Comment

Proof in observational medicine

Paolo Vineis

"*In it (the typical English book of physics) there are nothing but strings that move around pulleys, which roll around drums, which go through pearl beads, which carry weights, and tubes which pump water while others swell and contract... we thought we were entering the tranquil and neatly ordered abode of reason, but we find ourselves in a factory*" (P Duhem, 1914)

"*Duhem was writing about English physics, but the impression he would have of contemporary international molecular biology would surely be similar*" (KF Schaffner, 1993)

Two general philosophical models have been influential in medicine in recent decades. The first – *the realist* – relies upon mechanistic evidence, such as experimental physiopathology. In this model, proof of the effectiveness of a pharmaceutical treatment is based, for example, on molecular evidence suggesting that the drug has some desirable property. The second model – *the empiricist* – requires empirical evidence in patients as the only valid proof that treatment is effective. For the empiricist model, the "gold standard" is represented by the randomised clinical trial (RCT). These two models are not necessarily in conflict; for example, RCTs are conducted only when some molecular evidence of potential effectiveness of a drug is available. What they have in common is a high standard of proof – ie, the "Galileian" experimental rule.

Unfortunately, most evidence used in medicine is not of the Galileain type, either for the reason that experiments in humans are not ethically acceptable (except RCTs), or because experiments in animals may not be relevant. This is particularly true in the study of the following:

1. The natural history of human disease (which is not necessarily reproducible in animals);
2. Signs and symptoms within the logic of diagnostic inference;
3. Cause-effect relationships outside the context of drug administration.

Even in the context of drug effectiveness, the RCT has been questioned as the reference standard. In a classic paper, Jerome Cornfield has clearly shown the limitations of the RCT and the reasons why it should be considered as one of various inferential instruments, together with typically observational (non-experimental) procedures.[1]

The epistemological status of observational medicine should be assessed more carefully, avoiding a hierarchy such as randomised versus non-randomised studies – partly because the vast majority of medical knowledge is based on observation (or even narration), not on experiments.

To further illustrate how the border between hard sciences such as physics (particularly "classic" physics) and observational medicine is not so clear cut, I would like to describe briefly a classic experiment by Galileo Galilei. Before him, people usually accepted a direct relationship between the speed and weight of falling bodies. Galileo postulated that this relation was true only in the presence of air or another medium, whereas in empty space falling bodies would all have the same speed. In the language of observational medicine, the same hypothesis might be expressed saying that the medium's resistance is an effect modifier. In fact, the speed of falling bodies is not influenced by their mass in the empty space (any measure of association, eg, a regression coefficient would be equal to zero), whereas in the presence of a medium, a positive association between mass and speed is observed. Galileo was aware of the impossibility of complete experimental confirmation of this theory, since at that time no equipment was available for creating a vacuum. However, he found a practical solution by analysing separately observations stratified according to the density of the medium.

"... since only a completely empty space, with no air or other – although thin and rarefied – bodies, would be suitable to sensibly show us what we are looking for, and since we lack this space, we will observe what happens in the less dense and resistant mediums, in comparison with what happens in those more dense and resistant".[2]

Galilei's theory could be definitely proved on an experimental basis only when Lord Cavendish was able to create a vacuum, but it was accepted by the scientific community long before on the basis of Galilei's reasoning and "stratified analysis" of his observations.

The analogy with reasoning in observational medicine serves the purpose of showing the partial overlap of disciplines usually considered to be radically different – not to suggest that physics might be a model for medicine.

Unit of Cancer Epidemiology Department of Biomedical Sciences and Human Oncology, Main Hospital and University of Torino, via Santena 7 10126 Torino, Italy
P Vineis

Correspondence to:
Dr P Vineis.

Accepted for publication
July 1996

Elementary logic of observational inference

A statistical approach to causal inference was suggested by Reichenbach. Every relationship of "positive statistical relevance" concerning variables A and B had to be explained by common antecedent events in the following way:

(1) $P(A\&B) > P(A) \times P(B)$
(2) $P(A\&B \mid C) = P(A \mid C) \times P(B \mid C)$
(3) $P(A\&B \mid \bar{C}) = P(A \mid \bar{C}) \times P(B \mid \bar{C})$
(4) $P(A \mid C) > P(A \mid \bar{C})$
(5) $P(B \mid C) > P(B \mid \bar{C})$

where (1) logically follows from 2-5, and indicates that the joint probability of A and B is greater than the product of their separate probabilities; the product would be an expression of independence.

