and acknowledge the importance of the discovery by reclassifying some cases of stomach ulcer as a distinct disease. Whether we also include other causes such as acid in the classification depends on our preferences among contrast classes, something I have little to say about here; but at least the contrastive model as I have described it allows us to do so. Cases of ulcer may then split into those instances of this newly defined disease, and cases arising from other causes – just as fever or diarrhoea are been split into a number of different conditions.

Duodenal ulcer also provides a vivid example of the methodological dangers of a bare multifactorial approach. Initially the idea that they could be caused by a simple infection was met with resistance. (Moreover financial interests were opposed; lengthy courses of treatment are more profitable for drug manufacturers than a short course of antibiotics.) Angel draws a fascinating parallel between H pylori and
diabetes, citing an endocrinologist who suggests that the discovery of insulin has distracted attention from the pathogenesis of diabetes. Insulin provides a focus for symptomatic treatment (a profitable one, since treatment is lifelong) much as stomach acid did in the case of duodenal ulcers, Angel suggests. I have sought to argue that multifactorial thinking, without appropriate restrictions, can allow us to get too comfortable with the idea that a disease like diabetes has many causes. A multifactorial model of disease should explicitly guard against this danger.

The contrastive model is designed to do this. Diabetes (at least, Type 1) *looks like* a single condition: it appears to be a single disease. It satisfies SYMPTOMS. Yet in our current state of knowledge, it seems to flow from “the most varied possible sources”. Yet none of the known risk factors seems likely to satisfy CAUSES, either alone or in conjunction with other factors: we cannot identify any set of causes that are (jointly) present in every case of diabetes and (at least singly) absent in its absence. Thus we do not have a general explanation for diabetes. The contrastive model reminds us that however much we currently know about the risk factors for diabetes, we still do not understand it. And a link between this sort of understanding and effective intervention is both historically and conceptually plausible.

6. Objections

The most obvious shortcoming of the contrastive model I have proposed is that it leaves a crucial component unspecified: it does not tell us anything about the contrast class, and especially about the notion of health. This is one reason that the proposed conditions are necessary but not jointly sufficient. Perhaps it is an exaggeration to call this a model of disease until this gap has been plugged. But my proposal is no worse that the monocausal or multifactorial models in this respect, since they do not supply sufficient conditions for disease either. And I have argued that the contrastive model can solve some of the difficulties of those models.

The most obvious place to seek a source of appropriate contrasts would be in a theory of health. Although I have proposed no theory or health, the contrastive model of disease places certain constraints on such a theory. For example, defining health as the absence of disease is excluded (cf. Boorse, 1975). Like Christopher Boorse, I want to distinguish disease from illness or ill health; and like him, I am using the
distinction to mark a difference between ordinary and a certain (recommended) medical usage. But unlike him, I suggest that disease is not merely the absence of health as medically defined. It is also the presence of a certain cause or causes of that absence of health. It is this causal generality, I suggest, that makes the notion “medical” — not, as Boorse would have it, the absence of normative factors.

Philosophical answers to the question, “What is disease?” have tended to focus on the question of whether disease is value-laden (e.g. all the entries in Humber & Almeder, 1997). Although that debate is important, I fear it may have obscured a quite different and perhaps even more important methodological issue concerning the notion of disease, and its relation to the notion of general understanding. I have been trying to draw attention to this issue. Thus my model is neutral as to whether the medical notion of disease is determined by normative factors. On a view like Boorse’s, the notion of disease would not be medical if it were partly determined by normative factors. But on my view, even if the choice of contrast class were shot through with normative considerations (Kingma, 2007), the notion of disease could still be “medical” (Boorse) or “scientific” (Henle), in the sense of providing general causal explanations of contrasts.

Let me finish by anticipating two further objections, which I hope will serve as a representative sample of the many I will not have space to answer here. One is the charge of biological chauvinism, already raised against the monocausal model. The accusation was that the monocausal model directs attention away from social, economic and psychological causes of disease, because causes of this sort are very unlikely to be sufficiently general to satisfy the monocausal model. Even though the contrastive model lifts the restriction on the number of causes, it does not lift the generality requirement: to count as a disease, an illness must still be identified with some general causal differences between sufferer and a contrast class. And it was the generality required of causes, rather than their monocausality, that suggested a charge of biological chauvinism.

The other objection I anticipate also concerns my claims about general causal differences, but comes from the biological epidemiologist, in particular the genetic epidemiologist. The more we discover about the way genes work, the more complex their effects seem to be. Genes interact with each other and with the environment in extremely complex ways. There may be some traits, such as eye colour, for which a
definite genetic cause can be identified. However, other traits such as height simply are not like this: they are determined by a number of genes, and their interaction with each other and with environmental factors such as nutrition (a classic discussion of these issues is Lewontin, 1974). The same goes for the genetic contribution to health. There are genetic illnesses (such as phenylketonuria) for which general differences between sufferers and a contrast class can readily be identified. But the genetic contribution to illness, and to health, is far greater than that. The contrastive model of disease restricts our attention to a fraction of the ways in which genes influence health. To restrict our attention in this way is to radically underestimate the genetic contribution to health.

