Scand J Work Environ Health 1981;7(4):121-126
Issue date: 1981

"Negative" results in cohort studies--how to recognize fallacies.

by Hernberg S

Key terms: cohort study; epidemiology; fallacy; methodology; negarive study; negative result; occupational medicine; review

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/7330622

Print ISSN: 0355-3140 Electronic ISSN: 1795-990X

Scand j work environ health 7 (1981): suppl 4, 121—126

"Negative" results in cohort studies — How to recognize fallacies

by Sven Hernberg, MD [1]

HERNBERG, S. "Negative" results in cohort studies — How to recognize fallacies. *Scand j work environ health* 7 (1981): suppl 4, 121—126. Negative studies are important in occupational medicine because knowledge on noneffect levels for harmful exposures is pertinent. A truly negative study must (i) be large, (ii) be sensitive, and (iii) have well-documented exposure data. Small and/or insensitive so-called negative studies are uninformative, and negative results can only be related to the actual or lower exposure levels. Some of the causes for falsely negative results are inappropriate design (eg, cross-sectional instead of longitudinal), crude measuring methods, inappropriate type of examination, wrong categories of exposed workers, inclusion of workers with too short an exposure time and too low an exposure intensity in the exposed series, too short a follow-up for diseases with long latency times (eg, cancer), incomplete follow-up, wrong reference category (eg, the general population), poor precision of measuring methods, and insensitive or wrong statistical methods. Finally, the same data may be interpreted in different ways. Correct interpretation requires both knowledge of the subject and apprehension of the fact that errors in design and measurement often tend to mask existing differences. At the most, a small insensitive study may rule out very strong effects.

Key terms: epidemiology, methodology, negative studies, occupational medicine, review.

Cohort studies are designed to discover whether a certain exposure causes disease, and, if so, what kind of disease.

Conversely, the purpose of a study may also be to establish the absence of a hazard (in general or specifically, eg, for cancer). The human mind, especially when contaminated by too much academic dogma, is so constructed that positive answers to these questions are more exciting than negative ones. However, investigators engaged in occupational health research, and especially those involved in providing documentation for the setting of standards, find it very important to establish the *absence* of adverse effects also.

[1] Institute of Occupational Health, Helsinki, Finland.

Reprint requests to: Dr S Hernberg, Institute of Occupational Health, Haartmaninkatu 1, SF-00290 Helsinki 29, Finland.

The first stage is a qualitative one (yes/no), but, if the qualitative answer is positive, the next stage involves a quantitative dimension, ie, the definition of the nonresponse level (if such a threshold exists). Thus negative studies are indeed pertinent for both qualitative and quantitative purposes.

It is not always easy to judge whether a "negative" study is indeed negative or only "nonpositive." This distinction is important and requires both understanding and criticism. Theoretically, the negative requires an infinite number of observations to be proved. This requirement cannot, of course, be achieved in real life; "infinite" must be switched to "large."

A true negative study must fulfill three criteria: (i) it must be large; (ii) it must be sensitive; (iii) it must have well-documented exposure data. Obviously it must

0355-3140/81/080121-6

also be well designed otherwise, and there must be a clear awareness of whether only a specific condition is to be excluded (say, cancer) or if all types of adverse health effects are to be excluded. Only investigations that comply with these criteria can be considered true negative studies.

A small "negative" study is uninformative in the sense that it can exclude the presence of very powerful hazards only. Of course the terms "large" and "small" are diffuse concepts. If "large" is defined as a *"sufficient" number of exposed cases,"* some specification is provided. Then "large" becomes a function of both cohort size, length of follow-up time, and frequency of the disease under study.

Insensitive studies, ie, studies with a crude design or crude measuring methods, are also uninformative. They can be compared to an insensitive laboratory test in clinical diagnosis, eg, the sole use of hemoglobin determination for diagnosing lead poisoning.

Negative results can only be referred to the actual or lower levels of exposure; hence the availability of accurate exposure data is crucial. As is well known, this is seldom the case.

The requirement of a good study design applies of course to all epidemiologic studies. But negative bias is more easily obtained than positive errors; furthermore negative errors are often more difficult to detect. In addition, studies designed to exclude one or some conditions cannot be interpreted as demonstrating "complete safety."

