Epidemiologic evidence: to believe or not to believe

F.R. Rosendaal, MD PhD

Department of Clinical Epidemiology

Bldg 1, C0-P

University Hospital Leiden

P.O. Box 9600

2300 RC Leiden

In: Qual [illegible]
[illegible] Lester JW Perry, R,
[illegible]
London 1992

Abstract

There seems to be a controversy in the credibility given to the results of epidemiologic studies. On the one hand, some feel that more credibility should be given to the results from the 'basic' (laboratory) sciences than to those of epidemiologic studies. On the other hand, others have claimed that epidemiology offers the ultimate proof of a proposed mechanism or therapy.

In our view, there is only one body of empirical science; therefore, a hierarchy of fields of sciences is not only unjustified, but also counterproductive, since it may lead to wrong conclusions, based on only part of the available empirical evidence.

The information gathered in empirical studies may be false, because of bias, measurement error or chance. Therefore, the results from several studies may be contradictory. This forces us to weigh the results of different studies.

Generally, there are two approaches in judging whether an association, as found in a study, represents a true causal relation. One might apply criteria, for instance of plausibility and coherence, that might be helpful in considering whether an association is causal.

Nevertheless, the judgement will always remain subjective to some extent. One might apply an approach, that incorporates this subjectivity. In this Bayesian view, scientific studies are seen as procedures that modify our prior belief in certain associations into a posterior belief. With this view plausibility and coherence, as well as the unavoidable subjectivity are incorporated in the prior belief, whereas the posterior belief still allows for uncertainty.

Introduction

Empirical science is the gathering of information that may serve as evidence in our understanding of nature. In medical science, this will be understanding of normal biology, disease etiology, diagnostic tools and therapy. The information which is the result of a study may be false, because of bias, measurement error or chance. Therefore, the information obtained in one or several studies - that may often be contradictory -cannot be considered absolute and will have to be weighted.

There is a tendency to downweigh evidence from epidemiological studies, i.e. to consider evidence from epidemiological studies as intrinsically less convincing than evidence from other types of studies, for instance biochemical studies. A common phrase is that epidemiological studies will only yield 'statistical associations' which, apparently, are thought to differ from true associations.

To see what this view is based on, and whether it is justified, we will first examine how one may judge a association to be causal. A cause or a causal factor is a factor that brings about an event. This relationship between cause and effect is not necessarily one-to-one; on the contrary, it is rare that the effect will invariably occur in the presence of the cause and never in the absence of that cause. Usually, many causes are acting together and only when a group of minimally required factors are present the event will inevitably follow. Most of these conditions are still unknown and there may well be different groups of causal factors that produce the same disease. What will be observed is that each condition or cause will increase the probability of the event occurring.

Scientific studies: bias

Understanding nature, in science is the establishment of cause-effect relations by the conduct of scientific studies. When an association is observed in a study, this may be because there exists a true relationship, or this may be because the study was at fault, biased.

There are many different classifications of the possible types of bias, which vary from a simple dichotomy of selection bias and information bias [1] to extensive lists of different forms of biases, sometimes with not very obvious names [2,3]. Since these classifications do not essentially differ and we feel that in evaluating bias common sense is the best guideline, we propose the simple and easy-to-remember classification that is listed in table 1. Here we have classified bias according to the time frame of a scientific study.

Bias in the data is present when there was some form of distorting selection present. An example may be when a cluster of diseases is observed and beforehand associated with some exposure by these observers [4]. For instance, several child cancer cases occur in a small geographical area, in children who have all played near a chemical factory. A subsequent study, as often urged by the worried inhabitants will invariably show an association between the disease cases and the exposure to the chemical factory, which was known beforehand. Or even simpler: the inhabitants of one street are worried because of a cluster of cancer cases, and ask for a study to determine whether the incidence in their street is higher than the national average, which of course, in retrospect, it will be. As this example shows, one may also view chance occurrence as bias in the data, as well as confounding, i.e. spurious associations caused by distorting third variables.

Bias in the research is bias that occurs because something in the process of research, the collecting of data, has gone astray. This includes for instance measurement error and misclassification. An interesting example are the studies of a possibly increased cancer incidence around nuclear power plants in the United Kingdom. Whereas several studies seemed to show an increased frequency of cancers, another study found an higher than average number of cancer cases around sites where power plants had been intended, but never been built [5,6]. Although other explanations are possible and have been brought forward, a likely explanation for this intriguing result is that the research itself introduced a bias.

