Scand j work environ health 6 (1980) 163—169

Evaluation of epidemiologic studies in assessing the long-term effects of occupational noxious agents[1]

by Sven Hernberg, MD[2]

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. A review.

Key terms: epidemiologic methods, occupational toxicology, healthy worker effect, negative studies, review.

Epidemiology is predominantly a nonexperimental science, and therefore the evaluation of cause-effect relationships is difficult. While an experimenter can actively manipulate experimental conditions by (randomly) allocating the individuals into exposed and nonexposed groups, the epidemiologist has no control of these factors. His only option is to observe what nature has accomplished. This is the fundamental distinction between experiments and epidemiology, and it weakens the inference concerning a causal relationship between two phenomena. Hence the interpretation of the results always becomes a matter of probability.

The evaluation of an epidemiologic study is, to a great extent, dependent on how well the investigator has succeeded to avoid the errors that weaken the validity of nonexperimental research. The perfect epidemiologic study still remains to be done. All studies published to date contain errors, the magnitude and direction of which must influence the interpretation of the results. There are two main types of error, namely, random error and systematic error. A systematic error is one that distorts the results of a study in such a way that hypothetical replications of it would produce the same results so that a false conclusion is reached. A random error may distort the study on any one occasion, but the average distortion is predictable via a probability model. Random errors decrease the sensitivity of a study to detect an effect which actually exists.

Validity aspects

The validity of a study means the lack of systematic error. Validity has two dimensions, internal and external. The former refers to how "true" the results of a study are with respect to the study itself. The latter stands for the generalizability of the results beyond time and place, that is, to other similar situations, and, finally, to the sphere of scientific theories. Consider, for example, a mortality study which has shown an excess mortality of bronchial cancer for workers exposed to asbestos. The study has internal validity if systematic errors can be ruled out and external validity when it allows for the

[1] Lecture delivered at the Carlo Erba Foundation, Milan, Italy, on 15 September 1980.

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

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

0355-3140/80/030163-7

formulation of the hypothesis that exposure to asbestos *in general* causes bronchial cancer.

Internal validity can be broken down into the following three components (2):

Validity of selection
Validity of information
Validity of comparison
(a) validity of reference entity
(b) unconfoundedness of comparison.

When the results of epidemiologic studies are evaluated, it is useful to start the process with a check of the extent to which these conditions are met.

Validity of selection means that the probability of a subject being nominated for the study must not depend in a systematic way on the disease or exposure under study. This error especially distorts cross-sectional and case-referent (case-control) studies. In a case-referent study a selection bias may arise in the following way: Suppose that somebody at a clinic for occupational diseases wishes to study the effects of various environmental factors, among them occupational exposures, on the occurrence of gastritis. Patients with this disease are then defined as cases and patients with, for example, lumbar disease as referents. It would not be surprising if lead exposure would be overrepresented among the cases. However, this result would be an overestimate of the role of lead exposure in the etiology of gastritis *in general.* Workers exposed to lead and having epigastric pain are more likely to be admitted to this specialized clinic, because plant physicians often suspect lead poisoning in such a situation. In this example the reasons for selection bias were (i) the fact that the hospital specialized in occupational medicine and (ii) the connection between lead exposure and epigastric pain was known in advance. Selection bias must be avoided at the planning stage of a study since no method exists with which to control it during the data handling.

Validity of information means that the inaccuracy of the information gathered from both the cases and the referents, or from both the exposed and nonexposed, is similar. Only asymmetrical inaccuracy affects validity. In contrast, the sensitivity (the power to detect a causal association, if present) of the study suffers from symmetrically inaccurate information. Information bias may affect all types of epidemiologic investigations. However, case-referent studies, which rely on information from questionnaires and interviews rather than from measurements, are especially vulnerable to this source of error. Information errors can usually be overcome, providing the investigator is aware of this source of error and takes the necessary precautions, as, for example, the "blinding" of readings, interobserver error control, calibration of equipment, etc.

