Jan P. Vandenbroucke

Leiden University Medical Centre

The history of confounding

Summary

Confounding is a basic problem of comparability – and therefore has always been present in science. Originally a plain English word, it acquired more specific meanings in epidemiologic thinking about experimental and non-experimental research. The use of the word can be traced to Fisher. The concept was developed more fully in social science research, among others by Kish. Landmark developments in epidemiology in the second half of the 20th century were by Cornfield and by Miettinen. These developments emphasised that reasoning about confounding is almost entirely an a priori process that we have to impose upon the data and the data-analysis to arrive at a meaningful interpretation. The problems of confounding present their old challenges again in recent applications to genetic epidemiology.

Keywords: **Medical history – Epidemiologic methods – Confounding – Case-control studies – Causality – Genetics.**

The word "confounding" has over the past 20 years acquired almost mythical and even mystical proportions in the epidemiologic vocabulary. Originally, it was a plain English word – most probably of Norman origin, since one tends to hear some Latin in it. The Shorter Oxford English Dictionary on Historical Principles (3rd Edition, reprinted 1967) mentions that it is a medieval Latin word: "con-fundere", to pour together (mix together), that was taken over in medieval French, as "confondre". The same dictionary also mentions "to mix up in ideas, to fail to distinguish, to confuse", as meanings that are already distinct in the 16th century. From there other connotations come. Some of the oldest might go back to religion, when the help of the Lord was invoked, not only against pestilences, but also against human enemies. The Lord was asked: "Confound thy enemies", meaning: confuse them, bring them into disarray, a disarray so great that they will easily be dispersed, so that we, your loyal servants, will easily win the battle, and put the enemy to confounded shame. In this way, "confounded" also has other connotations, like doomed, hopeless, shameful etc. Since the 1700s it is regarded as a mild curse. Which should be telling to epidemiologists.

Where to start?

The history of confounding is a mirror image of the history of research design. Confounding is not a statistical or analytic concept. It is a concept that has to do with the logic of scientific reasoning. In particular the logic of inferring causality from observations. Therefore, the student of the history of confounding faces a dilemma that is common to historians: should one study the history of confounding only from the time that the word was coined in epidemiology with its specific methodological meaning? Or should one take the broader view and study the history of the underlying concept from time immemorial, i.e., all instances in which the concept might have been foreshadowed? The latter would include extremely varied sources, beginning with the Old Testament quotation that is often interpreted as "the first clinical trial", in which Daniel opposed the king of Babylon by adding a control group to verify the effects of the dietary precepts of the king upon the youths of Israel. The story is quoted in a paper on the history of the clinical trial, entitled "Ceteris paribus" ("other things being equal") by Lilienfeld (1982). Should we say that this emphasis on a comparison "ceteris paribus" showed that Daniel understood what "confounding" meant, and should we therefore see the bible as the first historical source on the subject?

Soz.- Präventivmed. 47 (2002) 216– 224
0303-8408/02/040216–09

However tempting the broader view, I have limited my inquiry to the more restricted option, for two reasons. Firstly, because the task would otherwise become unwieldy: all texts in which problems of comparisons were ever mentioned – not only the bible, but also ancient philosophers, medieval thinkers up to modern times, should be scrutinised. Secondly, because professional historians convinced me that the history of a concept does not *really* exist before it is more or less securely coined by a name in a particular context. Even worse, they say: going back to the times that neither the word nor its context existed, is nothing but a re-interpretation by hindsight, and is unscientific for an historian, since the re-interpretation only exists grace to the modern concept. The above example makes it clear: to say that biblical Daniel understood "confounding", whereby we imply that he understood the same concept as we do, really seems stretching our imagination too far. There is one exception, however: professional historians like to go back to the time *immediately before* the concept was coined, since that may give insight into its gestation and give clues to its overt as well as covert meanings.

For all this reasons, I will limit my search to the history of confounding in the past decades. Furthermore, my treatment of the subject will be quite personal, and therefore subjective. This aspect of my commentary might not be to the liking of professional historians, because I will trace the development of the concept as if it were a story of continual refinement and improvement until the present. Today's historians frown upon such stories wherein the world continually improves until the present, because this is typical of medical amateur historians who only want to describe the triumphs of present-day insights over a darker past. Yet, I must avow that it is difficult for me to do otherwise, because it is impossible for me to take sufficient distance from today's debates on confounding and their historical roots. I witnessed the aftermath of the development of the concept myself, during my training in epidemiology at the Harvard School of Public Health, and I feel involved with some of the actors in the debates. Finally, there still is something to be said for a mere history of the development of an idea – be it only as a first stepping stone for a more in-depth treatment of the subject, wherein the causes of the evolution of the concept are also traced. As a consequence, my treatment of the history of confounding should be seen as a first rough sketch, to be improved upon by others. The interested reader will find a selection of reprints of several papers on causality and confounding in one volume (Greenland 1987); some of these, besides others, will be mentioned as specific references in my text. The most recent authoritative treatment on the principles of confounding in epidemiology can be found in the textbook by Rothman and Greenland (1998: Chapter 8).

