Scand J Work Environ Health 1985;11(3):207-213

<https://doi.org/10.5271/sjweh.2233>

Issue date: Jun 1985

The case-referent study--some comments on its structure, merits and limitations.

by [Axelson O](#)

This article in PubMed: www.ncbi.nlm.nih.gov/pubmed/4035323

This work is licensed under a [Creative Commons Attribution 4.0 International License](https://creativecommons.org/licenses/by/4.0/).

The case-referent study — Some comments on its structure, merits and limitations

by Olav Axelsson, MD¹

AXELSSON O. The case-referent study — Some comments on its structure, merits and limitations. *Scand j work environ health* 11 (1985) 207–213. Case-referent (case-control) studies have become increasingly important in occupational health epidemiology. The concept of study base [Miettinen, *Scand J Work Environ Health* 8 (1982):suppl 1, 7–14] as referring to the health experience of the study population over time, and a clear recognition of open and closed populations, ie, with or without turnover, make the structure of the case-referent study clearer. The referents should represent the study base in terms of exposure and nonexposure and even with regard to other determinants of the disease under study. Case-referent studies refer to open populations unless nested in cohorts. Matching in cohort studies is straightforward and creates a symmetrical situation among the exposed and unexposed with regard to the matching factor, but this symmetry is not achieved in case-referent studies, in which matching tends to distort the representativeness of the referents with regard to the study base. This circumstance suggests that matching on determinants of the disease should be maintained in the analysis. However, since the referents are a sample of the base, there might be confounding in the data, which does not exist in the base and vice versa, ie, full control of confounding in case-referent studies is not achievable either through matching or through any other procedure, but a large sample of referents would more properly reflect the distribution of confounders in the base. Using hospital referents, one has to consider possible relations between the referent diseases and the exposure under study, and therefore evaluation of multiple exposures may even require different sets of hospital referents. The case-referent approach has a low cost and may permit the evaluation of several risk factors. It is convenient to apply induction-latency time requirements, but recall bias is commonly believed to be a serious flaw of the case-referent study even though the existence of this flaw has been poorly documented.

Key terms: cohort, confounding, matching, multiple exposures, nested case-referent, recall bias, study base, validity.

Epidemiologic studies have become an increasingly important tool for the identification of occupational health hazards. The traditional approach in work-related epidemiology has been through so-called cohort studies, ie, one group of workers with the exposure of interest has been studied with regard to health outcome in comparison to another group without the exposure. In studies of mortality or cancer morbidity, no specific reference group has usually been involved; instead, the observed numbers of cause-specific deaths or site-specific cancers have been compared to the expected numbers as derived from national or regional rates of mortality or cancer morbidity. However, also another study design has found its way into occupational health epidemiology, especially since the late 1970s, namely, the case-control or case-referent study. This type of investigation is somewhat complicated in structure and has been subject to some misunderstanding and controversy (9), particularly with regard to various aspects of validity. It seems justified therefore to try and elucidate the basic principles of the case-referent

approach, along with some of its merits and limitations.

The concept of study base and the case-referent approach

Every epidemiologic study has a base, which is the health experience of a specifically defined study population over a given period of time (11). The study population is either open, with a turnover of individuals, or closed, with only originally defined individuals, as in the traditional occupational cohort. The various measures of occurrence of disease might be connected with these two types of study bases. Hence, the incidence rate (or incidence density) has a person-year denominator and relates to an open base, the cumulative incidence is tied to a closed base, and the prevalence rate applies to both situations. Consequently, different types of studies might also be distinguished with regard to these various measures, as further elucidated elsewhere with regard to the case-referent design (3).

Irrespective of whether the base population is open or closed, it is possible and conceptually helpful to create a pictorial view of the base as shown in figure 1. The cases are symbolized with the plus signs, and the area might be taken as the number of person-years or

¹ Department of Occupational Medicine, University Hospital, Linköping, Sweden.

