0031-3998/85/1908-0787$02.00/0
PEDIATRIC RESEARCH
Copyright © 1985 International Pediatric Research Foundation, Inc.

Vol. 19, No. 8, 1985
Printed in U.S.A.

Uses of the Case-Control and Cohort Epidemiological Approaches in Pediatric Practice and Research

ROBERT J. GOLDBERG, HARRIS PASTIDES, R. CURTIS ELLISON, ROBERT W. TUTHILL, AND THOMAS DEWITT

Departments of Medicine and Pediatrics, University of Massachusetts Medical School, Worcester; and Division of Public Health, School of Health Sciences, University of Massachusetts, Amherst, Massachusetts

ABSTRACT. Increasing reliance is being placed on the use of quantitative epidemiological methods in the conduct and evaluation of pediatric research. The basic design features of two common types of observational studies, the case-control study and the cohort study, are reviewed. Advantages and disadvantages of these two study designs are discussed with emphasis on aspects such as the selection of comparison groups, avoiding selection and recall bias, gathering exposure information, controlling for potentially confounding factors, and methods of analysis. Appreciation of the salient features of these study design approaches should aid the clinician/researcher in the conduct of research endeavors as well as in critically reviewing the medical literature. (*Pediatr Res* 19: 787–790, 1985)

The current trend in clinical research is toward increasing reliance on quantitative epidemiological methods. Indeed, the pediatric literature is replete with studies based on epidemiological principles. In order to evaluate the validity of the results of these studies and their applicability to one's own practice or research endeavors, it is an asset for the clinician/researcher to have a working knowledge of epidemiological study designs. Furthermore, from the epidemiological vantage it is increasingly recognized that clinical observations constitute a fertile area from which to draw testable hypotheses of risk factors and occurrence of disease.

This article will review the two observational (nonrandomized) study designs that are most often used in clinical epidemiological investigations: the case-control (retrospective) study and the cohort or prospective study. This methodological review will not deal with the design and conduct of randomized controlled clinical trials (1–3) since a considerable literature on this study design already exists and clinicians are in general familiar with this design approach.

It should be pointed out, however, that in terms of the validity of its findings, the randomized trial is usually the strongest study design in the epidemiological arsenal. Its advantages include random allocation to treatment, which tends to "balance out" the distribution of factors other than the one being studied which may affect outcome; this helps assure that any differences between groups receiving and not receiving a treatment can be attributed to the treatment itself, and not to other confounding factors. Randomized trials are also frequently "double-blinded," in that the patients and physicians do not know what type of treatment is being given. This helps prevent bias on the part of the patient, as well as the physician, which may result from knowing which treatment is being administered and believing that it has an effect.

Despite the advantages of randomized controlled trials, they are quite difficult and expensive to conduct. These trials require a considerable degree of work in persuading both physicians and patients to participate in them and also require a high degree of monitoring.

Thus, for most questions that arise in the context of clinical pediatric practice and research, the observational studies may be a more practical approach to utilize in exploring clinical impressions or explanations of observed associations. Appreciation of the purposes and the inherent strengths and weaknesses of each of these study designs should help the practitioner in critically reviewing the medical literature as well as in developing and initiating research protocols.

THE CASE-CONTROL (RETROSPECTIVE) STUDY

The basic design of the case-control study is the comparison, in respect to some factor(s) of interest, of two sampled groups, one of which has a particular disease or condition under study and the other which does not. The first step is to select patients with the disease or condition of interest and then classify them as to whether or not they were exposed to a putative etiological factor. An appropriate comparison group of patients without the disease under study is assembled at the same time as the selection of the case group. These subjects are similarly classified according to exposure to the suspected risk factor. Exposure may have been in the recent or distant past.

Pertinent data for both groups are usually obtained through personal or parental interviews, medical records, or other sources, each approach having its own inherent strengths and limitations. A 2 × 2 table may then be constructed for descriptive and analytic purposes (Table 1). The four cells in table 1 consist of: cases who were exposed to the factor of interest (a), cases who were not exposed (c), controls who were exposed (b), and controls who were not exposed (d). In terms of analysis, a comparison is made of the proportion of cases exposed to the suspect factor (a/a+c) and the proportion of controls exposed to the factor (b/b+d). If exposure is positively associated with the disease in question, there should be a greater proportion of cases than controls exposed to the factor under study (4, 5).