In this historical excursion, I will treat firstly the basic problem of comparability, as originally described by Claude Bernard and John Stuart Mill, and the way in which these thoughts are still very much alive in modern epidemiology. Thereafter, I will concentrate on the evolution of the concept, starting with "desirable confounding" as described by R.A. Fisher, following with "undesirable confounding" as described by L. Kish and taken over in epidemiology. Next I will deal with the interpretation of confounding variables, about which very beautiful pages have been written by J. Cornfield, most notably in discussions on smoking and lung cancer, and in some acerbic debates concerning the interpretation of randomised trials. Today's theory on confounding will by highlighted from the writings on case-control studies by O.S. Miettinen and others. I will end by delineating how confounding is still very much with us, even in the most recent endeavours, the epidemiologic study of the role of genetic factors in the causation of disease.

Comparisons and comparability

The crux of research design, the crux of any observation, is a comparison. It can be a real comparison with data on two or more groups of subjects, or a mental comparison (against what we expect). That the essence of scientific observation always involves a comparison was already beautifully described by Claude Bernard, in the middle of the 19th century, in his "Introduction à l étude de la médecine expérimentale", published in 1865 (Bernard 1966). Although Bernard is mainly known for bringing physiologic experimentation to medicine, he also very clearly described his ideas about research methods in general. He explained that experimental research and observational research have one thing in common: that one thing is the comparison. In an experiment the researcher fiddles with reality to construct the comparison himself: for example, what happens to dogs with and without internal secretion of the pancreas. In observational research the researcher has to search for the comparison, he has to look and find where nature has made the data for him. Claude Bernard even gave an epidemiologic example: he wrote that, if a medical doctor observes that in a part of town, where hygienic conditions are appalling, some diseases are more prevalent, he might think that it is due to these conditions (Bernard 1966: 35). That initial observation is already a comparison, since the doctor compares poorer and richer parts of towns. Bernard called this observation "passive". Such initial observations are the source of later hypotheses and further "active" observation. In clinical

Soz.- Präventivmed. 47 (2002) 216– 224

medicine, they are often communicated as case reports and case series (Vandenbroucke 2001).

Karl Popper, who was not yet born when Bernard confined these thoughts to paper, much later remarked that anything that strikes us, always strikes us because it belies our expectations – again a comparison with the "expected". Claude Bernard made a great point in saying that any investigation always starts with some "preconceived idea" ("… une idée préconçue a toujours été et sera toujours le premier élan d'un esprit investigateur") (Bernard 1966: 59). The initial comparison that led to a new idea may have been made *passively*. Thereafter this new preconceived idea, e.g., the possibility of a greater disease incidence in the poorer parts of town, can be turned into an *active* observation. Claude Bernard wrote that to prove the point, the doctor starts to travel (he will probably mount his horse or carriage – we are still in the middle of the previous century), and that he will travel to another town, to see whether in similar conditions there are similar diseases. The doctor now makes an active observation, he actively seeks another comparison, still without being able to fiddle with reality – it is still non-experimental – but nevertheless he actively checks whether his initial impression is right (Bernard 1966: 35).

How comparisons should be made, be them experimental or observational, was described by J.S. Mill, in his 1856 canons on causality, as quoted in the relatively recent epidemiologic literature by MacMahon and Pugh (1970) and by Susser (1973: 70). The most important citation follows: "Second Canon: If an instance in which the phenomenon under investigation occurs, and an instance in which it does not occur, have every circumstance in common save one, that one occurring only in the former; the circumstance in which alone the two instances differ, is the effect or cause, or a necessary part of the cause, of the phenomenon."

This "method of difference" appeals most to us in medicine and epidemiology. We would wrong the genius of writers like Mill, however, to assume that this was the only way which he conceived to arrive at causal judgements. He described several others, like the "method of agreement", which says that if several circumstances in which a phenomenon occurs are completely different, except in one aspect, then the latter aspect is a likely cause. That is a type of reasoning that we also use in epidemiology: for example, we note that several different types of study in different circumstances all find the same association, which therefore strengthens our ideas about a causal interpretation. Then there is the "method of variation", which sounds very much like a dose-response argument. One might well say that these canons foreshadow Austin Bradford Hill's ideas about causality (Hill 1965).