These equations are very similar to those that are used in observational medicine to assess cause-effect relationships. A consideration, on very simple lines, of the logic of causal reasoning in observational medicine may shed light on its main properties.

Usually the product of a study is a complex and still confused observation. Alternative explanations have to be eliminated through eliminative induction, showing that an association is likely to be causal because it is not "con-founded" with others. A simple definition of alternative explanation (or, in the epidemiologic jargon, "confounder") is "an extraneous factor whose effect is mixed with the effect of the exposure to the factor we are interested in, thus distorting the estimate of the latter".[3] Confounding is one of the main reasons that have been traditionally invoked to state the superiority of the RCT compared with simple observation.

Let us suppose that we observe that exposure to tobacco smoke is associated with lung cancer, and then we consider the effect on this relationship of a concurrent exposure c (carrying matches). An alternative explanation takes this form:

a) $p(d \mid c\&e) > p(d \mid \bar{c}\&e)$
   $p(d \mid c\&\bar{e}) > p(d \mid \bar{c}\&\bar{e})$
(carrying matches causes lung cancer)
b) $p(d \mid e\&c) = p(d \mid \bar{e}\&c)$
   $p(d \mid e\&\bar{c}) = p(d \mid e\&\bar{c})$
(tobacco smoking per se does not cause lung cancer).

The concurrent exposure c is a determinant of disease (d), as the relationships in a) show. In addition, in the example when the original observation is stratified according to values of c, the association between d and e disappears (complete confounding).

When confounding is not present, then (b) becomes:

b) $p(d \mid e\&c) > p(d \mid \bar{e}\&c)$
   $p(d \mid e\&\bar{c}) > p(d \mid \bar{e}\&\bar{c})$.

(the apparently absurd example of carrying matches will be clarified later).

Valid induction is possible even in the absence of randomization, if observations are stratified according to levels of the confounder, as Galileo did in his experiment using different media.

The role of induction is "expanding knowledge" (ie, collecting new evidence), whereas deduction's role is "transmitting truth".[4] Deductive reasoning scrutinises the study design and identifies associations which are not empirically true, but logically true. Such associations are not a matter of fact, but logical necessities. Examples are the following:

a) $p(d \mid e\&d) > p(d \mid \bar{e})$
b) $p(d \mid e) > p(d \mid \bar{e}\&\bar{d})$

Both a) and b) represent a selection bias; in a) the exposed group is selected among ill people (for instance, we start a cohort study recruiting as "exposed" to tobacco smoke a cluster of lung cancer cases) and in b) the unexposed group is selected among healthy people. In both cases the association which is found between d and e is necessarily (logically) but not empirically true.

These short examples show that it is possible to design observational studies and analyze the data collected in a way which avoids serious misinterpretation, even in the absence of randomised experiments.

The nature of medical theories and conditionalised realism

What is the nature of an observational medical theory, such as "tobacco smoking causes lung cancer"? It can hardly be claimed that such theories represent universal laws of nature, comparable to the laws of thermodynamics. In the meantime, they cannot be dismissed as simple empirical generalizations. We believe that the statement "smoking causes lung cancer" is clearly related to some natural phenomenon. The feeling that it reflects something more than an empirical generalisation does not mean that we are ready to accept that such a statement is comparable to laws describing basic natural phenomena like the genetic code. Universality tends to be typically a feature at very low and very high levels of aggregation.

According to Schaffner,[5] biology is characterised by "middle range" theories – ie, laws that are intermediate between the simple observation of empirical regularities and universal statements about nature. Such middle range theories have the peculiarity of being strongly based on mutual reinforcement between different types of evidence, at different levels of reality and including some reference to basic laws of nature. The two main features of middle range theories are their being temporal models (ie, they refer to phenomena that undergo a process, like carcinogenesis) and their being "overlapping interlevel models" (ie, they serve to connect different levels of reality).