Some of the most common errors leading to falsely negative results will now be reviewed. A strict classification of the type of error according to the guidelines in the preceding outline is difficult, partly because some errors can be classified as belonging to more than one category and partly because some of them, in fact, reflect errors in interpretation of results rather than in design or methods.

Errors related mainly to sample size and type of subjects

Inappropriate design results in an inefficient study which fails to reveal an existing effect. A *small sample size* (and/or reference group) is the most common error of design (but it is often unavoidable because of reasons of availability). In general, such an error is easy to identify, but uncritical authors or unsophisticated readers of a report still often misinterpret the failure to obtain "statistical significance" as being evidence of negativity. In such a setting, only effects leading to very high rate ratios (of the magnitude of, say, 10—20—30) can be excluded. A "negative" result may thus emerge when the rate ratio is moderate only because statistical significance has not been established. But the error may be more subtle. For example, if the disease under study is rare in relation to the cohort size and follow-up time, only *few exposed cases* will be found. Such a study is in fact small, although its number of person-years may seem very impressive and may thereby cause confusion. When the disease is rare, of course, a case-referent study would be the only appropriate design.

Falsely negative results can also be produced if the study centers on *incorrect categories* of exposed workers. For example, if retired workers are left out of occupational cancer studies, crucial information will usually be lost and the "truth," which is actually positive, may appear negative. Likewise, if a study is restricted to retired workers only, it may fail to detect pathognomonic manifestations of the exposure in question in younger age groups (2, 4). For example, an American mortality study (focusing on bronchial cancer) of asbestos workers who had retired during 1941—1967 included those who retired normally at age 65, those who retired before age 65 for personal reasons but lived to age 65 and those who retired prior to age 65 because of disability, but who lived to be 65 (4). However, those who had died and those who had changed employment before the age of retirement were not included. Only one case of malignant mesothelioma was found. As the authors themselves pointed out, this study design was inappropriate for investigating mesothelioma, which often occurs in comparatively young persons. In fact, the authors said they were aware that a number of men who had been excluded by design from the study had died of malignant mesothelioma prior to age 65 (4). This study has

elicited considerable comment, since Borow et al (1) reported that they found 72 cases from hospital records. In a recent study, Henderson & Enterline (9) analyzed the reasons for this discrepancy. They concluded that, in their earlier study, they had indeed missed some cases, partly because malignant mesothelioma is not identifiable as a cause of death from death certificates. On the other hand, several of the cases did not qualify for their cohort (i) because they had died before retirement, (ii) because they were females, (iii) because they did not fulfill the exposure criteria, or (iv) because they had not worked at the plant in question. In any case, a study designed to include only retired workers seems doubtable even in the light of the authors' later explanations. One may also ask to what extent mortality from bronchial cancer was underestimated by this design.

Errors related mainly to insensitivity

All too often *workers with too short an exposure time* and *too low an exposure intensity,* or sometimes even nonexposed, misclassified workers, are forced into the exposed cohort. If the motive is only to increase the cohort size, it is honest, although inappropriate, because this practice usually yields quite the opposite of what the investigator had (probably) intended. In other words, the study becomes insensitive, and an existing effect becomes diluted as long as the cohort is analyzed as a homogeneous population. The only justification for such a procedure is the study of the exposure-response relationship, and this, of course, is important information provided the qualitative aspects of causation are already known. But basing a qualitative negative conclusion only on the outcome of subjects with very low exposure intensity and/or short exposure time is unjustifiable. For example, a recent German survey consisted of 1,960 workers engaged in the production of styrene and polystyrene and exposed to styrene at low intensities ranging from 1 to 8 ppm (6, 19). The fact that no excess cancer mortality was observed inspired the authors to make the absolute statement that "therefore it may be inferred that prolonged exposure to styrene is not a cancer hazard [p 231]" (6). I have used this conclusion, based on a study which also contained other effect-diluting features, first to exemplify how a study can be made insensitive by means of "dilution" of the exposure. Second, the conclusion shows severe misinterpretation of a negative finding which should be related to actual or lower exposure levels only. The fact that the exposure levels in the manufacture of, eg, laminated reinforced plastics may reach several hundred parts per million illustrates just how uninformative the German study was.