However, let me keep with the second canon: the idea of "ceteris paribus" that is present in that canon, applies equally well to observation as to experiment – and it applies even to thought experiments. Whenever the condition of "all other things being equal" is *not* met, the comparison might be wrong. Wrong information confuses, wrong information brings one into disarray, wrong information is confounded information.

Very crudely put: any departure from J.S. Mill's second canon, any departure of the "ceteris paribus" principle can lead to confounding. This is the essence of confounding. Nowadays, epidemiology has developed distinctions between several reasons why comparisons go wrong (the generally accepted terminology says that the comparison or the study is "biased") – of which confounding is only one.

Desirable confounding

The very first, at least to my knowledge, to apply the word confounding in thinking and writing about research designs was R.A. Fisher. He treated confounding at great length in his 1937 book on "The design of experiments" (Fisher 1937). However, in his treatise, confounding was *not* something that he always sought to avoid. On the contrary, he proposed to exploit confounding, by deliberately introducing confounding in agricultural experiments. He proposed to ignore higher order interactions between treatments by deliberately confounding the higher order interactions with some of the main effects in the design of the experiment. Of course, he presupposed that the investigator was certain that she was not interested in these higher order interactions, and also that she knew in advance that they would not add important effects over and beyond the main effects. Fisher seemed to have been very fond of this invention, which is quite complicated to read and understand. No less than 40 pages of the 260 pages of his book are devoted to "confounded designs". The book is not chiefly remembered for it. However, let me retain the notion that R.A. Fisher used the word confounding as a nuisance which he tried to turn into a benefit.

Undesirable confounding

The next important use of the word confounding, which we come across is by Leslie Kish, who devoted himself to methodological theory in sociologic research. His 1959 paper about "Some statistical problems in research design" is still worth reading, and a great source of contemporary references (Kish 1959) – it is indeed the time period in which the current use of the term confounding was born. In

thinking about research designs, he discerned the following four variables:

I. *Explanatory* variables, or "experimental" variables: the object or research, both "dependent and independent".
II. Extraneous variables which are *controlled* (in selection and estimation).
III. Extraneous uncontrolled variables, which are *confounded* with the Class I variables.
IV. Extraneous uncontrolled variables which are either actually *randomised*, or treated as if randomised. (Randomisation is a substitution of experimental control).

Kish's use of the word confounding derived from Fisher's: a confusion of two effects. The big difference, however, is his categorisation of the different variables that might influence the outcome of a study. Although not very explicit, he seems to make already a distinction between confounding and other types of bias: his second type of controlled variables have to do with measurement and selection. Nowadays, the word confounding is indeed used for one particular form of the confusion of two effects: the confusion due to extraneous causes, i.e., other factors that really do influence disease incidence, e.g., age, sex, habits, or living circumstances. The word confounding is not used to describe the problems that arrive by differences in measurement or selection. The latter we call nowadays "information bias" and "selection bias". They are artefacts of the design of the study. The separation of confounding from selection bias and information bias is in practice not always very clear-cut – the reasoning sometimes becomes difficult.

I am not certain how Kish's use of the word confounding entered epidemiology. Kish's views obviously were very influential, and it is possible that he influenced epidemiology via the writings of other social science methodologists like H.M. Blalock (1964) or via Campbell's writings on quasi-experimentation (Campbell & Stanley 1963).