Reprint requests to: Prof O Axelsson, Department of Occupational Medicine, University Hospital, S-581 85 Linköping, Sweden.

E-			E+	
+	+	+	+	+
+	+	+	+	+
+	+	+	+	+
+	+	+	+	+

	E+	E-	RR
Cases	16	12	
Base (units)	4u	12u	
Rate	16/4u	12/12u	4

Figure 1. An illustration of the study base concept with the presence and absence of an exposure (E) causing disease. The base might be represented either by a sample (case-referent study) or by a census (census study, of which the cohort is a special case referring to a closed population); u denotes units of the base. The rate ratio, $RR = (16/4u)/(12/12u) = 4$.

Figure 2. Creating an unconfounded base from an originally confounded base (or source population) in a census study through matching on a determinant of the disease, F, with which the exposure is associated. Notice the harmonious situation obtained with regard to the distribution of the exposed and unexposed sectors of the population with regard to F.

individuals underlying the cases that appear during the study period. Furthermore, several domains of a base might be identified with regard to exposure categories; in the figure only an exposed (E+) and an unexposed (E-) sector are distinguished. Now, the disease rates may be obtained for the different domains of exposure and nonexposure with the knowledge of the number of cases, as well as their denominators (either person-years at observation or individuals, depending on the character of the base and the involved measure of occurrence of disease) — in this instance, taken as proportional to the area units or base units as indicated in the figure. Then, with the rates obtainable, rate ratios and rate differences can also be calculated.

A study design with all the individuals or the person-years involved known in terms of exposure category and with regard to other characteristics should be thought of as a census study (11). The

traditional occupational cohort is a special case of this design, namely, representing a closed base, which has an “imaginary” unexposed sector, for which the number of expected cases has to be calculated from national or regional rates.

However, a further consideration of the base at hand in figure 1 makes it obvious that a sort of relative rate, and therefore also the rate ratio, might be obtained through a sample of the base population. This sample represents the person-years at observation in an open base, or, alternatively, the individuals enrolled at the beginning of the study period, if the base is closed. This is now the case-referent approach, which can be applied either to an open base, as is commonly seen, or to a closed base, resulting in a so-called nested case-referent study or a case-referent study within a cohort.

The referents are usually strictly taken as noncases, which is natural and unproblematic for an open

population, since it "emits" its cases and fills up the exposed and unexposed sectors with new, healthy individuals, ie, those who become cases no longer provide person-years at observation and risk (in principle at least; notice, though, that cancer incidence rates are usually based on the living population, including the cases). When dealing with cumulative incidence in a closed study base, however, one has to apply the condition that a rare disease be studied, since an increased rate for a common disease among the exposed individuals causes the relation of exposed to unexposed individuals to be in considerably changed among those remaining healthy in comparison to the original distribution of exposed and unexposed subjects in the base. However, if the sampling permits case individuals to appear also among the referents, should they happen to be caught in the sampling procedure, then there is no need for a rare disease condition in the context of a closed base. To the extent that the incidence rate ratio is wanted, which is more unusual however, then the rare disease condition might be relaxed even in the situation of a so-called nested case-referent study.

Comparability, but not through matching

With the view in mind that the referents represent the study base, it is obvious that they should be representative of the base population over the study period. This is an important point, since it is commonly believed that cases and referents, should, above all, be as similar as possible. However, this is apparently not a primary requirement but rather a misconception in view of the structure of the case-referent study. It is important, though, that the two (or several) compared domains of the study base, ie, the exposed and the unexposed sectors of the population, be comparable with regard to the various background variables which may affect the disease rate. Furthermore, it is obvious that the sampling procedure as such can do nothing to achieve this comparability, should it not exist in the base itself. Instead it is important to define a valid base in terms of this comparability through various judgements in the design phase of the study. Hence, since a case-referent base is usually indirectly defined through the cases, one should consider various restrictions on their enrollment that would produce a base that is likely to have comparable sectors of exposed and unexposed individuals.