Let us consider tobacco and cancer again. Even after the publication of persuasive evidence linking lung cancer to tobacco smoking, some investigators questioned whether the epidemiologic evidence incriminated smoking as a cause of cancer in humans. In particular, RA Fisher, an eminent statistician of this century, claimed that the early epidemiologic observations could not be interpreted as a proof of a cause-effect relationship, arguing that one could not rule out that a genetic factor increased

both the propensity to smoke and the risk of lung cancer. A key criticism was that exact knowledge of the mechanisms of tobacco carcinogenesis was necessary to establish a cause-effect relationship. Such criticism was at the root of skepticism towards epidemiological evidence and its applications in public health.

In fact, in addition to the (redundant) epidemiological observations linking tobacco to lung cancer in humans, we have several types of evidence at different levels. Tobacco smoke contains many mutagenic and carcinogenic substances. Both tobacco smoke and extracts induced tumors in experimental animals. A general trend in molecular studies is the increasing evidence that point mutations in tumour suppressor genes (ie, *p53*) and oncogenes (ie, *ras*) may be specific both for the type of tumour and for the critical environmental exposure. This is true also for tobacco. Overall, the consistency among investigations on oncogene/tumour suppressor gene mutations in lung cancer (and other tobacco-related cancers) in smokers is highly suggestive.

Furthermore, Fisher's hypothesis that genetic predisposition both induces smoking habits and increases the risk of lung cancer has been refuted on the basis of twin studies. Nevertheless, there is evidence of a different kind of genetic susceptibility – the ability to metabolise tobacco carcinogens, such as tobacco specific nitrosamines or 4-aminobiphenyl, depends on a genetically-based metabolic polymorphism.[6] In such cases, genetic susceptibility does not seem to be a risk factor per se, but, rather, an effect modifier of the exposure to carcinogens. The existence of genetically based susceptibility to the action of chemical carcinogens is strong proof in favour of the theory of chemical carcinogenesis.

The current model of causality in observational medicine (as described, for example, by K Rothman[7]) claims that cancer is caused by "sufficient" but not "necessary" constellations of single agents (component causes). This view has similarities to the model of causality that has been proposed by Mackie for physics, a model in which causes are defined as "INUS" conditions (Insufficient but Necessary components of an Unnecessary but Sufficient constellation).[8] Similarity rests on the idea of "causal constellation". The problem with epidemiology is that no "necessary" external cause has been identified at the individual level. If any, "necessary" steps could be involved in the mechanistic process – ie, at the DNA level (although the existing evidence seems to reject this interpretation). However, if we change scale, from the individual to the population, we may approximate the INUS model more closely also for external causes like smoking. In fact, Mackie considered an event to be the cause of another event if its elimination would have prevented the effect. This definition is certainly compatible with epidemiology, if we apply it at the proper, population level. Although eliminating smoking cannot predict the individual risk closely, it will eliminate more than 90% of lung cancers at the population level. In other words, at the individual level smoking is an insufficient and unnecessary component of a sufficient causal constellation, where the "necessary" part might be represented by the induction of DNA damage and the failure of DNA repair. At the population level, elimination of smoking is a necessary step to eliminate the vast majority of lung cancers which would arise in the next future.

To admit that smoking causes lung cancer one need not be either a realist or an empiricist, to refer to a long lasting debate in medicine. The realist postulates that empirical observations do refer to some reality in the external world (independently of theoretical models); the empiricist strictly sticks to observable entities, avoiding any judgement about the essence of reality. For example, realists in medicine tend to believe that basic biochemical or molecular mechanisms explain the effectiveness of therapies, while empiricists strongly advocate empirical evidence coming from RCTs.[9] Wide areas of observational medicine, and particularly epidemiology, clearly belong to the empiricist field. As a third alternative, Schaffner proposes a "*conditionalized realism*". This means that a "middle range" theory is held to be true if two conditions are met – that "auxiliary hypotheses" are also true and that no valid alternative explanation can be put forward. The second condition is well known to epidemiologists, since it corresponds to the concept of "confounding". The first condition is also easily understandable: examples of auxiliary hypotheses are that the design of a particular study did not introduce bias; that the evidence collected from animal experiments can be extrapolated to humans; that tobacco-related mutations in specific genes (oncogenes) actually are relevant to the carcinogenic process.