The interpretation of a confounding variable

Two leaps in the history of confounding are linked to the name of Jerome Cornfield. One is the epochal paper of 1959 in which he discusses, together with Haenszel, Hammond, Lilienfeld, Shimkin and Wynder, whether the 10-fold increase in lung cancer observed among cigarette smokers might be due to confounding with some other effect (Cornfield et al. 1959). The paper was written against one of the major initial objections to the idea that smoking would cause lung cancer. That objection was championed, amongst others by R.A. Fisher: his proposition was that there was some "underlying constitution" which caused both lung cancer and a propensity to smoke. Thus, the association between smoking and lung cancer would not be causal, but simply due to this underlying constitution which caused both. Cornfield and his colleagues who jumped to defend the causality of the association, did not use the word confounding in their paper; they spoke about a "non-causal agent" and a "causal agent". Today, we would call the underlying truly causal agent the confounder and the non-causal agent, with the apparent association that is non-causal, the confounded variable. Cornfield and colleagues demonstrated that a confounding variable, if any, would in itself need to have an even greater effect on the occurrence of lung cancer than a 10-fold increase to explain the association of smoking with lung cancer. They challenged the non-believers to come forward with such an agent. In general, epidemiologic reasoning admits that there might be differences between smokers and non-smokers, e.g., smokers drink more coffee. The crux of the question is, however, that to deny that smoking is a potential cause of lung cancer, one has to come forward with proof that something associated with smoking, e.g., coffee drinking, is a true cause of lung cancer. Moreover, it should even be a much stronger cause than the apparent association between smoking and lung cancer – otherwise it will never suffice to explain the association. What this historical example demonstrates, is that one has to *reason* about potential confounders, and that one should not take them as mythical or uncontrollable phantoms that destroy studies. People who propose that a certain study is confounded have to make clear why and how, and have to do so in logical and credible terms. Only if they do so, a meaningful discussion becomes possible (Vandenbroucke & de Craen 2001).

Quite recently, Cornfield's reasoning was perverted into its inverse. In the famous Science article on "Epidemiology faces its limits", by Taubes (1995), it is quoted, completely out of context that a relative risk should at least be elevated two or threefold, or even that the lower boundary of the confidence interval should be two or three, before being credible. The beautiful reasoning by Cornfield and his associates is turned into its opposite. Sometimes there is a vested interest in not wanting to believe the results of epidemiologic studies, for example when epidemiologic studies show side effects of medicines. Some persons like to teach that the possibility of confounding is so great when relative risks are low that they do not even need to name and articulate the confounder. They think they have a right to dismiss such a study without argument (Sackett et al. 1997). Wynder (1996) recently commented about "weak associations" showing the fallacy of this reasoning. After all, a twofold increase in the risk of disease is still 100 % more disease. Confounding is still with us.

Soz.- Präventivmed. 47 (2002) 216– 224

Cornfield's next contribution was even more subtle. It was his discussion about the results of the University Group Diabetes study (UGDP) in 1971 (Cornfield 1971). This discussion is very important. It showed that confounding can still exist after randomisation. After all, randomisation is only a game of chance, and it might only guarantee equality of "all other known and unknown" factors that influence the outcome of the study in the very long run or with very large sample sizes. Historically, randomisation was used principally as a means to *conceal* the allocation (Chalmers 1999). In theory, any type of allocation would be fine, except that fixed schemes like alternation, day of birth, etc. have the drawback that the physician knows in advance what treatment the next patient will receive. To circumvent that problem, randomisation was the solution. Thus, randomisation is only a guarantee against physician bias in the allocation. From a purely theoretical point of view, it has even been argued – and again Fisher was invoked – that randomisation can *never* guarantee complete equality between groups: one can always invent "a million ways to compare two groups" and there will always be something that is different (Urbach 1993). In modern times, this idea has new relevance when we think about genetic differences. Since humans have billions of base pairs, it is mathematically certain that randomisation will not guarantee equality. Even if you were to randomise tens of thousands of patients in an enormous randomised controlled trial, there will be tens of thousands of base pairs that will differ between the two groups. Some of these might be genetic polymorphisms that are important for prognosis. We will never know, but fortunately such unknown chance variation is taken care of by the confidence interval (Altman & Bland 1999).

Anyway, in actual practice it is quite possible, that by the luck of the draw one of the comparison groups in a randomised trial has different baseline characteristics, and has therefore a more favourable prognosis than the other. This was the case in the UGDP study (Cornfield 1971). In that study it was found that people who had been treated with certain oral glucose lowering tablets fared worse: they sustained more myocardial death than people treated with insulin or even with diet alone. Critics of the study, however, were quick to point out that the group that was randomised to tablets had a slightly less favourable prognosis: more people in that group had a history of angina pectoris or digitalis use, they were slightly older, with a little more males, somewhat more radiologic arterial calcification, and they were slightly more obese.

Cornfield took up the challenge (1971). In his treatment of the subject, again he did not use the word confounding; he spoke about "random and non-significant base-line inequalities". He took it up in the same spirit as in the earlier contribution, that is, that one has to reason about the strength of an alternative explanation. And he did so in a multivariate way. He constructed a multivariate prognostic model, and fitted the model on all groups with an indicator variable for the different treatments. Next he did two things. He showed how the base-line prognosis in each treatment group could be estimated, and how it differed a little, but not nearly as much as the real differences in outcome. His method of estimating overall base-line prognosis was ingenious: he had fitted an outcome model on the data with an indicator variable for the treatment groups, but thereafter he estimated the base-line prognosis of both groups after omitting the treatment indicator variable. Second, he stratified all groups according to their multivariate risks, and again showed that this stratification had little effect on the difference in outcome between the treatment categories. This foreshadowed Miettinen's multivariate confounder score (Miettinen 1976).