My response to both these objections is to emphasise the difference I have drawn between the concept of disease and the concept of illness. Epidemiology is concerned with public health. As such, any kind of ill health is a legitimate target of study. I would not for a moment assert that epidemiology, or medicine, ought to confine itself to the study of disease as I have defined it. My intention is not to confine the activities of epidemiologists to a certain kind of cause, or to illnesses with a certain causal structure. My model of disease is not a definition of what should be studied, but a conceptual tool to help study some of those things. Some public health concerns, such as suicide, may never fit the contrastive model. That does not mean that epidemiology ought not apply its techniques to studying suicide, obesity, and the rest. It just means that we should recognise a distinction between such things and what I refer to as diseases. To call something a disease requires, on my use of that term, that we have achieved a certain generality of understanding with regard to its causes, and that deserves recognition.

The effort to move from illness to disease corresponds to the effort to achieve general medical explanation. I do not assert that general explanation is the only kind, nor that it is always attainable, but I do think it has peculiar value. If the peculiar value of that kind of explanation is accepted then the value of the distinction between illness and disease should be clear; it marks the difference between those conditions for which we have achieved a general explanation and those for which we have not.

What if the peculiar value of general explanation is doubted? The relation between understanding and the ability to predict and intervene on events is tricky, but that there is some relation is hard to doubt. Russell’s chicken believed that the farmer who
had fed her every day would continue to do so; but he cut off her head. Had the chicken understood why the farmer was feeding her so well, she would not have fallen into that error. Evidence that administering beta blockers to non-cardiac patients could reduce their risk of heart attack was translated into a practical recommendation, but a subsequent study showed that the overall effect of the practice was an increase in mortality by various means (Deveraux, 2008). An understanding of the effects of administering beta blockers to non-cardiac patients, rather than a reaction to an observed correlation, might have enabled this to be predicted (and indeed, presumably some understanding of the effects of beta-blockers guided the corrective study). Note that this holds even if the observed correlation was correct and was causal. Epidemiologists continue to worry about how to infer causation (e.g. Rothman, 1976; Susser, 1991; Rockett, 1999; Rothman & Greenland, 2005; Rutter, 2007). It seems to me that much of this effort, though worthy, is misdirected. Identifying causes is not that hard; identifying explanatory causes is the more elusive goal.15

Acknowledgements

I am especially grateful to Ron Zimmern, who first drew my attention to this area. I would also like to thank Sorin Bangu, Richard Barnett, Jonathan Birch, Adam Bostanci, Kevin Brosnan, Vanessa Heggie, Hai Hong, Stephen John, Luis Nacul, Sridhar Venkatapuram, Caroline Wright, Caitlin Wylie, two anonymous referees, and audiences at the Brocher Foundation conference on the Science and Politics of Neglected Disease Research, and in the Department of History and Philosophy of Science at Cambridge, for many helpful comments and discussions. I am grateful to the PHG Foundation for a year’s research funding, during which much of this material was developed.

References


Rutter, M. (2007). Identifying the environmental causes of disease: how should we decide what to believe and when to take action? The Academy of Medical Sciences.


---

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.

K. Codell Carter’s 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 influenza, 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.

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.

Arguably, the monocausal approach also brings methodological advantages; hypotheses are easy to test and disprove. The dramatic success of these two interventions provided extremely strong evidence for an inference that a single causal factor offered an explanation of those cases, and that the efficacy of the intervention was due to its removing that causal factor.

I am indebted to Ron Zimmern for suggesting this example.

Nor can we get round the problem by just rejecting the sufficiency requirement represented by (ii): any given disease will have too many necessary causes, such as the presence of oxygen, for (i) alone to provide a useful constraint on disease classification (see note 2).

Paul Thagard makes a distinction between two kinds of medical explanation. One is “a diagnosis of a disease that explains the symptoms”, the other concerns “why the patient became sick with influenza, which we now know is caused by a virus” (Thagard, 1999, p. 20). Unfortunately this distinction does not serve my purposes. I want to understand what makes a diagnosis explanatory, in the right sort of way, with regard to a set of symptoms; and I want to understand what knowing that influenza “is caused by a virus” amounts to.

It may be that a theory of natural properties or natural kinds is needed to make the distinction between particular and general explanation. This is a general issue in the philosophy of science that I cannot pursue here.

Miasm did not exist, but even if it had, it would not have provided an adequate explanation of childbed fever in Vienna or cholera in London. Thus it is no defence to argue that genetic risk factors are different from miasm because they are real, and causal. That is not enough for them to be explanatory.

The philosophical inspiration for this approach is Peter Lipton’s contrastive theory of causal explanation (Lipton, 1990, 2004).

This is necessary to avoid including non necessary causes, for example, the restriction of a certain disease to a particular town. Merely requiring that the correlation be causal is not enough, for there might be a perfectly good causal link between living in that town and having that disease, e.g. via the water supply there.
Indeed, if the distinction were ever put into practice there might be good political reasons not to use my terminology, as Caitlin Wylie has pointed out to me. The consequences of denying that, e.g., heart disease is a disease are difficult to predict. Strictly, they may be present, but they may not be causes of the absence of any of $S$. As I have described the monocausal model, that too is incompatible with Boorse’s theory of health, for the same reason. It is striking that Austin Bradford Hill’s famous guidelines for inferring causation from correlation are all plausibly read as what a philosopher would call explanatory virtues (Hill, 1965). Hill’s discussion is excellent, but his guidelines are often cited as criteria for inferring causation, a misinterpretation which his explicit anticipation seemingly failed to prevent.