Fact and Fiction in Occupational Epidemiology

The Lucas Lecture 1982

SVEN HERNBERG, MD
Scientific Director, Institute of Occupational Health, Helsinki, Finland

A fact, says the Concise Oxford Dictionary, is ‘a thing certainly known to have occurred or to be true’; its definition of fiction is ‘a thing feigned or imagined, invented statement or narrative’.

The purpose of scientific research is to find out the facts about life without letting fiction blur the vision. In experimental research the conditions can be standardised and the subjects randomly allocated into exposed and unexposed. A proper experiment then gives a good opportunity of distinguishing between fact and fiction. Nevertheless, epidemiology is a non-experimental science. ‘Nature’, not the researcher, decides who will become exposed and who will not, without any random allocation. Therefore, epidemiological studies are much more vulnerable to systematic errors than are experiments. Actually, it is doubtful if any epidemiological study published so far is completely free from such errors.

The design of a valid epidemiological study is an important but difficult task for the investigator, requiring both skill and patience. No wonder that readers without specific epidemiological training find it hard to distinguish between fact and fiction in epidemiological articles. Evaluation of an epidemiological study requires a knowledge of the topic and of epidemiological principles to detect how well the investigator has succeeded in avoiding the errors that weaken the validity of non-experimental research.

Validity

A study is said to be valid if systematic errors are absent or insignificant. There are two aspects of validity, 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 general ‘truth’ of the results in other similar situations, and in the sphere of scientific theories. Consider a mortality study that shows excess mortality from coronary artery disease for workers exposed to carbon disulphide. The study has internal validity if systematic errors can be ruled out, and external validity when it allows for the formulation of the hypothesis that exposure to carbon disulphide in general is causative of coronary artery disease. Of course, external validity is meaningless if the internal validity is poor.

Internal validity has the following three major components [1].
1. Validity of selection.
2. Validity of information.
3. Validity of comparison: (a) validity of the reference group; (b) unconfoundedness of comparison.

Selection

Validity of selection means that the probability of a subject being nominated for study must not depend in any systematic way on the disease or exposure under study. Avoiding this error is a problem, especially in cross-sectional and case-referent (case-control) studies, and it must be eliminated at the planning stage since it cannot be controlled during data handling. For example, a cross-sectional study may show that there is no major difference in the prevalence of low back pain between lumberjacks and foremen. This, of course, does not free lumbering from having harmful effects on the back. The explanation is that those with a bad back cannot stay in the demanding lumbering work and therefore they move selectively away from this work, some of them perhaps becoming foremen.

Information

Validity of information means that the inaccuracy of the information received from both the cases and the referents, or from both the exposed and non-exposed, is symmetrical. Asymmetrical inaccuracy weakens validity. In contrast, symmetrically inaccurate information decreases the sensitivity (the power to detect a causal association, if present) of the study. Information bias may affect all types of epidemiological investigations. Nevertheless, case-referent studies, which rely on information from questionnaires and interviews rather than from measurements, are especially vulnerable to this source of error. In general, it is thought that the cases ‘remember’ better than do the referents, although I am not aware of any study that documents this belief. While the possibility of asymmetric recall of exposures must always be kept in mind, this source of error must not be exaggerated. For example, there must be a great difference between
remembering past medicine intake and, say, ‘exposure’ to lumbering work. In the setting of occupational health it is usually possible to double-check the information given by both cases and referents by contacting the employer. There are many other means of overcoming information bias, such as ‘blinding’ of readings, inter-observer error control, calibration of equipment, and so on.

Comparison

Validity of comparison affects all types of epidemiological studies. The idea is that the study group should be compared with a valid reference group which should be similar to the study group in all aspects relevant to the problem at issue, except for the exposure in follow-up studies and the disease in case-referent studies. In other words, the reference group is there to give an estimate of what would have happened in the exposed group if there had been no exposure in follow-up studies, or to show the general exposure pattern in the source population of the cases in case-referent studies.

Confounding is the second component of comparison bias. A confounder is an external factor that is intermixed with the scientific problem. Because of this intermixing, the confounder disturbs the assessment of the effect being studied. To have this confounding effect, a factor must be a causal risk factor of the illness in general, and statistically associated with the exposure, but only in the particular study. Thus, smoking is a causal risk factor of bronchial cancer. Nevertheless, it 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 positive or negative. Its effects are particularly difficult to evaluate when the relative risk is rather low (of the order of 1.1 to 3). When the relative risk is high, it is unlikely that a confounding effect strong enough to explain the difference will pass undetected. Confounding can be controlled either at the planning stage of a study (restriction, matching) or with data handling (stratification, modelling). In the latter instance this is possible only when the relevant data are available. For example, if alcohol use is to be controlled as a confounder, information on drinking habits is necessary, and the investigator must foresee this when planning the study.