Bias in the authors is probably always present, since authors can choose which subgroup analyses to perform, which tables to present etcetera. Although often the purpose and endpoints of a study will be decided in advance, very few study protocols give exact guidelines how the data should be presented. It is obvious that authors will not present tables and graphs that show their results in the most unfavourable way.

An example of how the authors can affect the conclusions of a study is formed by two papers, by the same authors and based on the same data, on the relation between AIDS in homosexual men and the use of stimulant drugs known as 'poppers' (amyl nitrite). Before the human immunodeficiency virus (HIV) was known as the causative agent of AIDS, a relation between AIDS and the use of poppers was reported in an epidemiological study. Several years later, when more was known about the viral etiology of AIDS, the authors reanalysed the data, controlled for more confounding variables, and the relation with poppers became less prominent [7,8]. In a way this is the normal process of science: one cannot but admire the authors who first came up with a bold hypothesis, and later tried to incorporate new knowledge in their analyses. Nevertheless, one has to realise that both results came from essentially the same data, which indicates the importance of an author's opinion or biases in analysing study results.

The problem of author's bias is not limited to retrospective or observational studies, since it may also occur in randomised controlled trials. A recent review by Altman and Doré [9] of 80 randomized controlled drug trials published in the leading medical journals showed that the total group of subjects receiving the active compound was consistently smaller than the group receiving placebo. Since randomisation of large numbers of patients would be expected to lead to, on average, even distributions over the groups, one might suspect that some patients in the experimental groups were not reported on in the original publications.

A completely different and extreme form of author's bias is scientific fraud, in which author's willfully and deliberately report on invented data. Although fraud perhaps belongs more in a discussion on criminology than epidemiology, one may wonder whether fraud is as prevalent in epidemiologic studies as in other studies. It seems, as noted by Vandenbroucke [10], that the famous examples of scientific fraud that are reported on in medical journals, are almost never epidemiologic. His explanation is that there is little need for epidemiologist to invent data, since they can *'churn out a paper, or at least some publishable unit, from almost any data set'.* This implies that author's bias is even more important in epidemiologic than in other studies. (Another explanation for the scarcity of epidemiologic fraud may be that it is easier to detect fraud in laboratory studies since these experiments are more readily repeated by other researchers.)

Bias in the journals is also known as publication bias [11,12]. It occurs when medical journals, that have to compete with one another for subscriptions and therefore have a clear commercial interest, preferentially publish reports of appealing results. It is understandable that medical journals will not publish a paper claiming that aspirin does not cure cancer, whereas they will not be able to resist publishing a paper claiming the opposite. This policy causes us only to see the top of the iceberg, or, in the theoretical extreme, only the five percent of studies that are statistically expected to show a significant effect in absence of any true effect, just by chance alone.

Non-causal relationships

When a study, or several studies convincingly show an association, the question remains whether the association is causal. A non-causal relationship may be the result of bias, of unknown confounding, or residual confounding after correction for confounding, or chance occurrence. These different scenarios are depicted in figure 1.

A spurious association is the product of some form of bias, which we have dealt with above. A confounded association results from the effect of a third variable, the confounder, that is associated with the putative risk factor as well as with the outcome variable. Chance may be viewed either as causing a spurious relation, or as equivalent to unknown confounding, depending on one's viewpoint being stochastically or deterministically inclined. A true causal factor is usually part of a causal pathway, in which intermediate factors act both as outcome variables and as effector variables.

When analysing or evaluating a scientific study, it is not possible to discern between these different possibilities with certainty. It is generally not possible to measure bias nor can one know whether a result was caused by chance or unknown confounder variables. Statistical and epidemiological methods may be helpful in elucidating the role of chance or in controlling confounding variables: when a study is well conducted, well controlled and confounding variables are adjusted for, bias may seem unlikely; when the appropriate statistical tests have been performed and a small p-value obtained, chance occurrence may seem unlikely, or, when a small confidence interval is obtained, the estimate may seem precise. Nevertheless, the possibility of bias, unknown confounding or chance cannot be ruled out. So, how then does one decide whether a association is indeed causal? It is clear that this 'decision', since no measurement is possible, is to some extent subjective.