The giant leap which Cornfield made was to make confounding a matter of judgement, even *after* randomisation. Even after randomisation the credibility of the comparison between the two treatment arms should be checked, and if necessary remedied. We do not care about the possibility that there are potentially innumerable differences between two groups after randomisation; we only care about the differences that matter in a causal explanation. Thus, we have to make a double judgement, based on prior knowledge: what are true prognostic variables, and do they differ between the groups. This philosophy, however, also leads to the idea that the randomised trial is not necessarily an instrument that delivers "true comparisons" automatically, by virtue of the randomisation itself. It makes the randomised trial only one of the study designs in epidemiology, about which one has to reason in exactly the same way as about the other study designs that are observational. As Cornfield later wrote himself, he placed "… emphasis on reasonable scientific judgement and accumulation of evidence and not on dogmatic insistence on the unique validity of a particular procedure" (Cornfield 1976).

It remains ironic that Cornfield made such great contributions to our thinking about confounding, but did not use the word. I wonder whether he avoided it on purpose. In this regard he is much like two other pioneers of epidemiology, Mantel and Haenszel, who wrote in 1959 a paper about the analysis of data from case-control studies, in which they proposed the currently very famous "Mantel-Haenszel test" as well as the "Mantel-Haenszel estimator" for the common odds ratio (Mantel & Haenszel 1959). In the treatment of the latter subject they speak of "factor control" and not about confounding. It seems that also in the earlier teaching of epidemiology at Johns Hopkins the word confounding

Soz.- Präventivmed. 47 (2002) 216– 224

was not used, but that it was denoted by the word "secondary association", i.e., secondary to something else that was a known cause of disease (personal communication, Milton Terris, Annecy France 1996).

Confounding in case-control studies

The idea that confounding is a matter of credibility of comparisons, hence to a certain extent subjective, was going to play an important role in the last developments of our insights into confounding. These have to do with case-control studies.

The problem faced by case-control studies, as they emerged as important tools in research, can be delineated by comparing them with follow-up studies, and in particular with the "idealised" follow-up situation, which is the randomised controlled trial like the ones we just discussed. In such a trial, and in any follow-up study, one can actually look at the data to see whether the exposed and the unexposed are different in their prognosis, at least as far as we know prognostic factors. We can tabulate the differences, which is always done in the famous "Table 1" of any randomised trial, the table with the baseline characteristics of the different treatment groups. As shown by Cornfield, one can then make a judgement: how much the groups differ and how importantly that difference will influence the outcome.

However, in case-control studies, that is not possible. Worse, when one looks at the baseline characteristics of the cases, they always have a poorer prognosis in all respects. If they are cases of myocardial infarction, for example, they will have more hypertension, more hypercholesterolaemia, more familial heart disease, more male pattern baldness, and whatever you wish to look for. It is never possible again to see whether at baseline the exposed and the unexposed, (say, smokers and non-smokers), differed. Even looking at exposed and unexposed in the control group is only a poor substitute, because the control group is only a sample (at best) of the combined population of exposed and unexposed people. Associations in the control group might be a matter of "chance" sampling variation. So, how should one go about the decision which factor is a confounder that needs adjustment – whatever the practical means: restriction, selection, matching, or multivariate analysis. The solution proposed by some is akin to the solution proposed by Cornfield on randomised trials: only adjust for *potential* confounders, i.e., other causes of the outcome that are potentially confused with the exposure of interest. This, however, is even more judgmental than with follow-up studies, because you cannot verify the baseline characteristics (the total population of exposed and non-exposed is not known). This situation may account for part of the long history of controversy that has accompanied case-control studies. Much of the theory that in the end the judgement about confounders is an *a priori* judgement has been developed in the department of epidemiology at the Harvard School of Public Health, among others by Miettinen in the 1970s (Miettinen & Cook 1981). Pivotal in the development of these thoughts were deeper insights in the role of "matching" in case-control studies (Miettinen 1970), and the idea of the "confounder summarising score" (Miettinen 1976).