In this context it is worth considering the ‘healthy worker effect’. This conceptually vague term describes the sum of the errors arising from comparing the mortality of an exposed cohort with that of the general population. The latter is heterogeneous and usually not completely free from the exposure under study. Furthermore, it rarely, if ever, represents the same social stratum as the exposed cohort. Hence the general population does not fulfill even the most elementary requirements of comparison validity, but such an invalid reference category is often used because it is practical and economical. An ad hoc reference cohort would double the costs of the investigation, and it is also difficult to find valid, suitable and available reference cohorts.

The healthy worker effect causes the standardised mortality ratio (SMR) of the exposed cohort to fall well below 100 if no life-shortening occupational hazard exists. In fact, figures of the order of 60–90 have often been reported. These figures are obvious under-estimates of the ‘true’ mortality. The main reason for the better-than-expected mortality 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 or handicapped during childhood and those ill when job seeking starts. All these groups have a higher mortality rate than the active population.

Table 1 shows some published SMRs from occupationally active cohorts. It is very difficult to interpret SMRs between, say, 90 and 110. Does no life-shortening hazard exist, or does the healthy worker effect mask such a hazard? In our Finnish foundry worker study[2] we were faced with this problem of interpretation. No certain conclusions could be drawn, but we intuitively favoured the explanation of the healthy worker effect masking a slightly increased overall mortality.

The healthy worker effect is not constant but varies according to a number of circumstances[8]. It is strongest in younger age groups, and declines with age until it is no longer significant in post-retirement groups. It is stronger for men than for women, stronger in higher social categories than in lower strata, and strongest at the start 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 during the first one or two years after cohort identification). Cancer is a typical example of such a disease. Coronary artery disease is an example of the opposite. A high risk of death due to this disease can be established in advance by early diagnosis or identification of its risk factors. Furthermore, coronary artery disease leads to self-selection because of symptoms or earlier warnings from physicians to avoid demanding jobs.

What still weakens the use of the general population is that no details of possible confounders are known. For instance, there is no information on smoking or drinking habits in national mortality statistics. Therefore the healthy worker effect poses a serious methodological problem whenever occupational groups are compared with a general population. The interpretation of such studies is difficult unless the effect being studied is very obvious. Because systematic errors of unknown strength are involved, tests of significance are meaningless. This is
so because they only measure random variation, giving no information at all about systematic errors. In this context they only mislead the unsophisticated reader.

Negative Studies

A very important and little considered problem is the evaluation of negative studies. Negative studies are at least as important as positive ones in occupational medicine because we are often interested in knowing a non-effect level for harmful exposures. Nevertheless, a clear distinction must be made between truly negative and ‘non-positive’ studies[9].

A true negative study must fulfil three criteria: (1) it must be large; (2) it must be sensitive; (3) it must have well-documented exposure data. Obviously, it must also be well designed. 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 to be true negative studies.

A small ‘negative’ study is uninformative in that it can only exclude the presence of very powerful hazards. 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. How sample size influences negative and positive conclusions can be illustrated by the following example. Suppose we are interested in deciding whether there are any black rabbits in a population of white ones. If, in a sample of 20 rabbits, even one is black, we know that black rabbits do exist. But if no black rabbits show up in our sample, can we state that no black rabbits exist? We would probably hesitate to do so, and require a larger sample. Would 100 be enough? Or 1,000, or 10,000? We can never be completely sure, but the absence of black rabbits in large samples tells us that these animals cannot be very common. We can say that the larger the sample, the more power it has to exclude the existence of black rabbits. We can use the same type of reasoning when interpreting negative epidemiological studies in terms of power. If a rather small study does not reveal any excess of, say, lung cancer in an exposed cohort, we can at most conclude that very strong carcinogenic effects of the exposure at issue are ruled out. If a large study shows the same negative result, we are more apt to reject the possible carcinogenicity of the exposure in question, but we can never be completely sure.