In census studies, where the investigator creates the study base by enrolling exposed and unexposed individuals in the study, matching is often suitable, since it leads to the selection of a reference population, ie, the unexposed sector of the base, that mirrors the distribution of the matching factors among the exposed. This procedure can be simply illustrated as in figure 2, showing a reduction of a primarily given study base through 1:1 matching so

as to obtain another, final and appropriate base for the study.

In the context of the case-referent approach, matching does not change the structure in the base itself but influences the distribution of the sample of referents with regard to the matching factor. Hence, matching in the sampling procedure requires that a referent be drawn from the same sector of the base from which the case comes, ie, the referents would tend to come more often from the sector where an extraneous risk factor or determinant for the disease operates and causes additional cases. To the extent that such an extraneous determinant tends to be positively or negatively associated with the exposure under study, so-called confounding would be present and might suggest a nonexistent but spurious effect of the exposure.

Irrespective of whether or not confounding is exerted by an extraneous determinant for the disease, the base is no longer adequately reflected in the data of a matched sample. Instead a distorted view is obtained, as shown in figure 3, ie, the case-densest sector becomes overrepresented among the referents.

It is also directly visible from the figure that no control of the existing incomparability or confounding in the base has been achieved. However, if matching has been applied, it should be maintained in the analyses, since the rate ratio is different in the unmatched design and in the dissolved, matched material. In this example the rate ratio is 4.1 and 1.9, respectively, both to be compared with the standardized mortality ratio of 3.6, as obtained from the stratified, unmatched material, as well as from the matched data with the matching maintained.

So, matching is useless in case-referent studies for the control of confounding, and, moreover, it eliminates the possibility to evaluate the effect exerted by the matching factor. Similarly, matching on a close correlate to the exposure under study might even more or less extinguish an existing effect of the exposure, a phenomenon known as overmatching. Even if matching might be thought of as contributing to the efficiency of a study through enrolling more subjects, for whom there is an extraneous and interfering determinant, it seems as if the drawbacks of matching in case-referent studies would predominate. The difficulties and costs involved in matching should namely be considered together with the analytical problems that arise through the distortion of the base as reflected in the data.

Because of the rather useless and merely disturbing effects of matching in case-referent studies, it should also be noted that matching is sometimes unintentional, eg, when the referents, for reasons of convenience and the structure of existing registers, are drawn from the same geographic area as the cases. Such unintentional matching is usually tantamount to random sampling, but exceptions exist, as in studies of the influence of geographically distributed

Figure 3. An illustration of the distorted reflection of the base as obtained through matching in a case-referent study. The stratum-specific rate ratios (RR) remain unchanged, as well as the standardized mortality ratio of 3.6, whereas the RR changes from 4.1 in the unmatched material to 1.9 in the dissolved, matched data. The effect of factor F is no longer visible in the total material but appears preventive in the sectors E+ and E- (the RR_F being 0.4 and 0.9, respectively).

Table 1. Case-referent data on a standardized mortality ratio (SMR) with the 95 % confidence interval in relation to number of referents. The study base consists of 900 and 90 000 without risk (ie, expected SMR = 1) and confounding but with varying strength of the extraneous risk factor.

Sample size of base (%)	Number of referents	Risk ratio of extraneous factor			
		1.0	2.0	4.0	8.0
8	72	0.47—1.53	0.43—1.57	0.30—1.70	0.09—1.91
0.08		0.62—1.38	0.54—1.45	0.39—1.60	0.17—1.82
16	144	0.62—1.38	0.67—1.33	0.65—1.35	0.57—1.43
0.16		0.53—1.46	0.54—1.46	0.52—1.47	0.50—1.49
32	288	0.76—1.24	0.74—1.26	0.71—1.29	0.67—1.33
0.32		0.74—1.26	0.71—1.29	0.67—1.33	0.62—1.37
64	576	0.89—1.11	0.89—1.11	0.89—1.11	0.88—1.12
0.64		0.84—1.16	0.81—1.18	0.77—1.22	0.73—1.26

factors, of which background radiation could be an example (6). In such situations a strictly applied random sampling would be preferable.