Whenever diseases with long latency periods (eg, occupational cancer) are studied, a falsely negative result emerges if the follow-up time is *too short* and if allowance is not made for the *latency period* in the computation of person-years. These are also examples of how insensitivity is produced. Unfortunately, however, most latency periods are unknown; therefore no fixed rules can be given. It is believed that many cancers have latency periods of several decades, especially if the exposure is low; this is a difficult practical problem especially in prospective cohort studies. The urgency for obtaining quick results must not interfere with the requirement of allowing for a minimum latency period before the computation of person-years is begun. An arbitrary rule would be half the supposed mean latency period of the cancer in question. In this context it should be stressed that the latency period should be computed separately for each individual. The "general" 5-a latency period proposed by Fox (5) is very likely too short in a biological sense.

Errors of selection rarely bias cohort studies. These errors are typical of cross-sectional and case-referent studies; they occur in cohort studies only when there has been selective turnover and simultaneous failure to trace the dropouts. Incomplete follow-up may then result in negative bias, especially when the disease is accompanied by subjective symptoms. Examples of such diseases are chronic bronchitis, coronary artery disease, lumbar disease, etc. This error is so obvious that it deserves no further comment in this context.

Errors of information, if systematic, cause bias which can be either positive or

negative. Again, such errors are fairly easy to recognize. In cohort studies the most common source of these types of errors is a more intense observation of the exposed cohort than of the reference population; such a situation usually results in a positive, not a negative, bias. On the contrary, *random errors* mask existing effects. Random errors often arise from poorly standardized measuring methods. Poor analytical precision, especially of the determinant variable, decreases existing group differences and levels off the regression slope in exposure-response studies.

A choice of *insensitive indicators* of illness also results in falsely negative results when the purpose of the study is to demonstrate "complete" safety. For example, mortality is too crude an indicator of health risks associated with exposure to many conditions in today's industry. Failure to find increased mortality is often misinterpreted as total absence of health hazards. This misinterpretation becomes especially dangerous when the general population is used as the reference category. (See the following discussion.)

Insensitive or otherwise incorrectly selected statistical methods may also fail to detect statistically significant differences which actually exist. There has recently been a dispute in *Science* about the justifiability of the methods used in a study of the mortality of lead workers (3, 7, 12).

Errors of comparison are indeed common in cohort studies. The general outline of such a study design is, as we all know, to compare the exposed cohort with a nonexposed one, which should be similar to the exposed cohort in all *relevent* aspects. Actually, the investigator is not at all interested in the morbidity of the reference group as such. That group is studied only to give an estimate of *what would have happened in the exposed cohort had there been no exposure.* Finding valid reference groups is by no means easy. Furthermore, the reference group doubles the costs of the investigation.

These two practical considerations have resulted in a vast number of studies which do not meet the criteria for validity of comparison. First, the reference group may not be *completely nonexposed.* For example, a recent study by Spivey et al (18) was designed to investigate the effect of lead upon the peripheral nervous system. In this investigation the reference group was exposed to 8-h time-weighted-average concentrations of lead in air ranging up to 51.0, 32.2, and 35.2 mg/m^3. The mean blood lead level of the reference group was 22.0 μg/100 ml (1.06 μmol/l), with a standard deviation of 5.9 μg/100 ml (0.28 μmol/l). Furthermore, the reference group was obtained from three plants in which occupational exposure to aluminum occurred, and this metal is suspected of being neurotoxic (17). No major differences in peripheral nervous function were found, but because of a substantial error of comparison leading to insensitivity (the reference group not being nonexposed), no conclusion can be drawn from this study. This example also illustrates how a reference group can be exposed to agents which may possess properties similar to the one under study. Second, in cohort studies, *social or other factors* with no causal connection to the problem at issue may distort the comparison. One may find an unexpectedly "high" morbidity in an incorrectly selected reference category in which, for various reasons, workers leaving more demanding jobs have accumulated. The "detection limit" of the study is thereby elevated.

