

Critical reading of epidemiological papers

A guide

MARIA BLETTNER, CARSTEN HEUER, OLIVER RAZUM *

Many clinicians, medical practitioners and decision makers have no formal training in epidemiology but need to understand and sometimes evaluate results in epidemiologic studies. This paper attempts to give guidance to non-epidemiologists on how to read and evaluate the quality of epidemiologic studies and their results critically. Different methodological issues for evaluating whether the results of a study are causal or by bias, chance or confounding are given. This includes criteria for the choice of an appropriate study design followed by problems of the definition of the study population and the sample selection. We will also point to potential sources of bias in the data collection procedure and list some principles for statistical analysis. Finally, we include comments on how the results should be presented and issues which are related to public health and ethical questions. Although it is not usually possible to perform a perfect study and the correct approach to study design and analysis is often highly dependent on specific features of the population and the surroundings, our paper should help to distinguish between weak and strong investigations and papers.

Keywords: criticism, epidemiology, reading

Many clinicians, medical practitioners and decision makers have no formal training in epidemiology. Nevertheless, they are users of the results of epidemiologic studies. Their understanding of the results may have an impact on political or regulatory board decisions, on the setting of occupational limits for exposure to chemical or physical carcinogens and on therapeutic decisions. This paper attempts to give guidance to non-epidemiologists on how to read and evaluate the quality of epidemiologic studies and their results critically. The focus is on 'classical' epidemiology which is primarily concerned with the statistical relationship between disease agents and non-infectious risk factors. An example is investigation of the association between alcohol consumption and breast cancer or between oral contraceptive use and ovarian cancer. We will use the case control design as our main example to bring specific issues to the readers' attention and to help them discover the potential biases and pitfalls as well as strengths of epidemiologic studies.

So far, only a few journal articles have offered help in systematically assessing the quality of epidemiologic studies. They vary from short checklists¹ to detailed papers focusing on specific aspects such as data quality and communication of results. Most work has been done in

the field of occupational medicine.^{2–5} Subject-specific guidelines have been published for the evaluation of health care technologies⁶ and for drug research.⁷ There are also guidelines specifically dealing with the quality of statistical analysis.⁸

This paper covers criteria for the choice of an appropriate study design and highlights potential sources of bias in the selection of subjects, data collection and analysis. Although it is not usually possible to perform a 'perfect' study and the 'correct' approach to study design and analysis is often highly dependent on specific features of the population and the surroundings, our paper should help to distinguish between weak and strong investigations and papers. Our paper is not sufficient for a clinician to carry out his/her own study. We strongly recommend consulting with an epidemiologist or biostatistician already in the planning phase of a study.

CRITERIA FOR EVALUATION

The main question in an evaluation of the results of any epidemiologic study is judging whether the observed association between the exposure and the disease is causal or whether it is induced by a systematic error, a random error or by confounding. Systematic errors can be caused by the way subjects are selected for the study or in the way the information is obtained from them. Random error plays a major role if the study size is small or the individual variation is high. An observed association could also be induced due to differences between subjects in one other variable (confounder) which was not measured or not taken into account in the analysis. The following sections will consider different methodological issues for evaluating whether the results are causal or not.

* M. Blettner¹, C. Heuer², O. Razum³

¹ University of Bielefeld, School of Public Health, Bielefeld, Germany

² University of Heidelberg, Department of Medical Biometry, Heidelberg, Germany

³ University of Heidelberg, Department of Tropical Hygiene and Public Health, Heidelberg, Germany

Correspondence: Prof. Dr. Maria Blettner, Department of Epidemiology and Medical Statistics, School of Public Health, Post Box 100131, 33501 Bielefeld, Germany, tel. +49 521 106 3838, fax +49 521 106 2968, e-mail: Blettner@uni-bielefeld.de

Criteria for selection of a study design

The main types of investigation in analytical epidemiology are case control studies and cohort studies or derivatives thereof. Many factors influence the decision for a study design; reasons for the choice should be presented – particularly when non-standard designs are used. *Table 1* provides a list which helps in evaluating whether a design is appropriate. Deviations from these rules are sometimes needed but the reasons for other designs should be well explained.

Definition of the study population and sample selection

The distinction between the study population (target population, i.e. population for which results are wanted), the sample population (brut sample, i.e. population selected to be included in the study) and the population for which data are available and on which the analysis is based (net population, i.e. sample population minus subjects which are eligible but do not participate) is helpful in evaluating for which population the results are valid and to what degree they can be generalised.