The odds ratio, or cross-products ratio (ad/bc), can also be calculated to measure the strength of the association between exposure and the clinical condition under study. This ratio will furnish an estimate of the risk of having the disease, given a particular exposure, as compared to the risk of having the disease without such exposure. An odds ratio of 1.0 implies no associa-

Received January 22, 1985; accepted March 13, 1985
Address for reprints Robert J. Goldberg, Ph.D., Department of Medicine, University of Massachusetts Medical School, Worcester, MA 01605.

Table 1. *Case-control study design*

	Study sample	
Risk factor	Cases (disease present)	Controls (disease absent)
Exposed	a	b
Not exposed	c	d
Totals	a + c	b + d
Proportions exposed	$\frac{a}{a+c}$	$\frac{b}{b+d}$
Odds ratio	$\frac{ad}{bc}$	

Table 2. *Frequency distribution of aspirin use in patients with Reye's syndrome and in controls*

Aspirin usage during prodromal illness	Reye's syndrome patients ("Cases") (n = 100)	Controls (n = 100)
Used	90	60
Not used	10	40
Proportion exposed	90%	60%
Example of odds ratio calculation:		
Odds ratio for aspirin Users *vs* nonusers	$\frac{(90)(40)}{(60)(10)} = 6.0$	

tion between the factor of interest and the disease in question. An odds ratio less than 1.0 suggests a negative or "protective" association, while an odds ratio greater than 1.0 implies a positive association between the risk factor and disease. To provide a measure of the degree of confidence one can attribute to the observed odds ratio, 95% confidence intervals are usually calculated. These provide an interval which quantitatively depicts the likelihood that the odds ratio is a reliable estimate of the true risk of disease.

As an illustrative example of the case-control approach, the association between aspirin intake and Reye's syndrome will be examined. In this hypothetical example which is based on actual data (6–8), the comparison groups under study consist of young children with recently diagnosed Reye's syndrome (case group) and young children without Reye's syndrome (control group), with the exposure factor being use of medications containing aspirin. As seen in the analysis of this example (Table 2), medications with aspirin were used significantly more frequently by cases (90%) than by controls (60%) during their prodromal illness, with the calculated odds ratio of 6.0 implying a strong positive association between Reye's syndrome and aspirin intake.

SELECTING THE CASE GROUP

Selection criteria for the case group in a case-control study are usually suggested by the question under study and, in general, should include predetermined diagnostic criteria, consideration as to the severity of disease, and consideration of the source of the case population (*e.g.* hospitals, clinics, private offices). Where possible only newly dignosed, or incident, cases should be included. There are two important reasons for preferring incident cases to long-standing or prevalent cases in a case-control study. One is that prevalent cases may be different from all cases with the disease merely by virtue of the fact that these patients still have the disease, but have neither been cured of it nor died as a consequence of it. Another reason for including only incident cases, particularly in the area of pediatric research, is that the passage of time can result in selective or biased recall of past events by either the child or parent. The use of of newly diagnosed cases tends to minimize the time lag between exposure and disease and helps avoid such recall bias, which could alter the etiological importance of the putative risk factor in either a positive or a negative direction.

The source of the case population strongly influences the extent to which the results cn be extrapolated to a population beyond that of the study group. Pediatric patients seen in a hospital setting may be quite different regarding factors such as disease severity, socioeconomic status, and other characteristics from those seen in physicians's offices or in neighborhood clinics; also children seen in one hospital may differ in important respects from those seen in another. Therefore, if the case and control populations are drawn from a single hospital or clinic, one has to consider the characteristics of the "captive" population that utilizes this health care setting. The more representative the study population is of the general population, the more likely it is that the results can be extrapolated to other children or adolescents with the condition in question.

SELECTING THE CONTROL GROUP

The selection of the control group is one of the most important and difficult aspects in designing a case-control study. The ideal control group would consist of children or adolescents who are representative of all children or adolescents without the disease in the community with respect to the exposure factor under study. General population controls, however, tend to be difficult to identify and are more likely to refuse study participation. Frequently therefore, case-control studies use two control groups, one consisting of patients hospitalized with conditions other than and unrelated to the disease in question in the case group, the second control group consisting of persons residing in the same neighborhood as the patients but without the disease under study. Each of these two comparison groups has logistic and methodological strengths and constraints that need to be carefully considered.