To summarise, which type of message does the "tobacco and cancer" example convey? First we believe that smoking causes cancer not only on the basis of empirical observations in humans (which are limited by their non-experimental nature), but also because we have independent proof referring to different levels of reality. Such proof includes reference to some of our most profound beliefs concerning nature, such as the crucial proof played by DNA damage in carcinogenesis. Therefore, prior beliefs in nature are crucial in the interpretation of empirical observations. For example, the above mentioned story of "match carrying" has been reshaped by Robins and Greenland to show that the choice of a statistical model based eg, on "statistical significance" – as in a "backward elimination strategy" – would lead to absurd conclusions.[10] In fact, if one starts with a saturated model that includes both cigarette smoking and match carrying, where the occurrence of lung cancer is the dependent variable, neither variable would be "statistically significant" if one adjusts for the other. This would make it impossible to determine the causal role of smoking or match carrying from the data alone, without any reference to prior beliefs about nature.

Secondly, the model of causality which is valid in observational medicine is compatible

with the models that have been proposed for physics (such as the "INUS" model). Thirdly, as in other fields of science, in observational medicine the truth of a theory is also conditionalised on auxiliary hypotheses and the lack of alternative explanations. This conditionalised nature of biologic realism (Schaffner) is an example of the interplay between direct evidence and interpretation, in that even an experiment – such as an RCT – will be interpretable only in the context of background knowledge concerning auxiliary hypotheses (although an RCT needs less auxiliary hypotheses than observational medicine).

How strictly interpretation and prior belief can be intertwined with the scientific practice of observational medicine is shown by the role of "model selection" in causal inference. According to Robins and Greenland,[10] "all modelling strategies contain implicit prior beliefs about nature". When we choose which "explanatory" variables, confounders, or effect modifiers to include into the occurrence function, we anticipate which of them make biological sense and are compatible with a reasonable interpretation of the data. "... a statistical model is a mathematical expression for a set of assumed restrictions on the possible states of nature ... For example, a linear ... logistic model for the dependence of subsequent fertility on dibromochloropropane exposure and parity implies the following restrictions about nature: 1. an exponential dependence of the fertility odds ratio on DBCP and parity; 2. a constant odds ratio across DBCP for the association of any parity level ... with subsequent fertility; and 3. a constant odds ratio across parity for the association of any DBCP level with subsequent fertility".[10]

The choice of the model, in fact, is a trade off between different and potentially conflicting goals, such as "saving variance" in the model by introducing few assumptions about nature, decreasing bias by introducing correct assumptions, and increasing bias by introducing incorrect assumptions. In fact, an occurrence function based on a "highly-saturated" model will have small bias provided that no confounding remains but may have large statistical variance.

Therefore, any opposition between scientific knowledge (based on the observation of facts within an "occurrence function"), and non-scientific prior beliefs would be misleading, since prior belief is clearly necessary for a correct building and interpretation of causal models.

A different perspective: observational medicine as a disciplinary matrix

According to Thomas Kuhn, a specific science in a given period can be described according to the concept of "disciplinary matrix", which contains four components as follows:

1. Symbolic generalisations, such as "f=ma" in mechanics, or Mendel's laws in genetics;
2. Anological models, such as billiard ball models of gases;
2. Values, by which scientific claims are assessed, such as predictive accuracy and simplicity, and
4. Examplars, which are "concrete problem-solutions that students encounter from the start of their scientific education".

According to the further developments by Imre Lakatos, a scientific research program contains a "hard core" – the essential hypothesis of a theory – and a "protective belt", a set of auxiliary hypotheses "which has to bear the brunt of tests and get adjusted and readjusted, or even completely replaced, to defend the hard core.[13] According to Laudan,[14] "1. Every research tradition has a number of specific theories which exemplify and partially constitute it; some of these theories will be contemporaneous, others will be successors of earlier ones; 2. Every research tradition exhibits certain metaphysical and methodological commitments which, as an ensemble, individuate the research tradition and distinguish it from others; 3. Each research tradition goes through a number of different, detailed (and often mutually contradictory) formulations and generally has a long history extending through a significant period of time".

Observational medicine is no different from other research traditions. Symbolic generalisations include for example the general idea of "genotoxic" damage as a key mechanism in carcinogenesis, as well as more specific models like the "multistage" model in cancer epidemiology. Values shared by the community of researchers encompass accuracy (unbiasedness) and the use of a formal (ie, meaningful) design in collecting data. "Exemplars" that are used as concrete problem-solutions to teach epidemiologic methods include the history of the link between smoking and lung cancer, ie, the use of both statistical analysis and biological considerations in causal inference.