Validity of comparison affects all types of epidemiologic studies equally. An ideal *reference group* should share all characteristics of the study group relevant to the problem at issue, except for the properties that define the groups. These are, of course, the exposure in a follow-up study and the disease in a case-referent study. This condition is sometimes extremely difficult to achieve, not the least because of practical circumstances, and in such a case the greatest emphasis must be placed on those properties that are likely to possess the strongest distorting effects.

Confounding is the other component of comparison bias. A confounder is an extraneous factor that is intermixed with the scientific problem. Because of this intermixing, the confounder disturbs the assessment of the effect under study. To have this confounding effect, a factor must (i) be a causal risk factor of the illness in general and (ii) be statistically associated with the exposure, but only in the particular study. For example, smoking is a confounding factor in the study of whether exposure to chromates causes bronchial cancer *only* when the smoking habits of the exposed cohort are systematically different from those of the reference group in that particular study. Confounding can be either positive or negative. It becomes especially problematic when the rate ratio is rather low (on the order of 1.1 to 3). When the rate ratio is high, it is not likely that such a strong confounding would pass undetected. The control of confounding takes place either at the planning stage of a study (restriction, matching) or at the data handling stages (stratification, modeling), self-

evidently under the assumption that the relevant data are available.

The healthy worker effect

In this context some comments about the "healthy worker effect" may be pertinent. This popular, but conceptually vague, term describes the sum of the errors arising from the comparison of the mortality experience of an exposed cohort to that of the general population. The general population is heterogeneous, not completely free from the exposure under study, and rarely, if ever, represents the same social stratum as the exposed cohort. In other words, the general population does not fulfill even the most elementary requirements of comparison validity. The reasons for using such a reference category so often, in spite of this, are the practicability involved and economy. An ad hoc reference cohort would double the costs of the investigation, and, besides, valid and suitable reference cohorts are difficult to find.

The healthy worker effect causes the standardized mortality ratio (SMR) of the exposed cohort to fall well below 100 if no life-shortening occupational hazard exists. In fact, figures on the order of 60—90 have often been reported; for example, 87 for rubber workers in the 40- to 64-year age range, 82 for steel workers, 89 for talc miners, 65 for workers in a communications company, and 90 for foundry workers. These figures are obvious underestimates of the "true" mortality (4). The main reason for the better-than-expected mortality experience of occupational groups lies in the fact that the general population also includes unemployable and unemployed persons. Among them are those in institutions, those with congenital anomalies, those handicapped during childhood, those otherwise ill at the time job seeking commences, and those unemployed or with unstable employment. All these groups have a higher mortality rate than the active population. Kitagawa & Hauser (5) reported that those American white males unemployed in 1950 or later (4.4 % of the population) had an SMR of 240. An additional 4 % with no occupation showed an SMR of 125. A Finnish study (7) showed that the SMR for occupationally inactive men was 275. It is quite clear that a comparison of an active, often health-selected group of workers with the general population, containing such fractions, cannot give correct results.

It has been proposed that a correction coefficient of 1.1 could be used to yield a "true" SMR (3). However, this procedure would be an oversimplification. The healthy worker effect is not constant but varies depending on a number of circumstances (6). It is strongest in younger age groups, and it declines with age until it is no longer significant in the postretirement age range. It is stronger for men than for women, stronger in higher social categories than in lower strata, and strongest in the beginning of employment. Even more important, it is different for different causes of death. In general, diseases with silent early stages and a rapid fatal course do not cause a healthy worker effect (except for during the first one or two years after cohort identification). Cancer is a typical example. The symptoms appear late, the early diagnosis of asymptomatic cancer is not well developed, and there are no means to predict who will get cancer before symptoms or signs appear. In contrast, a high risk for death due to coronary artery disease can be established in advance by means of early diagnosis of the disease or identification of its risk factors. Furthermore, self-selection occurs because of symptoms or earlier warnings from physicians to avoid demanding jobs.

