

Significance testing of potential confounders and other properties of study groups — misuse of statistics

by Sven Hernberg, MD¹

An editor sees much that the readers of scientific journals are spared of. One such thing is the surprisingly common malpractice of applying significance testing to differences between properties of a study group and its reference group. It is done to "investigate" whether such differences could be confounders. The reasoning then continues that, if the difference is "significant," the property is a confounder and should be controlled in the statistical model, otherwise not. This reasoning can lead to strange consequences.

Testing of potential confounders

Suppose, for example, that the issue is whether exposure to carbon disulfide causes coronary heart disease. Suppose also that the study design is a follow-up of two cohorts, one exposed, the other unexposed. Age could be a confounder because it is a risk factor of coronary heart disease; the question is now whether the age distributions of the study and the reference cohorts are similar or not. If they are similar, confounding by age is not a problem because, in order to be a confounder, the risk factor must be more or less common in the exposed group than in the reference group. However, suppose that the average age of the exposed group is three years higher than that of the reference group. Can this difference confound the effect of carbon disulfide on mortality from coronary heart disease? The problem is now that some authors try to answer this question by testing the difference for statistical significance.

If the study is large, the difference may easily become "significant," whereas the same difference may well remain "nonsignificant" in a small study. If the result of significance testing now decides whether or not confounding by age should be controlled in the data analysis, age will become controlled when the results of the large study are analyzed. But the consequence of relying on significance testing results is that the very same difference is left uncontrolled in the small study because it was not "significant."

Suppose next that the difference is three years ("significant") in the large study, but a "nonsignificant" six years in another small study. Using the same reasoning, the three-year difference in the large study will be controlled, but the

nonsignificant six-year difference in the small study is left uncontrolled. This would, of course, not be rational, but it would be the logical consequence of letting the result of a significance test decide whether control is needed or not.

A difference in the distribution of a potential confounder (age in this example) between the groups confounds the result exactly as much as such a *large difference* does, irrespective of whether it is "significant" or not. It is not only the magnitude of the difference, but also the size of the study material and to some extent the statistic chosen that decides if something is "significant." If age increases the incidence of coronary heart disease, say, by two cases per 1000 person-years per year of age, the study size has nothing to do with this point. An age difference of three years will, on an average, produce six extra cases, and one of six years will yield 12 cases per 1000 person-years, both in large and small studies, and significance testing will not provide any information on this point.

Moreover, if a confounder distorts a study, the strength of its effect is unique to that study. It depends on how asymmetrically the confounder is distributed across the exposed and reference categories and on how strong its biological effect is relative to the effect of the exposure under study (meaning that, if the exposure has a lower biological potential than the confounder, the confounding effect of a small difference can be strong, stronger than if the biological potential of the exposure were high). In another study the distribution of the same factor could be different, and its confounding effect consequently stronger or weaker. The latter may even shift from positive to negative or vice versa. There is no scientific hypothesis stating that a confounder of one study is a confounder in all other studies. In other words, confounding is bound to time and place and cannot be generalized. Therefore statistical significance testing is not the way to scrutinize potential confounding.

An extreme case is when authors first *match* for a potential confounder, such as age, and then carry out significance testing on the small difference that may prevail in spite of the matching. Almost always this difference is "nonsignificant," what else could it be, and then the authors conclude that there is no confounding by age. Fortunately such "*horreurs*" do not happen often, but I have seen them a couple of times.

¹ Scandinavian Journal of Work, Environment & Health, Helsinki, Finland.

Clear objectives would help one to judge when to test

Age is the simplest example of a potential confounder and thus the preceding discussion was hopefully easy to follow. Perhaps handling, for example, body weight, medication, smoking, drinking, and other life-style factors as potential confounders is slightly more difficult to comprehend. This difficulty would at least explain why such matters are so often described in the beginning of the "Results" section and, alas, tested for statistical significance. I believe such writing is the result of a poor conceptualization of what the study objectives really are. Let me repeat to those who may find this statement astonishing that this is what *editors and referees* find in many manuscripts, *not* what is published after the editing process. One sees exceptions in some journals however. Clear thinking when specifying the objectives of the study — prior to initiating it — would help authors differentiate between matters belonging to the "Materials" and the "Results." (Already in 1820, the great Swedish poet Esaias Tegnér said, freely translated, that "cloudy saying reflects cloudy thinking" and "what you cannot state clearly, you do not comprehend.")

If the objective is to study the health effects of some occupational exposure, then the study hypothesis is "Exposure A causes disease B," but definitely not "Exposed subjects smoke more" or "Exposed subjects drink less." If this turns out to be the fact in a specific study, it is not a "result" but confounding (provided smoking and drinking are risk factors for disease B). The distribution of potential confounders should be presented in the "Materials," not the "Results" section, because these distributions describe properties of the material. In this example smoking and drinking should not be tested for significance, but they should be controlled in the data analysis if their distributions between study subjects and referents differ much. (See below.)

Testing exposure levels for significance

Another rather common example of misusing significance testing is considering exposure levels as a "result." Such misuse often occurs if the exposure is measured by sophisticated and expensive methods, for example, biochemical tests of exposure. Suppose that the objective is to study the effects of lead upon some health parameter, say, nerve conduction velocity. The study hypothesis is then "Lead exposure is toxic to the peripheral nervous system." To test this hypothesis, one usually selects the exposed group from a plant with substantial exposure, say, a lead smelter, and the reference group from an "unexposed" plant. The concentrations of lead in the blood are measured from both groups, and so far this is fine. But then it goes wrong. The results of the blood lead measurements are often presented in the "Results," and the difference, which of course must be great if the groups have been well *selected*, is tested statistically.

Quite expectedly it turns out to be extremely significant. Was there a study hypothesis stating that the exposed workers were more exposed than the unexposed ones? Definitely not in this kind of study. The blood lead levels are hence not a "result" but a *description of the material* telling the readers how "exposed" the exposed group was.

Another matter is that one, having completely different study objectives, may be interested in surveying lead exposure levels in different types of industry. In such a case the blood lead levels are "results" and differences between different worker categories can well be tested. Likewise, if hygienists survey the exposure levels to some organic chemical occurring in paper mills, their study objective is to do just this, and their findings describing the exposure are indeed "results." In this case the "Materials" section is made up of descriptions of the plants, such as exhaust ventilation, amount of the chemical used, and so forth. It is the ability to define the study objectives clearly that helps the researcher to distinguish between "Materials" and "Results" and which explains why the blood lead levels sometimes belong to the former section, sometimes to the latter. As Shakespeare said: "There is nothing either good or bad, but thinking makes it so."

Concluding remarks

How should one then decide if confounding occurs? This is both a question of personal judgment and computing. The general rule is that confounders should be controlled, but that nonconfounders should be left without control to enhance the sensitivity of the study. Formally one can compare the crude relative risk and the relative risk resulting after adjustment for the potential confounder. A difference indicates confounding, and in that case one should use the adjusted risk estimate. If there is no or a negligible difference, confounding is not an issue and the crude estimate is to be preferred. Personal judgment comes into play when what is "negligible" is decided. Some authors show both estimates and leave the decision to the reader.

Why have I said all this? Because the wrong use of significance testing is all too common, while really understanding why significance testing is performed and what it means is all too uncommon. I believe that both the conceptualization of the research problem and the writing of the manuscript would improve if potential researchers and authors would give the matters I have discussed more thought. Editors and referees would be happier, and the probability of getting one's manuscript published would increase.

Acknowledgments

I would like to thank Olav Axelsson and Kari Kurppa for their constructive criticism.