Insensitive studies, i.e., studies with a crude design or crude measuring methods, are uninformative. They can be compared with trying to identify beautiful girls through spectacles that are too strong, or with an insensitive laboratory test in clinical diagnosis, for example, the sole use of haemoglobin determination for diagnosing lead poisoning. Moreover, negative results can be referred only to the actual or lower levels of exposure; hence the availability of accurate exposure data is crucial. We know that this condition is seldom met. Good study design is necessary in all epidemiological studies. But more errors lead to false negativity than to false positivity; furthermore, negative bias is often more difficult to detect. In addition, investigations designed to study one or a few conditions cannot be interpreted as demonstrating ‘complete safety’. Some common errors leading to false negativity are as follows[9].

Inappropriate design often results in an inefficient study that fails to reveal an existing effect. A small sample size (and/or reference group) is the commonest error of design (but it is often unavoidable for reasons of availability). Such an error is usually 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 relative risks (of the magnitude of, say, 10–20–30) can be excluded. Thus, when the relative risk is moderate, a ‘negative’ result may emerge only because statistical significance has not been established. Hence, one should never automatically classify results into ‘positive’ and ‘negative’ on the basis of whether or not a statistical significance has been reached. If the \( P \)-value is based on rather small numbers, one case more or less can decide whether or not the study is ‘positive’ or ‘negative’. A few wrong diagnoses can then severely distort our judgement. Consider a 20-year follow-up of 500 men aged 45 to 59 years. Based on figures for the Finnish male population, the expected number of pancreatic cancers is 1.5. Table 2 shows the relative risks and the \( P \)-values for different numbers of observed cases.

| Observed No. | Relative risk | \( P \) (Poisson) |
|--------------|--------------|------------------|
| 3            | 2.0          | 0.191            |
| 4            | 2.7          | 0.062            |
| 5            | 3.3          | 0.019            |
| 6            | 4.0          | 0.004            |

By conventional reasoning, four observed cases would yield a ‘negative’ result, while five cases would result in a ‘positive’ one. It is quite obvious that one should not be fixed by the artificial cut-offs of statistical significance. Instead, one should consider the number of exposed cases, the magnitude of the relative risk, the confidence limits, and above all, how consistent the data are with previous studies and if they are biologically plausible.

Falsely negative results can also arise if the study centres on incorrect categories of exposed workers. Thus, if retired workers are left out of occupational cancer studies, crucial information will usually be lost and the ‘truth’, which may actually be positive, may appear as 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[10].

All too often, workers with too short an exposure time and too low an exposure intensity, or sometimes even non-exposed misclassified workers, are forced into the exposed cohort. Usually, the motive is to increase the cohort size. Nevertheless, this may be inappropriate
because the study becomes insensitive; i.e., an existing effect becomes diluted if the cohort is analysed 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.

Whenever diseases with long latency periods (e.g., occupational cancer) are studied, a falsely negative result emerges if the follow-up time is too short. There is no point in searching for occupational cancer before the subjects have been followed long enough to allow for the latency time. Of course no occupationally-induced excess can be found before the latency time has passed—it is biologically impossible. It should be noted that this error has nothing to do with lack of statistical power; hence it cannot be overcome by increasing the sample size. Fig. 1 usually gives a positive, not a negative bias. In case-referent studies, information errors form an even more serious problem, their main effect being to produce false positive results.

In contrast to systematic errors, whose effects may be positive or negative, random errors always mask existing effects. Such errors can arise from poorly standardised methods of measurement. Poor analytical precision, especially of the independent (exposure) variable, always decreases existing group differences and levels off the regression slope in exposure-response studies.

A choice of insensitive indicators of illness also causes falsely negative results when the purpose of the study is to demonstrate complete safety (as for standards of hygiene). Mortality is too crude an indicator of health risks associated with exposure to many conditions in today’s industry. Failure to find increased mortality has been misinterpreted as total absence of health hazards. This misinterpretation becomes especially dangerous when the conclusion is based on a comparison with the general population.

Errors of comparison are common in cohort studies. Finding valid reference groups is by no means easy. Furthermore, the reference group doubles the costs of the investigation. These two practical obstacles have often forced investigators to design studies that do not meet the criteria for validity of comparison. The errors that arise when the general population is used as a reference category have already been discussed. In case-referent studies, in which the referent group should be one whose condition is neither caused nor prevented by the exposure(s) at issue, insufficient knowledge of the aetiology of that condition can cause a comparison bias. The greater the number of exposures examined in the same study, the greater is the possibility that the reference condition will not fulfil this requirement for some exposures.