Each quantitative study requires a careful decision about the required sample size. Power calculations should be based on requirements either for significance testing or for precision. The investigation should be large enough to find important associations but should not be larger than needed. A good report should include a formal power calculation and provide pragmatic reasons for the final decision on the study size. The power calculation should take non-response rates into account, as well as a possible subgroup analysis. For a case-control study with a 1:1 ratio of cases:controls and a significance level (probability of type I error) of 5%, the sample size required to avoid a type II error is shown in *table 2*.

Note that large numbers are needed when the prevalence of the exposure is rare and/or the relative risk is small. If subgroup analysis is performed, the numbers in *table 2* are needed for each subgroup. The value of extensively small studies is questionable and they should not be performed in the first place.

Most epidemiological studies are of observational nature and based on comparisons between exposed and non-exposed or between diseased and non-diseased subjects.

Table 1 Appropriate design for different investigations

Purpose of the investigation	Study type
Investigation of a rare disease such as cancer	Case-control study
Investigation of a rare exposure such as industrial chemicals	Cohort study in population where exposure is present
Investigation of multiple exposure such as combined effects of oral contraceptive and smoking	Case-control study
Investigation of multiple outcome such as mortality risk for several causes	Cohort study
Estimating incidence rates in exposed populations	Cohort study only
Investigation of covariates varying over time	Preferably in cohort study

Evidently, the correct choice of groups to be compared is crucial in order to achieve valid results.⁹ In case-control studies, the presence of exposure among cases should be compared with the presence of exposure among persons which are similar in all aspects except that they do not have the disease. In theory this would be best achieved by drawing a random sample from the population in which the disease occurs. As this is not always possible, an important issue is to judge whether the selection of the sample population may have introduced a bias.

To evaluate this, authors should provide a comprehensive description of the study population and the sample population in terms of age distribution, place of residence, time period, nationality and ethnicity. In addition, the total numbers of the brut and net sample populations and of important subgroups should be given. The reader should then check whether the sample population is a random sample of the study population, i.e. that all members of the study population had the same chance of being included in the sample. This condition is possibly violated in case-control studies if, for example,

- controls are selected from an incomplete list (telephone directories, lists, and drivers licence registries),
- control selection does not include institutions, e.g. homes for elderly persons,
- diseased subjects are excluded from control selection,
- healthy subjects are excluded from control selection ('hospital controls'), and
- interview times are not suitable for persons on shift work.

It is often not feasible to draw a random sample from the total eligible population. However, any compromise in sampling may introduce bias. The magnitude of the bias can be assessed by checking whether

- the same selection and inclusion criteria are used for cases and controls, e.g. for age, race, place of residence, social class, ownership of telephone, etc.,
- the percentage of persons included in the sample is the same for major subgroups such as for females and males and for different age groups (unless this was explicitly planned differently), and
- for hospital controls, a large variety of distinct diagnoses are included in the control group.

A good paper should contain a description of persons that are not included in the study.

Data collection, exposure assessment and information bias

The outcome variable, which is usually the diagnosis of the disease of interest or the cause of death, will be

Table 2 Required sample size for a case-control study with case-control ratio of 1:1 (alpha =0.05; power =80%)

RR	Prevalence of exposure among controls		
	50%	30%	10%
1.2	1,521	1,745	3,923
1.5	311	340	731
2.0	110	113	227
3.0	50	55	80

abstracted from medical records, pathological reports or from death certificates. In case-control studies exposure data for individuals are collected via questionnaires, interviews or by abstracting records. For each source of information, accuracy and validity of data poses different problems.

Measurement errors are a major source of variability in epidemiologic results. Errors are so-called 'non-differential' if they are the same for diseased and non-diseased persons in case-control studies. In this situation the estimates are biased towards no effect, the precision is decreased and the power is reduced.¹⁰ If misclassification is not the same for cases and controls, the bias may have any direction and even reverse the true effect. In cohort studies, differential misclassification can occur when death certificates are used for one group but health insurance data are used for the comparison group. In case-control studies, this problem can arise when cases are interviewed by a questionnaire and controls by telephone but also if they recall life events differently. The reader should pay major attention to this problem when evaluating the results.

The correct technique of exposure assessment is highly dependent on the risk factors of interest (e.g. diet, occupational exposure to chemicals, radiation and smoking habits). However, a few criteria are applicable to most measurement problems, including diagnosis and exposure assessment. Quality of data collection is often improved if the sampling procedures and the data collection methods have been tested in a pilot study and if additional investigations to validate the instruments, e.g. the questionnaire or the diagnosis procedures, are performed.

To assess whether misclassification may have occurred and to decide whether it is likely to be non-differential or not, one should examine whether the data collection is identical (or at least similar) for cases and controls. For example, the following.

- The same persons should interview cases and controls.
- The interview techniques (e.g. whether person-to-person or telephone interview) should be the same for everybody.
- Cases and controls should be interviewed at the same place (at home or in hospital).
- If possible, interviewers should not know the disease status of the interviewees.
- The data sources used for cases and controls should be the same, e.g. if blood group is needed, a differential error is introduced if controls are asked about their blood group and medical records are inspected for cases.