The relative strength of all these various components is not similar in all situations, and hence no general rule can be given. It should also be noted that the healthy worker effect and the effect under study may sometimes mask each other so that SMRs on the order of approximately 95 to 105 result. What still weakens the use of the general population is that no details of possible confounders are known. For example, one cannot obtain information on smoking habits from national mortality statistics. For these reasons it is evident that the healthy worker effect poses a serious methodological problem when occupational groups are compared with a general population. The interpretation of such studies is therefore very diffi-

cult unless the effect under study is very outspoken. It is quite clear that tests of statistical significance are completely uninformative, even misleading, in such situations. Whenever possible, the use of a more appropriate ad hoc reference group is the best solution to the problem. If such a group cannot be utilized for practical or economical reasons, the *active* general population is a better reference category than the total general population. The healthy worker effect can be further decreased if the calculation of person-years does not start immediately at the time of cohort identification but, say, five or ten years later. Alternatively, separate comparisons can be made at different intervals after the identification. The mortality rate should also be compared separately for different causes and within different age groups. Finally, whenever possible, comparisons should be made within the cohort itself between different categories of exposure intensity (8). These rules apply to the investigator; the reader of the report unfortunately has no such possibilities left and must rely on his intuition and his knowledge of the substance. However, these are not always sufficient, and hence many mortality studies using the general population as the reference remain rather uninformative.

Negative results

The healthy worker effect is by no means the only problem rendering the interpretation of cohort studies difficult. Another central problem is the evaluation of "negative studies." Negative studies are at least as important as positive ones in occupational medicine because it is extremely important to be able to define a noneffect level for harmful exposures. However, a clear distinction must be made between truly negative and "nonpositive" studies. A true negative study must fulfill three criteria. It must be *large* and *sensitive* and have *well-documented exposure data.*

Only investigations that comply with these criteria can be considered true negative studies. Small so-called negative studies are more or less uninformative, and the same is true for insensitive studies, that is, studies performed with a crude design or crude measuring methods. It is necessary that the exposure level be well documented because a result can be negative only in relation to actual or lower exposure intensities and durations. For example, if no excess cancer mortality can be found among workers exposed to styrene for a few years at intensities ranging between 1 and 5 ppm, the finding is completely uninformative regarding higher exposure intensities and longer exposure times and can therefore under no circumstances be used as an argument in favor of the assumption that styrene possesses *no* carcinogenic properties. Small materials also cause nonpositive results in case-referent studies. This situation occurs when the *exposure under study* is "rare" in the source population. In this context "rare" is a relative concept, the prevalence of detectable exposure being inversely proportional to the number of subjects in the case and referent series. Hence, if the number of cases and referents is small, the exposure must be common to have a chance to be detected. The larger the number of subjects, the "rarer" the exposures can be and still be detected. For example, in a study comprising 200 lung cancer cases, chromate exposure is not likely to show up because of the rare occurrence of this exposure in the general population, whereas cigarette smoking certainly will emerge.

False negative and false positive studies

It is regrettable that epidemiologic studies are such blunt instruments for detecting long-term effects of noxious agents. This insensitivity produces errors in the direction of negative; in other words, the studies fail to detect existing effects. With this in mind, the interpretation of negative studies is especially difficult. As for all epidemiologic studies, the prerequisite is that the authors have described their material and methods so thoroughly that the reader is able to make his own independent evaluation. Poorly documented articles are suspect and should be considered as uninformative.