Although observational medicine includes a variety of points of view (some "realist", like the assumption of genotoxic damage, some "empiricist', like the central role of the RCT – model), it appears that the empiricist tradition prevails by far. The methodological "gold standard" in medicine, which serves both as an "exemplar" for students and as an "ethical" reference, is in fact the RCT. Its role of "ethical" reference is due to the fact that the RCT embodies, by definition, the values of objectivity and impartiality. The RCT is also a kind of cornerstone for statistical theory. Without randomisation, in fact, statistical theory makes little sense. Sander Greenland has clearly shown that inferential statistics have limited relevance when the study exposure has not been randomised. Remedies that Greenland suggests include: "(a) restrain our interpretation of classical statistics by explicating and criticizing any randomization assumptions that are necessary for probabilistic interpretations; ... c) de-emphasize inferential statistics in favour of pure data descriptors, such as graphs and tables ...".[15]

Observational medicine could perhaps reframe its influential metaphors along the same lines. Firstly, one possibility is to decrease the

emphasis on the RCT as gold standard, and the consequent emphasis on the related statistical tools. Secondly, the importance of our prior beliefs on nature should be stressed, as a counterweight to the empiricist attitude of many medical researchers. From this point of view, cooperation between epidemiologists and laboratory scientists is to be welcomed: too many empirical observations about different kinds of human behaviour and the risk of chronic diseases are published each year, and too few attempts to corroborate them with mechanistic evidence are made. The story of tobacco smoking is an exception to this rule.[6] Thirdly, the interplay between proof and interpretation should be recognised: observational medicine is not just an exercise of description of natural phenomena, occurring in a social vacuum, but is influenced by the elements of a broad context which include general models of health and disease. To build its "disciplinary matrix", observational medicine can refer to hard sciences like physics or, on the other side, to medical anthropology as models. However, the most original role of observational medicine could just consist in bridging the gap between natural science and social science.

This paper was presented at the International Congress of Logic, Methodology and Philosophy of Science, Florence, 21 August 1995. I am grateful to Professor Paul Humphreys and Dr Michael Eve for advice. The study was partially supported by the Associazione Italiana per le Ricerche sul Cancro.

1 Cornfield J. Statistical relationships and proof in medicine. *American Statistician* 1954;**8**:18–21.
2 Galilei G. *Dialogo sopra i due massimi sistemi del mondo tolemaico e copernicano*. Edizione nazionale delle Opere. Firenze: Barbera, 1890–1909, rist 1964–66.
3 Henneken CH, Buring JE. *Epidemiology in medicine*. Boston, Toronto: Little Brown and Co, 1986.
4 Giere RN. *Understanding scientific reasoning*. New York: Holt Rinehart and Winston, 1979.
5 Schaffner KF. *Discovery and explanation in biology and medicine*. Chicago: University of Chicago Press, 1993.
6 Vineis P, Caporaso N. Tobacco and cancer: epidemiology and the laboratory. *Environ Health Perspect* 1995;**103**:156–60, 1995.
7 Rothman K. *Modern epidemiology*. Boston, Toronto: Little, Brown and Co, 1986.
8 Mackie JL. Causes and conditions. *American Philosophical Quarterly* 1965;**2**(4):245–55.
9 Wulff HR, Andur Pedersen S, Rosenberg R. *Philosophy of medicine*. London: Blackwell, 1986.
10 Robins J, Greenland S. The role of model selection in causal inference from nonexperimental data. *Am J Epidemiol* 1986;**123**:392–402.
11 Good B. *Medicine, rationality and experience*. Cambridge (MA): Cambridge University Press, 1994.
12 Vineis P. Environmental risks: scientific concepts and social perception. *Theoretical Medicine* 1995;**16**:1–17.
13 Lakatos I, Musgrave A. (Eds.). *Criticism and the growth of knowledge*. Cambridge (MA): Cambridge University Press, 1980.
14 Laudan L. *Progress and its problems*. Berkeley: University of California Press, 1977.
15 Greenland S. Randomization, statistics, and causal inference. *Epidemiology* 1990;**1**:421–29.