In addition, non-differential errors may appear if non-response rates differ between cases and controls or the percentage of missing values for some variables differ between cases and controls.

In some circumstances, the above-mentioned deficiencies cannot be avoided, but may be reduced. For example, if interviews with next of kin (proxy interviews) were performed, then proportions of these interviews should be reported, a validation study regarding the quality of the

proxy interviews should be performed and the analysis with and without proxies should be compared.

Statistical Analysis

The statistical analysis should summarise the crude data in a clear way; assess how likely it is that differences between groups (e.g. in exposure levels) are merely due to chance (random variability) and assess whether differences persist (or appear) when confounding variables are controlled for. Statistics cannot make up for samples which are too small or control bias in the study design.

The protocol for the statistical analysis should have been fixed prior to the analysis. A clear distinction should be made between formal hypothesis testing and a more explorative analysis. In most investigations, explorative modelling will be performed and the limitations of these results should be discussed.¹¹ The statistical analysis should always start with a comprehensive description of the data, including response rates, brut and net samples and deviation of the achieved from the intended sample population. In the second part, crude and stratified analyses should be performed and statistical models used to investigate the association between the risk factors and the disease. The following are some criteria for assessing whether an acceptable standard of statistical analysis is achieved.

- The overall strategy for analysing as intended should be outlined.
- A clear distinction between planned and *ad hoc* analysis should be given in the paper.
- The description of the analysis should be sufficient to understand precisely what has been done and to be reproducible. For example, statistical tests and estimation procedures should be described, variables used in the analysis should be listed, transformations of continuous variables (such as logarithm) should be explained, rules for categorisation of continuous variables should be presented, deleting of outliers should be elucidated and how missing values are dealt with should be mentioned.
- Is the analysis adjusted for relevant confounders? Although it is rather difficult to assess whether the confounder adjustment is appropriate, clarification of several points are needed: Which confounders are mentioned? How are confounders defined as being relevant? and Was adjusting done, e.g. by matching, multivariate modeling or stratification?
- If statistical models are used, the model building process should be transparent. It should be stated how variables were included or excluded from the multivariate model and results from different models should be described and compared. Sensitivity analyses could be used to investigate the model assumptions.
- Confidence intervals should be given for all estimations.
- If numbers are small (say below 30), exact tests should be used, if available.
- Potential bias and its influence on the results should be investigated, e.g. by using results from a validation study or by performing some sensitivity analysis.

*Author's critical assessment of the results**– Critical discussion of the findings*

The last part of a publication should include two major parts. First, the authors need to discuss the results, i.e. whether they confirm or contradict existing knowledge about the association between the disease and risk factors and what the study has added to existing knowledge. The authors should discuss the consistency between their findings and those of other studies, but also unexpected observations. Second, the authors should carefully discuss the limits of as well as biases and potential flaws in their study design, data management and analysis. Some important points are as follows.

- The findings should be compared with current knowledge.
- Not only significant but also non-significant results should be presented. (There is a danger of so-called 'publication bias', meaning that significant results are more likely to be presented. It is important for further research that negative studies and non-significant results are also reported in an adequate way.)
- The limitations of the study should be outlined and the influence of these potential limitations should be discussed (but not only with some standard phraseology).
- A clear distinction between 'no effect' and 'no observed effect' should be made (i.e. non-significant results due to a too small sample size).
- Suggestions for further research which arise from the study should be presented.

– Public health relevance

The quality of an epidemiologic study goes beyond technique.¹² Of equal importance is its actual benefit to the health of the population: do the findings contribute towards prevention or improved management of cases? An epidemiological study may have been carried out in a methodologically sound way and achieved statistically significant results. This does not imply its importance for public health or clinical care.¹³ The term 'significant' has a meaning in statistics which is different from its everyday meaning, namely 'relevant' or 'important'. Even a small increase in relative risk can be statistically 'significant' if the sample size is large enough; however, if the exposure is rare, the association may be of little (if any) practical importance. Nevertheless, such findings (e.g. a slightly elevated risk of childhood leukaemia near nuclear installations) often have enormous emotional impact on experts as well as on the public. The additional information required from a public health point of view is the fraction of cases which could be prevented if the exposure were removed. A prerequisite for this calculation is which the association under study is causal, which would have to be supported by additional evidence.^{14,15} Finally, to be of public health interest, the exposure under study has to be vulnerable to intervention.

- How prevalent is the exposure under study in the study population (or in the general population if the study population is a representative sample thereof)?

- How large is the attributable risk among the exposed or the population attributable fraction?
- Can the exposure under study be modified or removed?