It may be of some relevance to discuss some of the most common causes for falsely negative results. The intention is of course not to a provide a recipe of how to produce such results, rather to help the critical reader in his evaluation. An *inappropriate design* may result in an inefficient study which fails to reveal an existing effect. For example, if the disease under study is rare in relation to the cohort size and follow-up time, only a few cases will be found. In such a setting, only very high rate ratios (of the magnitude of 20—30) can be shown. Similarly, if the exposure in the source population is rare in relation to the number of cases in a case-referent study, the likelihood of that exposure showing up among the cases and referents is extremely low, as discussed before. *Crude measuring methods* may also result in falsely negative results. For example, mortality is too crude an indicator of health risks associated with exposure to organic solvents. The *type of examination* used may also be inappropriate. For example, the lack of finding effects on liver function or cardiovascular performance does not prove that a certain intensity of lead exposure is *completely* without health effects. Falsely negative results can also be produced by means of using *wrong categories* of exposed workers. For example, if retirees are left out of occupational cancer studies, it is very likely that much information will be lost. Moreover, if workers with too *short an exposure time* and too *low an exposure intensity,* or even nonexposed workers, are included in the exposed cohort, a dilution of the effect results. This is a very common error, and such inclusions are usually made in order to increase the cohort size and thereby get a larger material. However, the effects of this procedure may be quite the opposite of what the investigator intended. If the follow-up time is *too short* when diseases with a long latency time are studied (for example, cancer) or if the follow-up is *incomplete* (a high proportion of persons who have not been traced), the result can also be falsely negative. In some instances the cohort may be exposed to a mixture of agents, among them antagonistically acting materials (such as lead and zinc or cadmium and zinc). Negligence to allow for latency times in person-year computations in cohort studies, and in exposure histories in case-referent studies, may also introduce negative bias, although the duration of the study as such may be appropriate.

The choice of *reference category* is critical. The healthy worker effect is by no means the only possible source of error. The reference group may not be *completely nonexposed* (for example, lead) and can also include subjects exposed to *other agents* with effects similar to those of the exposure under study. In mortality studies, *social or other factors* which have no causal connection with the problem at issue may distort the comparison. One may find an unexpectedly "high" mortality in a wrongly selected reference category, or alternatively the mortality may be unexpectedly "low." Not only the exposed cohort but also the *reference group* may be *too small,* especially in relevant strata. The possibilities for hidden negative confounding are also manifold.

Poor precision of the *measuring methods* tends to mask existing effects. The same can be said of an *insensitive design,* and of *random errors* in general. Insensitive or wrong *statistical methods* may also fail to detect statistically significant differences which actually exist. And, finally, the same data may be *interpreted in different ways.* For example, if an SMR of 95 is found for foundry workers, the result can be interpreted so that no life-shortening exposure exists. On the other hand, another interpretation is that the effects of exposure and the healthy worker effect have masked each other and that, consequently, the SMR of 95 suggests an increased mortality. Unfortunately both authors of original reports and, especially, those who write review articles are seldom able to interpret the results of small negative studies correctly. How often have you not read that "A could show an increased mortality but B could not," without any further discussion! The truth may be that B's study was so small that its negative result did not prove anything at all. At the most, small "negative" studies may rule out very strong effects (rate ratios of, say, over 7).

Self-evidently also *falsely positive studies* exist. The most common reasons for this are information bias and com-

parison bias. In the same manner as one can select too "unhealthy" a reference group, one may also select too "healthy" a one. Then falsely positive results arise. Falsely positive findings may also arise through *confounding*. Large studies are the most dangerous in this respect, since even a small difference yields statistical significance. Then even a weak confounding factor may be decisive. By contrast, the difference must be so large in smaller studies that a statistical significance due to confounding is easily revealed, as already discussed. For example, a rate ratio of 1.5 is significant ($p < 0.01$) in a study of 1,000 exposed and 1,000 referents if the mortality is 10 %, whereas the same level of significance in a study of 50 exposed and 50 referents requires a rate ratio of 11. Hence, in *large studies* the magnitude of the rate ratio is more important than the statistical significance of the difference, while in *small studies* the magnitude of the rate ratio is very much influenced by chance and the statistical significance is therefore more important.