Confounding in the base versus confounding in the data

There are also other aspects of the representativeness of the referents that need to be dealt with. First there is the notion that, since the referents are only a sample of the base, random variation in the sampling may distort the reflection of the base in the case-referent data obtained so as to suggest confounding, even when not present in the base and vice versa. This phenomenon can be studied through computer simula-

tions, ie, letting a computer draw a number of sets of referents from a predefined base, each set corresponding to one case-referent study. Thereby the influence of the number of referents might be studied in situations with and without the presence of another risk factor for the disease, as displayed in table 1 and also shown elsewhere (5). As can be seen in the table, the standardized mortality ratio might deviate considerably, even in the absence of any true effect of the exposure, especially if the number of referents is small and a strong extraneous determinant is present, eg, when only 72 or 144 referents are enrolled as in the table. It seems advisable therefore to have at least some 200 referents, but preferably even more,

especially when relatively strong extraneous factors can be suspected to operate.

Hospital referents as representative of the base

Another aspect of the representation of the base is involved in the utilization of other disease entities as referents to reflect both the distribution of the exposure and other determinants of the disease. Hence, it is commonplace to draw both the cases and the referents from a hospital register instead of utilizing a population register as the source of referents. The hospital register as the only source of subjects for a study is often both adequate and desirable, since, first of all, the cases and referents would be likely to spring from the same socioeconomic and ethnic sector of the population. This probability is important for the comparability of the exposed and unexposed individuals in the base, since social, economic, or ethnic factors could have been involved in the recruitment of the patients to the particular hospital or to health care in general, ie, from the study point of view, primarily the recruitment of the cases. Secondly, referents who are also patients might be apt to respond more similarly to the cases in interviews and questionnaires than healthy individuals, drawn as referents from a general population register.

However, in the utilization of hospital referents, it should be realized that various disease entities themselves might be related to the exposure under study. For example, individuals suffering from cardiovascular disease would not be adequate referents for a study of lung cancer and smoking, since cardiovascular disease also seems to have causal relations to smoking. In other words too many referents with this type of disorder would be smokers in comparison to the smoking frequency in the base that delivered the cases.

It is noteworthy that the case-referent approach easily permits several exposures to be studied in the same series of cases and referents, eg, a set of lung cancer cases and referents might be scrutinized for smoking, as well as for exposure to asbestos, nickel, chromium, etc. One particular set of hospital referents does not always permit an unbiased evaluation of all types of exposures, however. For example, cardiovascular disorders would be questionable among the referents for the evaluation of the effect of smoking, but they would do for studying some of the occupational exposures. Furthermore, people with, say, a physical handicap already from childhood would hardly have been exposed to certain occupational hazards, eg, to asbestos in the construction industry.

These aspects of multiple exposure evaluations in case-referent data have been further elaborated elsewhere (4) and the problem of selecting representative disease entities as referents is somewhat more illustrated in figure 4. It is obvious also that, if the exposure is associated with a determinant for a referent disease entity, then a sort of confounding results on the referent side, a phenomenon that further complicates the utilization of hospital referents. A procedure similar to that of a hospital-based study would be to utilize a register of death records, from which both the cases and the referents are drawn, ie, a study approach closely related to the so-called proportional mortality study (1, 12).

Merits and limitations

One merit of the case-referent study is immediately obvious, namely, that a large study population can be represented through a rather limited number of referents, ie, the case-referent approach is usually comparatively cheap and therefore also effective in economic terms. A special aspect refers to the applica-

	E-			E+	E-	RR
	$\begin{array}{c} + \\ + \\ 0 \end{array}$	$\begin{array}{c} + \\ + \\ 0 \end{array}$	$\begin{array}{c} + \\ + \\ 0 \end{array}$	$\begin{array}{c} + \\ + \\ 0 \end{array}$		
Cases	16	12				
o-referents	8	12	2			
•-referents	8	24	4			
Popul ref's or base units	4u	12u	4			

Figure 4. The + -disease is taken as the case entity, and, as referents, are drawn either individuals with the o-disease or with the •-disease, the latter being representative of the base as unrelated to the exposure and therefore giving the same results as "population referents" (popul ref's). The rate ratio (RR) is given for the various alternatives of reference series.