Finally, inappropriate interpretation of the results of a study can account for falsely negative (and also falsely positive) conclusions. Failure to appreciate that a negative finding is negative only in relation to actual or lower exposure levels is a common error in ‘negative’ epidemiological studies. Wrong thinking is also involved when a general negative statement of ‘safety’ is made on the basis of too few, too insensitive or even inappropriate measures of effect. For example, if no changes are seen in the blood picture or in hepatic or renal function at a certain intensity of exposure to lead, it does not follow that such a level is safe in all respects[11].

The most common reasons for falsely positive studies are comparison bias and information bias. In the same manner as one can select too ‘unhealthy’ a reference group, one can also select too ‘healthy’ a group. Falsely positive results then arise. Falsely positive findings may also arise through confounding. Large studies are the most dangerous ones in this respect, since even a small difference yields statistical significance. Then even a weak confounding factor may be decisive. In contrast, in smaller studies the difference must be so large that a statistical significance due to confounding is easily revealed. For instance, a relative risk of 1.5 is significant ($P<0.01$) in a study of 1,000 exposed and 1,000 referents

![Fig. 1. Schematic outline showing that a cancer study can be informative only when the latency time has passed.](image-url)
if the mortality is 10 per cent, whereas the same level of significance in a study of 50 exposed and 50 referents requires a relative risk of 11. Hence, in large studies, the magnitude of the relative risk is more important than the statistical significance of the difference, while in small studies the magnitude of the relative risk is very much influenced by chance and the statistical significance is therefore more important.

The most important cause of falsely positive results in case-referent studies is probably a special type of information bias, the so-called ‘recall’ bias. It arises whenever cases ‘remember’ better than referents.

Discussion

This review of some common reasons for fictitious results in epidemiological studies may give a rather pessimistic overall picture of epidemiological research. Undoubtedly, in many instances more skill would decrease errors, but very often various realities of life are effective obstacles to achieving validity: unavailability of proper populations, poor exposure data, frequent turnover of workers, mixed exposure patterns, and so on. In many instances the options are either to do nothing at all or to try to make the best of a defective situation. The first option should be given serious consideration if the indications are strong. But since completely perfect epidemiological studies probably are impossible to design, it is often necessary to initiate a study even if all validity aspects cannot be completely met. Then the key problem becomes one of interpreting the results correctly. 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 about the author’s level of understanding (sometimes also the editor’s), not about the problem at issue. Honest and thorough discussion of possible sources of error 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 reveal or exclude. Minor errors do not completely invalidate a study, especially if their direction and approximate magnitude can be estimated.

This recapitulation has tried to illustrate how difficult the distinction between fact and fiction can be in occupational epidemiology. The beginning of wisdom is the realisation of this difficulty. The rest develops from knowledge of the subject matter, epidemiological insight and the sceptical attitude of a critical mind.

Acknowledgement

I am indebted to Professor Olli S. Miettinen, MD, PhD, for having taught me to understand the concepts of validity during his annual Advanced Courses in Epidemiology in Occupational Health, given in Helsinki since 1972.

References

1. Campbell, D. (1975) Psychological Bulletin, 54, 297.
2. Koskela, R. S., Hernberg, S., Kärävä, R. et al. (1976) Scandinavian Journal of Work and Environmental Health, 2, suppl. 1, 73.
3. Lloyd, J. W. and Ciocco, A. (1969) Journal of Occupational Medicine, 11, 299.
4. McMichael, A. J., Spirtas, R. and Kupper, L. L. (1974) ibid., 16, 438.
5. Ahlman, K., Backman, A. L., Hannunkari, I. et al. (1975) Kansanterveyslaitoksen julkaisuja AL, 4, 128.
6. Ahlman, K., Backman, A. L., Koskela, R. S. et al. (1975) Työterveyslaitoksen tutkimuksia, 105, 64.
7. Ott, M. G. (1976) Journal of Occupational Medicine, 18, 171.
8. McMichael, A. J. (1976) ibid., 18, 165.
9. Hernberg, S. (1981) Scandinavian Journal of Work and Environmental Health, 7, suppl. 4, 121.
10. Enterline, P., De Couflé, P. and Henderson, V. (1972) Journal of Occupational Medicine, 14, 897.
11. Ramirez-Cervantes, B., Embee, J. W., Hine, C. H. et al. (1978) ibid., 20, 610.