The most important reason for falsely positive results in *case-referent studies* is probably a special type of information bias, namely, the so-called memory bias. Consider, for example, a case-referent study of congenital malformations. The cases are mothers of malformed children and the referents are mothers of healthy babies. Data on possible harmful exposures during pregnancy are collected by means of interviews. Such exposures would be use of medicines, past infections, occupational exposures, radiological examinations, traumas, etc. It is a well known fact that mothers of malformed babies often feel guilty and brood about why such a disaster has happened. Hence, they remember and report all possible exposures in detail (provided that their feeling of guilt does not force them to hide some facts and provided there is some advance information of a possible connection between certain exposures and the birth defects). However, mothers of healthy babies have no reason to dwell on such matters and therefore give a less-detailed history. Consequently, such information causes a positive bias by being asymmetrically inaccurate. Insofar as occupational exposures are concerned, this bias can and should be overcome if other sources of information are utilized — namely, employers. If the exposure history relies mainly on the employers' files and other data obtained at the workplaces, this type of memory bias can be overcome or at least substantially reduced.

Epidemiology and causality

It may once more be repeated that definite cause-effect inferences cannot be derived from nonexperimental research. Causality, therefore, can be viewed in terms of probability only. The following circumstances support the causality of a relationship (1):

1. The *exposure must precede the illness.* When latency periods are involved, the exposure must have commenced early enough, at least earlier than half the mean latency period. Since the exact length is rarely known, this, too, becomes a matter of judgement.

2. The *stronger the association* (a high rate ratio), the greater the probability of a causal association. An exception is "small" series because their rate ratios are subject to substantial random variation.

3. The presence of an *exposure-response relationship* is often said to support the causality of an association. However, it should be noted that a confounding factor may often be so intermixed with the exposure that the effects of the two cannot be separated. Concomitant exposure to a variety of metals in many smelters is a typical example. In this case, the carcinogenic effects of one metal cannot be separated from the possible effects of the others, and the exposure-response relationship may become confounded and cannot be considered to support a causal connection between the disease in question and a particular exposure.

4. The credibility of results in the light of *prior knowledge* usually supports causality. For example, if animal experiments

have shown an agent to be carcinogenic, epidemiologic evidence of excess cancer among exposed workers becomes more credible.

5. Known or possible *explanatory biological mechanisms* speak strongly in favor of the causality of an association. In contrast, it is difficult to have much confidence in associations that have no explanations.

6. Perhaps the most conclusive evidence of causality is provided when a *change in the exposure* brings about a *change in morbidity*. This type of evidence can be obtained in interventive epidemiologic studies, and its strength of evidence is connected with the similarities between intervention and experiment.

There is a strong need for good epidemiologic studies. Unfortunately, however, epidemiologic research is so demanding and the sources of error so manifold that conducting a good and valid epidemiologic study requires great skill on the part of the investigator. There is a great lack of opportunities for education throughout the world. Self-education, to a great extent using the trial-and-error method, which is an expensive game, has been the school for many epidemiologic investigators. Training facilities for epidemiologists should therefore be developed as soon as possible in as many countries as possible.

Acknowledgment

As usual, when writing epidemiologic articles, I have benefited greatly from Prof Olli S Miettinen's advanced courses on epidemiologic methods in occupational health, given 1972—1980 in Helsinki.

References

1. Bradford Hill A. Principles of medical statistics. 8th ed. The Lancet Limited, London 1967, chapter XXIV.
2. Campbell D. Factors relevant to the validity of experiments in social setting. Psychol bull 54 (1957) 297—312.
3. Goldsmith J. What do we expect from an occupational cohort? J occup med 17 (1975) 126.
4. Hernberg S. Epidemiology in occupational health. In: C. Zenz, ed. Developments in occupational medicine. Year Book Medical Publishers, Chicago IL 1980, pp 1—40.
5. Kitagawa EM, Hauser PM. Differential mortality in the United States. Harvard University Press, Cambridge, MA 1973.
6. McMichael AJ. Standardized mortality ratios and the "healthy worker effect": Scratching beneath the surface. J occup med 18 (1976) 165.
7. Sauli H. The socio-economic aspects of occupational mortality in Finland. Nord företagshälsovård 2 (1979) 72—85.
8. Tola S, Hernberg S. "The healthy worker effect." Paper presented at the Conference on Epidemiologic Methods for Occupational and Environmental Health Studies, 2—5 December 1979, Washington, DC.