Matching was a time-honoured way of tackling confounding in case-control studies, already mentioned by Mantel and Haenszel (1959). It was originally seen as the equivalent of "blocking" in a randomised design. Blocking in a randomised design means that one first assigns the subjects to various "blocks" depending on characteristics in which they are equal. Only thereafter randomised allocation to the treatment arms is performed, separately for each block. This assures that for the characteristics of the blocks, the two treatment arms will be perfectly equal. It led to the old experimental maxim: "Block where you can and randomise where you cannot" – meaning that known prognostic factors should be used for blocking, to assure their equal distribution over the treatment groups, whereas the unknown factors should be taken care of by randomisation. Superficially, making controls alike to cases in case-control studies seemed similar: e.g., if the first case of myocardial infarction is an elderly gentleman, the first control should be a man of the same age. By doing this, however, something else also takes place: by making controls alike to cases, they will also become much more alike in the exposure that one wants to study, e.g., elderly gentlemen all tend to smoke. As a matter of fact, matching on confounding factors, which is intended to make the comparison series alike to the cases in case-control studies, introduces its own "bias" – a bias towards no association, be it in a controlled way (Miettinen 1970). The solution is to perform a stratified analysis. Indeed, a "matched analysis" wherein each case-control pair is seen as a single stratum is the same as a Mantel-Haenszel analysis with stratification for the confounding variable (Mantel & Haenszel 1959). A nice recent explanation can be found in Rothman's textbook (Rothman 1986). This pivotal insight, that "matching" in a case-control study performs something totally different from blocking in a follow-up study opened the way to a deeper understanding of confounding in case-control studies.

The "confounder summarising score" was developed by Miettinen as an extension of Cornfield's analysis of the UGDP study (Miettinen 1976). It calculated for each individual in a study his or her "baseline probability" to get

diseased, or to have been exposed (depending on whether an outcome or an exposure model was used). Thereafter, the individuals were grouped in strata with similar probability, and the analysis proceeded by simple stratification. Although the use of the confounder summarising score was abandoned later, because of the wide-spread use of the logistic model in "canned software packages", it made confounding very insightful, and was therefore again an important intermediary step in our understanding.

The reason for much of the ongoing discussions about case-control studies is that case-control studies are often about side effects, be it of drugs or of exposures of daily life (like putting babies to sleep in the prone position, or eating cookies, or being exposed to cigarette smoke).There are always parties with a strong interest to challenge such findings and dream up all possible biases and confounders. An attempt at bringing people with different views on case-control studies together was the so-called "Bermuda Peace Conference" organised in 1978, whose proceedings were published in the Journal of Chronic Diseases (Ibrahim 1979). It still makes useful reading, especially in the light of ongoing debates about confounding and selection in case-control research.

Causal pathways

Arguments what constitutes a "proper" confounder have even become more difficult because of the added complexity of "causal pathways". Causal pathways were described by Wright, Wold (1956) and Blalock (1964), and brought into epidemiology by Susser (1973: 111–35).

When we want to study the relation between some exposure and some disease, a true confounder has an association with the exposure that we want to study, and is at the same time a determinant of the disease. (Thereby it confounds the relation between the exposure and the disease.) However, any "intermediary causal variable", in between exposure and disease also answers that definition. Nevertheless, it is wrong to control for this intermediary or "intervening" variable. If one does, it will take away some legitimate association of the exposure with the disease, because the intermediary variable is always linked somewhat closer to the disease than an exposure that is more remote in the causal chain. What variable is the original exposure (and ultimate cause), and what variable is only intermediary, are matters of judgement. Which is which is a decision by the investigator. No statistical model can discriminate between true confounding, spurious associations or variables that are intermediary in causal pathways. Again, the a priori reasoning predominates.

Confounding and genetic markers

Let me end with today's fashion in clinical epidemiology, which is the advent of genetics in epidemiology and the resulting problems posed by confounding.

Two decades ago, the study of genetic traits looked simple: there was no confounding involved. There is a classic example, dating from the early 1970s, that is often used for teaching. The teacher asks the students: "If you study the influence of ABO blood group on the occurrence of venous thrombosis in middle aged women, can you use new-born male babies as controls" (Hardy & White 1971). Students more or less immediately answer: "Of course not". Then the teacher explains that the true answer is: "Yes, you can, because ABO blood group is not linked to age, nor sex". The distribution of blood groups in new-born boys is the null distribution (or the expected distribution) among middle aged women; new-born boys will serve very well for a blood group comparison.