The advantages of using hospital controls include ease of access, similarity of the setting in which patients and/or their proxy respondents (*e.g.* parents, friends) are examined and questioned, and increased likelihood of participation. The major disadvantage is that hospitalized controls may not be representative of the population at large without the disease in question by nature of the fact that they are hospitalized for some conditon; furthermore, their condition may unknowingly be related to the etiological factor under study. Careful consideration is necessary regarding the diagnoses to be included or excluded from consideration as a hospital control; the disease(s) in this group must not be related etiologically to that of the case group. Due to these and other concerns, it may be desirable to have a neighborhood control group as an additional measure of the exposure factor in the community. However, using neighborhood controls involves some type of survey (door-to-door interviews or mail or telephone questionnaires) which makes obtaining exposure data logistically difficult and relatively more expensive.

GATHERING EXPOSURE INFORMATION

In gathering data on the exposure factor, it is important to have some means of validating exposure. In the previously described study of Reye's syndrome and administration of aspirin, for example, parents of the children with the disease, or the children themselves, could have been asked to supply the specific trade name of medication used, or even to furnish any unused samples of medication if they were still available. Confirmation of exposure by review of physician's records or prescriptions may be necessary, but can be difficult and expensive; for over-the-counter medications this is not possible. Furthermore, to avoid bias in collecting exposure information, persons conducting the interviews with study subjects should not know if the individual being interviewed is a "case" or a "control" and, whenever possible, should also be "blinded" to the major hypothesis under study.

In addition to the principal exposure factor under study, other factors that may explain its association with the disease in question must be defined and considered. These factors are called confounding factors. For instance, in examining the relationship between Reye's syndrome and aspirin use, the presence of other factors possibly related to the development of Reye's syndrome, such as a viral infection prompting use of aspirin or the simultaneous ingestion of other medications, can be considered as potentially confounding factors, particularly if the distribution of these characteristics differs between case and control groups. To control for the effect of these additional risk factors, statistical adjustments of the data need to be made. Stratified subgroup analyses (*e.g.* looking separately at cases and controls taking and not taking other medications) or multivariate analyses are some of the techniques frequently used for this purpose.

ADVANTAGES AND DISADVANTAGES

The case-control study is the most appropriate and feasible design to use when the disease or condition in question occurs infrequently. Among its advantages are that it is a relatively inexpensive study to conduct, the number of subjects needed is relatively small, and the study can be conducted reasonably quickly. This design is ideally suited for the initial testing of hypotheses and for exploratory ventures suggested by clinical observations.

The case-control study, however, has a number of disadvantages. It is rarely possible to obtain a truly representative control group, and the information obtained about past events or exposures may be limited. In addition, incidence rates of disease cannot usually be calculated from this study design since one is not monitoring an entire population for development of disease. The case-control study is typically the initial design approach used in examining a potential association between a suspect risk factor and disease. If the results of this study suggest some kind of relationship, the investigator may then want to proceed to a cohort study, which is also known as a prospective or longitudinal study.

THE COHORT (PROSPECTIVE) STUDY

The basic design of the cohort study is illustrated in Table 3. The investigator typically selects a sample of healthy individuals according to whether or not they were exposed or not exposed to some factor of interest or of diseased individuals according to whether or not they were nonrandomly treated or not treated with some therapeutic agent. In the usual cohort study, these comparison groups are followed over time to see whether those exposed (or treated) are more likely or less likely to develop the selected endpoint(s) than those not exposed (or not treated). Similarly to the calculation of the odds ratio in the case-control study, the relative risk is calculated to obtain an estimate of the strength of the association between exposure (or treatment) and the disease in question. The method of calculating the relative risk is shown in Table 3.