Criteria for causality

In 1965, A.B. Hill proposed a list of nine items that might be helpful in considering whether an association is causal, although he cautioned that the list was neither exhaustive nor a sine-qua-non for causality [13]. The items are listed in table 2.

Strength of association: strong associations, i.e. those with high relative risks or risk differences, are more likely to be causal than weak associations. Although one cannot rule out strong unknown confounders, or even chance occurrence, this appears a reasonable suggestion, the idea being that the strong bias or the strong confounding needed to produce a spurious strong association, would usually be obvious. On the other hand, bias and especially measurement error may weaken the effect estimators in a study, which has the result that a factor which is in reality strongly associated with disease, appears only weakly correlated in a study. Finally, there may also be true causal associations that are weak.

Consistency: when an association is observed repeatedly, in several studies, with different designs in different populations, this lends credibility to the causality of the association. It is evident that a relation between for instance diet and disease that is only found in Dutch clergymen, and not in Dutchmen of other professions or foreign clergymen, does not deserve much credibility. If an association is found in many different settings, however, the same confounder may well be present in all these different populations and research designs, and the association may still be spurious. On the other hand, there may be true effects that are only present in certain subpopulations, for instance only in Asian males of a particular age.

Specificity: a factor is highly specific for a certain effect when it is associated with only that effect, as opposed to factors that are related to a wide variety of effects. If a particular drug when taken by pregnant women causes a well-defined syndrome, which is not seen in children from women who did not take this drug, as was the case with thalidomide, a causal relation is almost certain. High specificity may, however, be the result of a highly specific relation between the putative causal factor and a confounding factor. On the other hand, it is not impossible that some factors do cause a variety of diseases, for instance in case of birth defects, high parental age increases the risk of many congenital disorders.

Temporality: a cause should precede the effect. The evidence should be examined to see whether or not the putative causal factor might have been brought about by the outcome. This can be ruled out in prospective studies, or in factors as blood group, HLA-type, but in many studies this possibility of the supposed causal factor in fact being the result of the outcome should be considered.

Biologic gradient: the presence of a dose-related response, when a greater effect (or risk of effect) is observed with a higher amount of exposure, makes a causal relation more plausible. There may be causes, however, that do not produce a dose-response relation, for instance those with a threshold effect. When a confounding variable has a biologic gradient, i.e. when the association between the putative causal factor and the confounder is dose related, a spurious dose-related response will be the result. One of the finest examples of a spurious dose-response relationship was given by Skrabanek, who commented on a list risk factors for scurvy, made several centuries ago, that included exposure to 'sea air' [14]; he pointed out that this association between exposure to sea air and scurvy must have shown a strong dose-response relation.

Plausibility: a cause-effect relation should be biologically plausible. On the one hand this criterium is quite vague, on the other hand it is probably the one we tend to give most weight to in our judgement on causality. It is clearly one of the more subjective items on this list.

Coherence: this requirement reflects the idea that new findings should fit in the general body of science. When accepting a relation as one of cause and effect is in conflict with what is already known, one has to weigh the new evidence against the evidence for the present knowledge that will now have to be rejected. For instance, if we accept that extreme dilutions still carry some activity, as in homoeopathy, we have to reject the most fundamental ideas of physics and chemistry.

Experimental evidence: in experiments it is possible to manipulate the putative causal factor, which possibility is absent in observational studies. In experiments one approaches the scientific ideal of 'ceteris paribus', all other things being equal except the factor that is being manipulated. This does not exclude the possibility that, by chance, like is not compared with like in a controlled experiment or trial, which introduces confounding. Spurious results may be produced in experimental as well as all other studies by incomplete correction for confounding or by unknown (and therefore uncorrected) confounding. Of course, bias may be also present as author's bias and publication bias. In addition, in research in humans, experiments are often impossible out of ethical reasons, for instance in all issues regarding potential risk factors for disease, or even logically impossible, for instance when genetic factors are studied.

Analogy: A certain factor is more likely to cause an effect, when we are aware of similar factors that cause similar effects. The classical example is that if one drug is capable of causing birth defects, other drugs also may have this effect.

Although these items should not be used as a checklist to establish causality, especially since there are, as given above, arguments against each of them, they offer a useful way to evaluate a possible causal relation.