Life has become more difficult. Take the example of a case-control study demonstrating that homozygotes for the angiotensin T235 variant are at increased risk for cardiovascular disease (Katsuya et al. 1995). The argument hinges on two odds ratios: firstly the simple age and sex adjusted odds ratio for homozygozity for T235 was 1.6, but the odds ratio increased to 2.6 after adjustment for multiple risk factors (besides age and sex, multivariate adjustment was carried out for smoking, diabetes, cholesterol, systolic and diastolic blood pressure, body mass index, current alcohol consumption, treatment for hypercholesterolaemia or hypertension, and the other genotypes studied in the same study). The increased odds ratio upon multivariate adjustment is emphasised since it seems to strengthen the conclusion of an independent causal role for T235 homozygosity.

How can we understand this result? Like in the classic blood group example, it seems evident that T235 is not linked to sex, nor age-linked, nor linked to any of the other things for which the authors adjusted. For example, it is highly unlikely that in the population at large this genetic marker is linked to smoking, or to cholesterol or to blood pressure. Thus, age and sex matching in the study design, or any other adjustment during the analysis is in principle not necessary. If adjustment has any effect on the odds ratio, that must be a result of some association with age and sex *within the data* which is not present in the population at large. Such an association can be completely spurious, but if it is in the data, that might either be like a randomisation that ends with baseline imbalances (smokers or hypertensives or alcoholics are over-represented among the people with the mutation), or be due to a sampling accident of the controls of the study. Then we are back to Cornfield: should we adjust, given that

we cannot check the base-line imbalance? There are no easy solutions.

In principle, at least in a genetically reasonably homogeneous population, we expect no associations between genetic polymorphisms and environmental variables. The next point of discussion is: what constitutes a genetically reasonably homogeneous population? Population geneticists and epidemiologists seem often slightly at odds about this issue. Population geneticists maintain that even within populations that look genetically homogeneous when considered superficially, there might be genetic substrata. If these genetic substrata are also associated with personal or environmental characteristics, this might lead to confounding when studying gene-disease associations. However, epidemiologists have argued that this will only happen in extreme situations. By actual examples and simulations it has been shown that even in situations where high genetic diversity was expected, like among Caucasians of European origin in the US, the assumption that there is no confounding by admixture of genetic subgroups is quite tenable (Wacholder et al. 2000). The study of risk factors at the DNA-level brings back all old discussions and controversies about the nature of confounding in epidemiology. Confounding is still very much with us.

Sources/Acknowledgement

This text is based on a talk given at the symposium "Measuring our scourges – the history of Epidemiology" at Annecy, France, July 1–10, 1996.

Zusammenfassung

Die Geschichte der Störfaktoren oder des Confounding

Confounding ist ein grundlegendes Problem der Vergleichbarkeit und somit schon immer Teil der Wissenschaft. Ursprünglich ein einfaches englisches Wort, hat es in der experimentellen und nicht experimentellen epidemiologischen Forschung spezifischere Bedeutungen angenommen. Der Gebrauch des Wortes geht auf Fischer zurück. Das Konzept wurde in den Sozialwissenschaften unter anderen von Kish noch umfassender entwickelt. Historische Entwicklungen in der Epidemiologie der zweiten Hälfte des 20. Jahrhunderts sind auf Cornfield und Miettinen zurückzuführen. Diese Entwicklungen verdeutlichten, dass die Argumentation mit Störfaktoren/ Confounding ein fast ausschliesslich deduktiver Prozess ist, den wir für die Daten und Datenanalyse anwenden müssen, um zu einer aussagekräftigen Interpretation zu gelangen. Die Probleme des Confounding sind auch in der neueren Anwendung auf dem Gebiet der genetischen Epidemiologie dieselben geblieben.

Résumé

L'histoire de l'effet de confusion

L'effet de confusion est un problème élémentaire de comparabilité et a donc toujours été présent en science. C'était à l'origine un simple mot d'anglais, mais il a acquis une signification spécifique dans la pensée épidémiologique par rapport à la recherche expérimentale et non expérimentale. L'utilisation du mot remonte à Fisher. Le concept a été approfondi dans la recherche en science sociale, entre autres par Kish. Le développement du concept en épidémiologie dans la deuxième moitié du 20ème siècle a été assuré par Cornfield et Miettinen. Ces développements ont mis l'accent sur le fait que le raisonnement sur l'effet de confusion est presque entièrement un processus a priori que nous devons imposer aux données et à l'analyse afin d'aboutir à une interprétation qui ait du sens. Les vieux défis liés à l'effet de confusion se représentent dans leurs applications récentes en épidémiologie génétique.