Perhaps the most serious comparison invalidity arises when the *general population* is used as the reference category. This unfortunate practice is mainly caused by economic necessity, since the trivial bias pompously called the "healthy worker effect" is by now so well known to all epidemiologists that everybody tries to avoid it. Unfortunately, however, most researchers, including myself, must often commit this error for practical reasons (eg, 13). The "healthy worker effect" has been discussed in several contexts (8, 10, 11, 14, 20); hence it is no longer necessary to go into detail about it. In this context it should be stressed that the bias introduced is almost always one that hides existing differences. Efforts to analyze the nature of the "healthy worker effect" have given rise to the image of a manifold and very interesting phenomenon. In my opinion, however, there is nothing exciting about the "healthy worker effect"; the term could actually be abandoned and replaced

by its real name, ie, *comparison bias.*

The general population does not fulfill even the most elementary criteria for a reference group, especially if one keeps in mind that such a group should reflect what would have happened in the exposed cohort had there been no exposure. The general population includes inactive persons, some exposed workers from other plants, etc, and it is socially heterogeneous. There are no ways to control confounding, since data on possible confounders (eg, smoking habits, use of medicines, leisure-time activities, etc) cannot be obtained. No general rules for adjusting for the so-called "healthy worker effect" can be given, since its strength varies according to the length of follow-up, the age structure of the cohort, race, the type of disease of interest, and a multitude of occasional factors such as the opportunity for turnover, the economic situation prevailing in the nation, etc. Hence the "healthy worker effect" poses a serious methodological problem whenever occupational groups are compared with the general population. The interpretation of such studies is therefore extremely difficult unless the effect under study is very obvious. Tests of statistical significance are completely uninformative, even misleading, in such situations. The only recommendable remedy is to avoid comparisons with the general population.

Errors related mainly to insufficient exposure data

Failure to relate a negative finding to well-documented exposure data is the most common type of error in "negative" epidemiologic studies. This omission is due to the well-known fact that good exposure data are rarely available, at least not in retrospect. The problem is far less serious when qualitative positive conclusions are concerned, although exposure-response data will become inaccurate or even impossible to derive. But for negative conclusions poor exposure data are disastrous.

Miscellaneous errors

The effects studied may be *inappropriate* or irrelevant to the exposure in question. For example, a recent American study compared lead workers with present or 5-a mean blood lead levels of 60—80 μg/100 ml (2.89—3.86 μmol/l) to those with levels below 60 μg/100 ml (2.89 μmol/l) (16). Many biochemical tests and other indicators of illness were impressively employed, among them a "comprehensive" list of no less than 29 laboratory tests. No statistically significant differences were found between the groups. The authors concluded that "there were no significant differences in the health of workers with blood lead concentrations between 60 and 80 μg/dl and those whose blood lead concentrations were lower than 60 μg/dl [p 617]." They also stated that "it is our opinion that the current blood lead standard of 80 μg/dl can be kept, unless more data will support the OSHA proposal [p 617]." A very remarkable flaw in this study was that no tests whatsoever describing direct disturbances of protoporphyrin synthesis were included and that no measurements of neurophysiological functions or psychological performance were performed. Making a general negative statement without studying these parameters, which are generally known to be the critical effects of lead toxicity, is a flagrant example of how a biased selection of examinations can be abused to substantiate some nonscientific cause.

In some instances the cohort may be exposed to a *mixture of agents,* among them *antagonistically acting* materials (eg, lead and zinc, cadmium and zinc). For example, a Danish study on the peripheral nervous function of workers exposed to lead did not show any significant slowing of conduction velocities, but closer scrutiny of the report reveals that the workers were concomitantly exposed to zinc, which is a known antagonist to lead toxicity (15).

Concluding remarks

This recapitulation of some common errors producing falsely negative results raises the question of why such studies are carried out and published at all. In many cases the reason is apparently a lack of epidemiologic education combined with the naive belief that epidemiology is an easy game. In some cases it may be

that the researcher is indeed biased and for some reason or other intentionally tries to fabricate a "negative" study. But the vast majority of so-called negative results arise from various realities of life: from the unavailability of enough exposed subjects and/or valid reference groups; from too short an industrial use of some agent to demonstrate long-term effects; from difficulties in tracing dropouts; from a lack of good exposure data, etc. In many instances the options then are either to do nothing at all or to try to make the best of a defective situation. Of course the first option should be given serious consideration if the indications are strong. But since the perfect epidemiologic study has yet to be carried out, it is often quite correct to initiate a study even if all validity aspects cannot be totally met.