– Ethical issues

Many ethical imperatives in epidemiological research are self-evident, e.g. minimising harm or comparing treatments in a clinical trial (or interventions in a community trial) only in the absence of evidence that one intervention is better than the other.^{16,17} An outright unethical study should not have been conducted in the first place. Still, a reader may wish to see whether i) the most important ethical aspects are discussed (consent, benefit to the population under study and potential for harm), ii) an ethics committee had overseen the study, and iii) whether the study was large enough to investigate the question of interest (small studies are unethical if it is known in advance that they would not be able to investigate the question of interest and are a waste of money and resources).

FINAL REMARKS

This paper is intended as a guide for non-epidemiologists evaluating publications of epidemiological studies. It focuses on the most important issues regarding scientific quality and is intended to help users distinguish between methodologically strong studies and particularly weak and biased studies which can be identified by carefully inspecting a published paper along the lines presented here. Decision makers may wish to exclude studies with many flaws or deficiencies from their considerations.

This guide is not sufficient for planning and conducting epidemiological research. For this purpose readers will need to refer to textbooks to obtain background information. We recommend the books by Rothman and Greenland¹⁸ and dos Santos-Silva¹⁹ for epidemiological methods, Kirkwood²⁰ for basic statistical methods and Clayton and Hills²¹ for further statistical modelling. Breslow and Day^{22,23} have provided a comprehensive description of statistical methods in cohort and case-control studies.

REFERENCES

- 1 Oxman AD. Checklists for review articles. *BMJ* 1994;309:648-51.
- 2 Collins JJ, Buncher CR, Halperin W. Managing the quality and conduct of epidemiologic studies. *J Occupat Med* 1991;33:1213-5.
- 3 Buncher CR, Collins JJ, Halperin W. Possible progress and unresolved conflicts resulting from guidelines on good epidemiologic practices. *J Occupat Med* 1991;33:1261-4.
- 4 Tomenson JA, Paddle GM. Better quality studies through review of protocols. *J Occupat Med* 1991;33:1240-3.
- 5 Cook RR. Overview of good epidemiologic practices. *J Occupat Med* 1991;33:1216-20.
- 6 Guyatt G, Drummond MF, Feeny DH, et al. Guidelines for the clinical and economic evaluation of health care technologies. *Soc Sci Med* 1986;22(4):393-408.
- 7 ISPE Notice. Guidelines for good epidemiology practices for drug, device, and vaccine research in the United States. International Society for Pharmacoepidemiology, 1996. Published in the internet: <http://www.pharmacoepi.org/policy/goodprac.htm>.

- 8 Thompson WD. Statistical criteria in the interpretation of epidemiologic data. *Am J Public Health* 1987;77:191-4.
- 9 Wacholder S. Design issues in Case control studies. *Stat Methods Med Res* 1995;4:293-309.
- 10 Armstrong BG. The effects of measurement errors on relative risk regressions. *Am J Epidemiol* 1990;132:1176-84.
- 11 Blettner M, Sauerbrei W. Influence of model-building strategies on the results of a case control study. *Stat Med* 1993;12:1325-38.
- 12 Susser M. Epidemiology today: a thought-tormented world. *Int J Epidemiol* 1989;18:481-8.
- 13 Feinstein AR. Scientific standards in epidemiologic studies of the menace of daily life. *Science* 1988;242:1257-63.
- 14 Hill AB. The environment and disease: association or causation? *Proc R Soc Med* 1965;58:295-300.
- 15 Glynn JR. A question of attribution. *Lancet* 1993;342:530-2.
- 16 CIOMS. International ethical guidelines for biomedical research involving human subjects. Geneva: Council for international organisation of medical sciences, 1993;1-63.
- 17 CIOMS. International guidelines for ethical review of epidemiological studies. Geneva: Council for international organisation of medical sciences, 1991;1-31.
- 18 Rothman KJ, Greenland S. *Modern epidemiology*. Philadelphia: Lippincott Raven Publishers, 1998.
- 19 dos Santos Silva I. *Cancer epidemiology: principles and methods*. IARC, Scientific Publication, 1999.
- 20 Kirkwood B. *Essentials of medical statistics*. Oxford: Blackwell Scientific Publication, 1988.
- 21 Clayton D, Hills M. *Statistical Methods in Epidemiology*. Oxford University Press, 1993.
- 22 Breslow NE, Day NE. *Statistical methods in cancer research, vol. II: the analysis of cohort studies*. IARC Scientific Publication, 1987.
- 23 Breslow NE, Day NE. *Statistical methods in cancer research, vol. I: the analysis of case control studies*. IARC Scientific Publication, 1980.

Received 29 October 1998, accepted 14 February 2000