References

Altman DG, Bland JM (1999). Treatment allocation in controlled trials: why randomise? BMJ *318*: 1209.

Bernard CI (1966). Introduction à l étude de la médecine expérimentale. Reprint of 1865 ed. Paris: Flammarion.

Blalock HM (1964). Causal inference in non-experimental research. Chapel Hill: University of North Carolina Press.

Campbell DT, Stanley JS (1963). Experimental and quasi-experimental designs for research on teaching. In: Gage NL, ed. Handbook of research on teaching. Chicago: R. McNally: 171–246.

Chalmers I (1999). Why transition from alternation to randomisation in clinical trials was made. BMJ *319*: 1372.

Cornfield J, Haenszel W, Hammond EC, Lilienfeld AM, Shimkin MB, Wynder EL (1959). Smoking and lung cancer: recent evidence and a discussion of some questions. J Natl Cancer Inst *22*: 173–203.

Cornfield J (1976). Recent methodological contributions to clinical trials. Am J Epidemiol *104*: 408–21.

Cornfield J (1971). The University Group Diabetes Program: a further statistical analysis of the mortality findings. JAMA *217*: 1676–87.

Fisher RA (1937). The design of experiments. London: Oliver and Boyd.

Greenland S, ed. (1987). Evolution of epidemiologic ideas: annotated readings on concepts and methods. Chestnut Hill, MA: Epidemiology Resources.

Hardy RC, White C (1971). Matching in retrospective studies. Am J Epidemiol *93*: 75–6.

Hill AB (1965). The environment and disease: association or causation? Proc Royal Soc Med *58*: 295–300.

Ibrahim MA, ed. (1979) Proceedings of symposium on the case-control study. J Chron Dis *32*: 1–139.

Katsuya T, Koike G, Yee TW, et al. (1995). Association of angiotensin gene T235 variant with increased risk of coronary heart disease. Lancet *345*: 1600–3.

Kish L (1959). Some statistical problems in research design. Am Sociol Rev *26*: 328–38.

Lilienfeld AM (1982). Ceteris paribus: the evolution of the clinical trial. Bull Hist Med *56*: 1–18.

MacMahon B, Pugh TF (1970). Epidemiology, principles and methods. Boston: Little, Brown and Co.

Mantel N, Haenszel W (1959). Statistical aspects of the analysis of data from retrospective studies of disease. J Natl Cancer Inst 22: 719–48.

Miettinen OS (1970). Matching and design efficiency in retrospective studies. Am J Epidemiol *91:* 111–8.

Miettinen OS (1976). Stratification by a multivariate confounder score. Am J Epidemiol *104*: 609–20.

Miettinen OS, Cook EF (1981). Confounding, essence and detection. Am J Epidemiol *114*: 593–603.

Rothman KJ (1986). Modern epidemiology. Boston: Little, Brown and Co.

Rothman KJ, Greenland S (1998). Modern epidemiology. 2nd ed. Philadelphia: Lippincott-Raven.

Sackett DL, Richardson WS, Rosenberg W, Haynes RD (1997). Evidence-based medicine: how to practice and teach EBM. New York: Churchill Livingstone: 148–9.

Susser M (1973). Causal thinking in the health sciences: concepts and strategies of epidemiology. New York: Oxford University Press.

Taubes G (1995). Epidemiology faces its limits. Science *269*: 164–9.

Urbach P (1993). The value of randomization and control in clinical trials. Stat Med *12*: 1421–31.

Vandenbroucke JP (2001). In defense of case reports and case series. Ann Intern Med *134*: 330–4.

Vandenbroucke JP, de Craen AJ (2001). Alternative medicine: a "mirror image" for scientific reasoning in conventional medicine. Ann Intern Med *135*: 507–13.

Wacholder S, Rothman N, Caporaso N (2000). Population stratification in epidemiologic studies of common genetic variants and cancer: quantification of bias. J Natl Cancer Inst *92*: 1151–8.

Wold H (1956). Causal inference from observational data. J Royal Stat Soc (A) *119*: 28–61.

Wynder EL (1996). Invited commentary: response to *Science* article "Epidemiology faces its limits". Am J Epidemiol *143*: 747–9

Address for correspondence

Jan P. Vandenbroucke, MD PhD
Professor of Clinical Epidemiology
Leiden University Medical Center
PO Box 9600
NL-2300 RC Leiden

e-mail: j.p.vandenbroucke@lumc.nl

Soz.- Präventivmed. 47 (2002) 216– 224