Table 2. Exposure to phenoxy acids among cases of soft-tissue sarcomas by occupation^a (7).

	Agriculture/forestry			Other occupations		
	Phenoxy acids	Chlorophenols	Unexposed	Phenoxy acids	Chlorophenols	Unexposed
Cases	13	1	17	1	10	67
Referents	5	3	39	—	5	167
Rate ratio	6.4	—	1.1	—	5.0	(1.0)

^a The reference level, in parentheses, was obtained from the unexposed in other occupations.

tion of the so-called nested case-referent study. This approach would sometimes permit the omission of processing all information about the individuals in a large study population. Hence, primary information might be stored as raw material, say, unevaluated electrocardiograms, laboratory test results, or even frozen blood samples, to be processed only for those who turn out to be cases or finally are enrolled as referents.

Studying multiple exposures and allowing for induction-latency time

Another merit of the case-referent approach relates to the possibility of evaluating several potentially hazardous factors within the same set of cases and referents, given that the referents adequately reflect the exposure situation for the base population, as already discussed. It should be appreciated, though, that various risk factors or determinants might be associated with one another, exerting mutual confounding. Furthermore, even in the absence of confounding, other determinants than the potential one being evaluated would influence the baseline morbidity and thereby affect the rate ratio (unless there is multiplicative interaction). This is a fact that speaks in favor of census studies as directly providing disease rates and therefore also a less strongly affected measure of effect, namely, the rate difference.

Especially in cancer studies it is well motivated to account for an induction-latency period, usually from first exposure to appearance of disease. In brief, the period from first exposure to the induction of the disease might be thought of as the induction period, whereafter follows the latency period until the disorder is clinically diagnosed or until death, if dead cases are considered. The case-referent approach easily permits the ineffective last period of exposure before inception to be disregarded, say, the last five years. Also very remote exposure might be of less importance and can be omitted so that merely a "time window" is created with regard to what exposure is considered relevant. [Compare the paper of Flodin et al (8).] For the cases, it is natural and easy to anchor such an induction-latency allowance at the point of diagnosis or death. Since the referents are usually enrolled over a period of time concomitantly with the appearance of cases, the year or month of their acquisition would be the natural reference point, or, alternatively, in mortality studies the same would

apply for the time of death. In this way, the exposure might be time-weighted conveniently so as to improve the sensitivity of the study.

Some disadvantages and problems of classifying exposure

There are however not only benefits, but also disadvantages connected with the case-referent study. From a technical point of view there is first the limitation that only the rate ratio, but neither rates nor rate differences, can be directly obtained. To the extent that the absolute rate for the base population is known, either directly or indirectly, there are indirect methods, however, to derive the rates for the exposed and unexposed subpopulations and consequently also the rate difference (10). Another drawback has already been discussed, ie, the problem of coping with confounding as perhaps appearing in the data in spite of not being present in the base and vice versa.

Furthermore, case-referent studies are usually subject to much criticism because of an implied risk of differential assessment of exposure for the cases and the referents. Hence, a too frequent or too high exposure is usually believed to have been assigned to the cases, whereas the opposite is suspected for the referents, the reason being memory difficulties for the subjects and/or the researcher's too intensive interest in the (exposed) cases. There is little if any information in the literature, however, that really indicates that such bias operates. Instead, a comparison of questionnaire information about occupational exposure with the employer's files has resulted in almost total agreement (13). A greater problem with regard to validity in the assessment of exposure might actually be at hand in situations when there is already some toxicologic or epidemiologic evidence for an effect, but a case-referent study comes out "negative" or uninformative, ie, with a rate ratio in the vicinity of unity. Such a result might be due to just random misclassification of exposure, a viewpoint which is rarely discussed, however.