Then the key question becomes one of correctly interpreting the results. The reader of the report must also have this opportunity; hence a thorough description of the design, the material, and the methods is an absolute requirement. Articles with deficient data on these matters are suspect and informative only of the author's level of understanding (sometimes also the editor's), not of the problem at issue. Honest and thorough discussion of possible sources of errors is an indication of maturity and insight, not weakness. Interpretation of the results is also a test of these properties. Interpretation must take into account both the impact of errors and exactly what type of effects the study was designed to exclude. General qualitative negative statements based on observations which are too scant, too crude, incorrectly selected or unstandardized and which do not allow for duration and intensity of exposure are something we can well do without.

References

1. Borow M, Conston A, Livornese L, Schalet N. Mesothelioma following exposure to asbestos: A review of 72 cases. Chest 64 (1973) 641—646.
2. Collins JF, Redmond CK. The use of retirees to evaluate occupational hazards. J occup med 18 (1976) 595—602.
3. Cooper WC, Gaffey WR. Mortality of lead workers. J occup med 17 (1975) 100—107.
4. Enterline P, DeCoufle P, Henderson V. Mortality in relation to occupational exposure in the asbestos industry. J occup med 14 (1972) 897—903.
5. Fox AJ. Vinyl chloride and mortality? Lancet 2 (1976) 416—417.
6. Frentzel-Beyme R, Thiess AM, Wieland R. Survey of mortality among employees engaged in the manufacture of styrene and polystyrene at the BASF Ludwigshafen works. Scand j work environ health 4 (1978): suppl 2, 231—239.
7. Gaffey WR. Occupational lead exposure: What are the risks? Science 208 (1980) 130—131.
8. Goldsmith J. What do we expect from an occupational cohort? J occup med 17 (1975) 126.
9. Henderson VL, Enterline PE. Asbestos exposure: Factors associated with excess cancer and respiratory disease mortality. Ann ny acad sci 330 (1979) 117—126.
10. Hernberg S. Epidemiology in occupational health. In: Carl Zenz, ed. Developments in occupational medicine. Year Book Medical Publishers, Chicago, IL 1980, pp 1—40.
11. Hernberg S. Evaluation of epidemiologic studies in assessing the long-term effects of occupational noxious agents. Scand j work environ health 6 (1980) 163—169.
12. Kang HK, Infante PF, Carra JS. Occupational lead exposure and cancer. Science 207 (1980) 935—936.
13. Koskela R-S, Hernberg S, Kärävä R, Järvinen E, Nurminen M. A mortality study of foundry workers. Scand j work environ health 2 (1976): suppl 1, 73—89.
14. McMichael AJ. Standardized mortality ratios and the "healthy worker effect": Scratching beneath the surface. J occup med 18 (1976) 165.
15. Paulev P-E, Døssing M. Motor nerve conduction velocity in asymptomatic lead workers. Int arch occup environ health 43 (1979) 37—43.
16. Ramirez-Cervantes B, Embree JW, Hine CH, Nelson KW, Vernes MO, Putnam RD. Health assessment of employees with different body burdens of lead. J occup med 20 (1978) 610—617.
17. Rozas VV, Friedlich KP, Rutt WM. Progressive dialysis encephalopathy from dialysate aluminum. Arch int med 138 (1978) 1375—1377.
18. Spivey GH, Brown CP, Baloh RW. Subclinical effects of chronic increased lead absorption — A prospective study: I. Study design and analysis of symptom. J occup med 21 (1979) 423—429.
19. Thiess AM, Friedheim M. Morbidity among persons employed in styrene production, polymerization and processing plants. Scand j work environ health 4 (1978) suppl 2, 203—214.
20. Tola S, Hernberg S. "The healthy worker effect." In: Proceedings of the conference on epidemiologic methods for occupational and environmental health studies, Washington DC, December 2—5 1979. (in press).