Epidemiologic studies versus studies from other fields

What might be the reason to consider evidence from epidemiological studies as generally less convincing than evidence from other studies, e.g. laboratory studies? A likely explanation is that in those studies usually an experimental design is used, i.e. the factor under study can be manipulated, which is only rarely possible in etiologic epidemiological studies. As we have seen, however, the availability of evidence from experimental studies is only one of the many criteria one may use to consider causality. As epidemiological studies, all other type of studies may be at fault because of chance occurrence and different forms of bias (measurement errors or misclassification, author's bias, publication bias). It is unjustified to speak of 'statistical associations', since the phrase is meaningless. All associations as observed in a study, of whatever type, are essentially the same, and may subsequently prove to be true or false.

The Bayesian view

The judgement whether a factor is causal is to a large extent subjective, since the items mentioned above are not exhaustive nor absolute, and may not carry equal weight. It seems that the most vague items, are also the most important ones, and the ones one tends to give the most weight to: plausibility and coherence. How does this fit in with the idea that when in a study a significant result is found, the null-hypothesis should be rejected?

Plausibility and coherence are reflections of our prior belief in a hypothesis, i.e. the credibility we are prepared to give to a hypothesis before a study is performed. This prior belief is based on our knowledge of the biological mechanisms involved in the topic of our research, our knowledge of previous studies either on the same research question, or to questions closely related to it. The scientific study is performed to test this prior belief [15]. In this respect, a study can be seen as analogous to a diagnostic procedure. The posterior probability of disease in a diagnostic tests depends on tests characteristics - sensitivity and specificity - and the prior probability of disease of the patient.

Example: suppose the sensitivity of a tests is 80 percent (% positive tests among diseased individuals) and the specificity 95 percent (% negative tests in normal individuals). Suppose the prior probability of disease is 50 percent: the patient may as well be healthy as diseased (or, the prevalence of disease in the population this patient originated from is 50 percent). If we tested 1000 individuals, of whom 50 percent were normal and 50 percent diseased (the prior probability), we would find 0.80 x 500 = 400 true-positive tests among the patients with the disease, and (1 - 0.95) x 500 = 25 false-positive tests among the patients without the disease. Therefore the posterior probability of disease, given a positive test, becomes 400 / 425 = 94 percent. In this population, this proves a useful tests, since it raises our suspicion from disease from an uncertain even odds to an almost certainty.

Now suppose the prior probability of disease is very low, for instance 1 in 1000. If we apply the same test to 1000 individuals, one of whom diseased and 999 healthy, we will find 0.80 x 1 = 0.80 true-positive test results, and 0.05 x 999 = 50 false positive *results; the posterior probability becomes 0.80 / 50 = 1.6 percent. In this instance, the test is not very useful: we did not suspect disease before we performed the test, and we still do not think it at all likely that the patient has the disease after the test turned out positive.*

When we view a scientific study as a diagnostic test, the sensitivity is now called the power (the probability of finding a positive result when there is an effect) and the specificity is (one minus) the p-value (the probability of finding a positive result when there is no effect). Let the p-value be 0.05 (specificity 95%), and the power 0.80 (sensitivity 80%), then we will find positive (significant) results in 80 percent of the cases when the alternative hypothesis is indeed true, and in 5 percent of the cases when the null-hypothesis is true, i.e. when no effect exists. The reasoning to incorporate prior probabilities is exactly similar to that of diagnostic tools. If we have a prior belief of 50 percent in a hypothesis, say that a new drug is superior to an old one, and a study with a 80 percent power shows a positive result, our posterior belief in the superiority of the new drug will increase to 94 percent. If subsequent studies are performed, this posterior belief will become the prior probability, and this will in its turn be affected by the study result, be it positive or negative [16].

When our prior belief is extremely low, say 1 in 1000, a significant test result will have very little effect on our belief: since most positive studies would be false positives, our posterior belief would still be near to only one percent (i.e. 0.80 / 50) = 1.6%). This indicates that implausible, contradictory study results may well occur, and that incorporation of these studies in our probability system need not at all lead to contradictions, but only to, sometimes only slight, modifications of our belief in a hypothesis. In addition, this view makes it clear that studies into the very improbable are useless, and should not be conducted, since they will not have much influence on posterior probabilities.