It might be possible to reveal whether a systematic misclassification of an occupational exposure has taken place in a study if another, more extensive exposure indicator is available, eg, all potentially exposed job categories (2). Then, if the rate ratio appears high among the exposed and low among the unexposed individuals within the potentially exposed jobs in comparison to those in jobs without potential ex-

posure, a misclassification is likely to be at hand. As an example of the application of this principle, table 2 is taken from a study of soft-tissue sarcomas (7) and exposure to chlorophenols and chlorinated phenoxy acids. There appears to be no indication of a misclassification, however, since the rate ratio for the unexposed workers in agriculture and forestry is close to (and even slightly above) unity. Had the increased risk among the exposed been due to misclassification, the unexposed within agriculture and forestry might have shown a rate ratio considerably below unity. A precondition for this principle to work effectively is however that the exposure frequency in the base population be relatively high, say, 10–15 %, and that relatively high rate ratios, ie, about 2–3 or more, are being dealt with.

Epilogue

The case-referent approach, as discussed in this presentation, is certainly a valuable and rather inexpensive tool for the identification of occupational health hazards, as well as other determinants of illness. There are also drawbacks with this method, however, and its complex nature tends to create difficulties both in application and in interpretation, but also means a challenge and a stimulation to the researcher. In view of the arguments presented, it seems as if commonly applied matching could be, and preferably even should be, abandoned in case-referent studies. The full control of confounding is unattainable in principle even for recognized confounders, ie, in the base of the study, but still reasonable control can be achieved in the analysis of the data, given a relatively large number of referents. The problem of misclassification of exposure in case-referent studies has perhaps been exaggerated, but

this question warrants further elucidation through comparisons of subjective and objective exposure information.

References

1. Axelson O. The case-referent (case-control) study in occupational health epidemiology. *Scand J Work Environ Health* 5 (1979) 91–99.
2. Axelson O. A note on observational bias in case-referent studies in occupational epidemiology. *Scand J Work Environ Health* 6 (1980) 80–82.
3. Axelson O. Elucidation of some epidemiologic principles. *Scand J Work Environ Health* 9 (1983) 231–240.
4. Axelson O, Flodin U, Hardell L. A comment on the reference series with regard to multiple exposure evaluations in a case-referent study. *Scand J Work Environ Health* 8 (1982): suppl 1, 15–19.
5. Axelson O, Johansson B, Axelson T. On the problem of controlling confounding in case-referent studies. *Ann Acad Med Singapore* 13 (1984): suppl, 308–311.
6. Edling C, Kling H, Axelson O. Radon in homes — A possible cause of lung cancer. *Scand J Work Environ Health* 10 (1984) 25–34.
7. Eriksson M, Hardell L, Berg NO, Möller T, Axelson O. Soft-tissue sarcoma and exposure to chemical substances; a case-referent study. *Br J Ind Med* 38 (1981) 27–33.
8. Flodin U, Andersson L, Anjou C-G, Palm U-B, Vikrot O, Axelson O. A case-referent study on acute myeloid leukemia, background radiation and exposure to solvents and other agents. *Scand J Work Environ Health* 7 (1981) 169–178.
9. Ibrahim MA, Spitzer WO. The case-control study: Consensus and controversy. Pergamon Press, Oxford 1979.
10. Miettinen OS. Estimability and estimation in case-referent studies. *Am J Epidemiol* 103 (1976) 226–235.
11. Miettinen OS. Design options in epidemiologic research — An update. *Scand J Work Environ Health* 8 (1982): suppl 1, 7–14.
12. Miettinen OS, Wang J-D. An alternative to the proportionate mortality ratio. *Am J Epidemiol* 114 (1981) 104–108.
13. Pershagen G, Axelson O. A validation of questionnaire information on occupational exposure and smoking. *Scand J Work Environ Health* 8 (1982) 24–28.