Table 3. *Cohort study design*

	Follow-up			
	Develop disease	Do not develop disease	Totals	Incidence rates of disease
Exposed population	a	b	a + b	$\frac{a}{a + b}$
Nonexposed population	c	d	c + d	$\frac{c}{c + d}$

$$\text{Relative risk} = \frac{\text{incidence of disease among exposed}}{\text{incidence of disease among nonexposed}} = \frac{\dfrac{a}{a + b}}{\dfrac{c}{c + d}}$$

The cohort study design is particularly attractive to the clinical researcher because it enables the investigator to calculate incidence rates of disease among exposed and nonexposed (or treated and untreated) comparison groups and, hence, to measure directly the risk of disease or other health-related outcome. In this design, either an entire cohort can be followed over time or only subsets within the cohort (*e.g.* groups of adolescents who smoke cigarettes and groups of those who do not smoke). Unfortunately, many illnesses of clinical and public health concern have long latency periods betwen exposure and disease and the investigator usually cannot wait years before obtaining results from a concurrent cohort study.

As a means of condensing the years of follow-up in a cohort study, a nonconcurrent, or historical cohort approach is frequently used. The features of the standard prospective design format are retained, but the starting point of the study is set back in time by selection of the study population from past medical records or other sources. The subjects are then traced from that point up to the present or some recent date. As with all cohort studies, individuals assessing outcome should be unaware of exposure status and follow-up should be carried out equally in the exposed and nonexposed groups.

Illustrative of the cohort design is a study examining the relationship between breast- and bottle-feeding and the development of respiratory illness during the first year of life (9) (Table 4). The sampling in this investigation was restricted to a residential suburb of London which had 2365 livebirths occurring in the community's hospital during a defined 2-yr period. Of these, 2205 families were available for participation in the study. Study interviewers visited each family within 14 days of birth to determine the exposure of interest, breast- or bottle-feeding, as well as to collect information on a variety of additional factors (including birth weight, health at birth, housing conditions, social class, and parental smoking and respiratory symptoms) that could influence the outcome under study, namely, subsequent development of respiratory illnesses. Three distinct infant cohorts were created as a result of these family interviews: breast-fed only; breast- plus bottle-fed; and bottle-fed only. At a follow-up visit at the time of the infant's first birthday, detailed information was collected on the child's history of respiratory illness during the past 12 months, with a primary focus on the occurrence of bronchitis and pneumonia. Validation of this information for a sample of infants was conducted by reviewing general practitioner's records.

Incidence rates of selected respiratory conditions according to initial feeding status were then estimated. As seen in Table 4, a significant trend in bronchitis or pneumonia occurrence was observed according to feeding status: 8.1% among breast-fed only, 12.8% among breast- plus bottle-fed, and 14.8% among bottle-fed only infants. Relative risks and the calculations thereof are shown, indicating excess risks of respiratory illness of approximately 60% (relative risk = 1.6) and 80% (relative risk =

Table 4. *Incidence rates of bronchitis or pneumonia in the 1st yr of life, by feeding pattern*

Feeding pattern	Incidence rates (per 100 infants)
Breast-fed only (*n* = 958)	8.1
Breast-plus bottle-fed (*n* = 274)	12.8
Bottle-fed only (*n* = 842)	14.8
Total (*n* = 2074)	11.5

Relative risk (breast plus bottle-fed/breast-fed only) = 12.8/8.1 = 1.6
Relative risk (bottle-fed only/breast-fed only) = 14.8/8.1 = 1.8

1.8) in breast- plus bottle-fed infants, and bottle-fed only infants, respectively, in relation to breast-fed only infants.

The advantages of the cohort approach as used in this study are 2-fold. The first is the accurate and unbiased collection of infant feeding information. By inquiring at two distinct points in time about this (*e.g.* early in life and 1-yr later), data should be more precise than in a case-control design where a mother would be asked, retrospectively, about feeding patterns during the infant's 1st yr of life. The second advantage is in the ability to compute incidence rates. The case-control approach would not have allowed an estimate of the frequency of occurrence of respiratory conditions during infancy in this defined population.

As with retrospective studies, the characteristics of the study population will limit the validity of any generalizations that can be drawn. One must therefore consider the sociodemographic and clinical characteristics of the population studied, the system of referral to the health care facility from which the study population has been assembled, and other relevant characteristics.