Conclusion

In the evaluation of scientific evidence, it seems unjustified to give more or less weight to the evidence depending on the branch of science it originated from. It is far more important to consider the quality of the individual studies, regardless of the field they were conducted in, and to view the results in the light of what is plausible, and in the light of what is known from previous studies and from other fields.

References

1 Rothman KJ. Modern Epidemiology. Little, Brown and Company. Boston 1986.

2 Sackett DL. Bias in analytical research. J Chron Dis 1979; 32: 51.

3 Feinstein AR. Clinical Epidemiology: the architecture of clinical research. Saunders. Philadelphia 1979.

4 Rothman KJ. A sobering start for the cluster busters' conference. Am J Epidemiol (Suppl) 1990; 132: 6-13.

5 Forman D, Cook-Mozaffari PJ, Darby SC et al. Cancer near nuclear installations. Nature 1987; 329: 499-505.

6 Cook-Mozaffari P, Darby S, Doll R. Cancer near potential sites of nuclear installations. Lancet 1989; 1145-1147.

7 Marmor M, Friedman-Kien AE, Laubenstein L et al. Risk factors for Kaposi's sarcoma in homosexual men. Lancet 1982; i: 1083-1087.

8 Marmor M, Friedman-Kien AE, Zolla-Pazner S et al. Kaposi's sarcoma in homosexual men: a seroepidemiologic case-control study. Ann Intern Med 1984; 100: 809-815.

9 Altman DG, Doré CJ. Randomisation and baseline comparisons in clinical trials. Lancet 1990; 335: 149-153.

10 Vandenbroucke JP. How trustworthy is epidemiologic research? Epidemiology 1990; 1: 83-84.

11 Dickersin K, Chan S, Chalmers TC, Sacks TC, Smith H Jr. Publication bias and clinical trials. Controlled Clin Trials 1987; 8: 343-353.

12 Rosendaal FR. Fate of manuscripts rejected for publication in the AJR (letter). AJR 1991; 157: 1352.

13 Hill AB. The environment and disease: association or causation? Proc R Soc Med 1965; 58: 295-300.

14 Klevay LM. The role of copper, zinc and other chemical elements in ischemic heart disease. In: Metabolism of metals in Man. Rennert OM, Chase WY (eds.). CRC Press 1984.

15 Howson C, Urbach P. Scientific reasoning: the Bayesian approach. Open Court. La Salle (Ill) 1989.

16 Stijnen Th, Houwelingen JC van. Empirical Bayes methods in clinical trial meta-analysis. Technical Report No 2. University of Leiden, 1987.

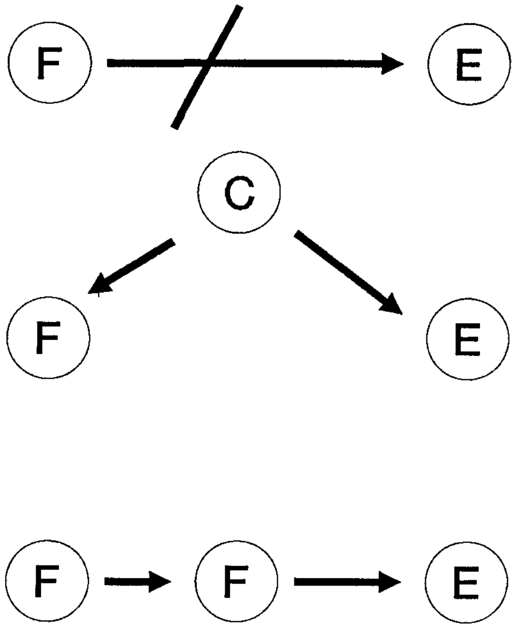
Table 1. Types of bias

type	example
bias in data	selection
bias in research	measurement error, misclassification
bias in authors	selective subgroup analysis
bias in journals	publication bias

Table 2. Hill's criteria for evaluating causality

strength of association	association has a high relative risk
consistency	association is found in different designs/populations
specificity	association exists between only one factor and one effect
temporality	the cause should precede the effect
biologic gradient	association has a dose-response relation
plausibility	association is biologically plausible
coherence	association fits in what is known already
experimental evidence	association can be demonstrated experimentally
analogy	association resembles similar associations

Figure 1. Factors (F), Confounders (C) and Effects (E).



spurious relationship
(chance, bias)

confounded relationship

causal factors