The major problem in the cohort study is the loss of individuals to follow-up. Losses clearly must be kept to a minimum; they not only reduce the number of subjects available for analysis, but the reasons why individuals are lost to follow-up (*e.g.* illness, death) may be related to the outcome under study and thus add a potential source of bias. As a means of assessing this bias, those lost to follow-up should be compared in terms of baseline sociodemographic or clinical characteristics to those remaining under follow-up to determine if there were initially any systematic differences between these two groups. If no consistent differences were noted, the investigator would be somewhat reassured that the results were not biased. However, it may be advantageous to use special approaches (*e.g.* review of death certificates) to try to collect some outcome data on a random sample of those lost to follow-up, and then to compare results in these patients with those of patients under follow-up. However, the underlying principle should always be to keep the number lost to follow-up to a minimum.

ADVANTAGES AND DISADVANTAGES

Compared with the case-control study the cohort study has a number of distinct advantages. This design allows the estimation of incidence rates in exposed and nonexposed individuals; it introduces considerably less bias in the assessment of the exposure factor, as comparison groups are classified according to this factor prior to endpoint ascertainment; it provides meaningful results when exposure is rare; it provides better data on the time relationship betwen exposure and onset of disease, which might be clouded in a case-control study; and it allows for the assessment of multiple outcomes. The latter is illustrated by the Collaborative Perinatal Study (10), in which the original purpose was to identify perinatal risk factors for neurological defects among children but which also permitted examination of the relation of certain risk factors to the development of congenital heart disease (11, 12).

The cohort study also has a number of disadvantages. It is not a practical design for studying diseases that are of rare occurrence (*e.g.* aplastic anemia), as it would require following too large a population to detect enough children with the disease. Prospective studies also are inexpensive and involve a variety of logistical concerns related to maintaining contact with the cohort and assessing for the occurrence of the disease of concern.

In summary, both the case-control and cohort study designs have inherent strengths and weaknesses. The practitioner should keep in mind the salient features of these two observational approaches whenever contemplating the conduct of a research endeavor as well as when critically reviewing the medical literature. Particular attention should be directed to the biases that may arise in carrying out these observational studies. These may result from inadequate control for potentially confounding variables, selection bias, and bias in the assessment of outcome in the comparison groups under study. The potential for such biases should be carefully considered before undertaking a clinical/epidemiological study and steps taken to avoid them as much as possible.

REFERENCES

1. Spodick DH 1982 The randomized controlled clinical trial. Scientific and ethical bases. Am J Med 73:420–425
2. Sackett DL, Gent M 1979 Controversy in counting and attributing events in clinical trials. N Engl J Med 301:1410–1412
3. DerSimonian R, Charette LJ, McPeek B, Mosteller F 1982 Reporting on methods in clinical trials. N Engl J Med 306:1332–1337
4. Hayden GF, Kramer MS, Horwitz RI 1982 The case-control study. A practical review for the clinician. JAMA 247:326–331
5. Sartwell PE 1974 Retrospective studies. A review for the clinician. Ann Intern Med 81:381–386
6. Halpin TJ, Holtzhauer FJ, Campbell RJ, Hall LJ, Correa-Villasenor A, Lanese R, Rice J, Hurwitz ES 1982 Reye's syndrome and medication use. JAMA 248:687–691
7. Starko KM, Ray CG, Dominquez LB, Stromberg WL, Woodall DF 1980 Reye's syndrome and salicylate use. Pediatrics 66:859–864
8. Waldman RJ, Hall WN, McGee H, Van Amburg V 1982 Aspirin as a risk factor in Reye's syndrome. JAMA 247:3089–3094
9. Watkins CJ, Leeder SR, Corkhill RT 1979 The relationship between breast and bottle feeding and respiratory illness in the first year of life. J Epidemiol Community Health 33:180–182
10. Niswander KR, Gordon M 1972 The women and their pregnancies. WB Saunders Co, Philadelphia
11. Mitchell SC, Korones SB, Berendes HW 1971 Congenital heart disease in 56,109 births: incidence and natural history. Circulation 43:323–332
12. Mitchell SC, Sellman AH, Westphal MC, Park J 1971 Etiologic correlates in a study of congenital heart disease in 56,109 births. Am J Cardiol 28:653–657