The retrospective cohort study during an epidemic field investigation

Victor M. Cardenas, Editor-in-Chief, AJFE

Abstract

Retrospective cohort studies in the context of epidemic field investigations do not receive sufficient attention in most epidemiology texts despite being the most commonly employed. Moreover, it is erroneously argued that the case-control study should be used more frequently. I review the indications in which the retrospective cohort study is most appropriate: identification of an at-risk population that can be studied by census or sampling and that experiences a substantial attack rate, say 10%. I also address other methodological issues regarding measurements of exposure and disease occurrence, the sample size required, and the analysis of data from a retrospective cohort study.

Keywords: epidemics; disease outbreaks; cohort studies; epidemiologic methods

Introduction

The retrospective cohort study (sometimes referred as historic cohort study) is the most commonly epidemiologic study design used during the investigation of an epidemic outbreak. This fact does not receive due attention in epidemiology textbooks where only planned epidemiological studies are conceived as epidemiological studies [1], and even a preponderant or special place is wrongly proposed for case-control studies in the investigation of epidemic outbreaks [2]. In common scenarios encountered in epidemic field investigation a population with a risk of at least 10% has already occurred, and all or part of the population at-risk can be enumerated to assess the occurrence of disease according to the exposures postulated to be associated with such occurrence. These are the necessary circumstances when one should consider carrying out a retrospective cohort study, as we present below in more detail.

Earlier in this section of the Trainer's Corner it was argued that in most instances, the investigation of an outbreak is part of the public health response, and a full epidemiological investigation is not warranted [3]. However, there are times when an analytic epidemiologic investigation is warranted to advance knowledge and apply it immediately, for as Schaffner and La Force wrote: "Natural outbreak experiments only later appear self-evident because an epidemiologist seized the circumstance in the field to pose a question and structure an investigation in order to learn something new" [4].

To describe disease occurrence during an epidemic, if data on the population at-risk is available, such as population estimates by age and sex or other characteristics, specific rates by age group, sex, place of residence, or other characteristics should be calculated to characterize the epidemic and to postulate hypotheses about agent, host, and environmental factors that may be associated with disease occurrence. This part of the research is descriptive, although it may have elements of an ecological study by testing hypotheses on the existence of differences, say by age, sex, place of residence or

other characteristics. For the description of the occurrence of the disease in this descriptive phase, the proportions of affected persons according to such characteristics are used in the form of attack rates (AR):

$$\text{Attack rate} = \frac{new\ cases}{Population\ at-risk}\ x\ k$$

where *k* is a constant usually 100 or any number that allows the rate to be at least one whole digit. During an epidemic of short duration, the follow-up period of time is understood to be the period that the epidemic lasts, i.e., the epidemic period. The AR is a specific instance of a more general frequency measure called cumulative incidence or risk, which has an implicitly or explicitly associated follow-up period of time. Assuming that the disease was not present before the onset of the epidemic period, i.e., that all cases considered for the calculation are newly onset cases, such a ratio is actually the risk of disease during the epidemic period. One should keep in mind that, to obtain the population at risk, one should exclude those who have already had the disease or who are not susceptible to it (i.e., who have antibodies). Attack rates are preferably compared on an age-adjusted basis and such a comparison, even if it is a descriptive phase of the research, is consistent with an analytical study of groups of people, even if the case data represent a series of cases and thus individuals. The crude or adjusted ARs are compared according to these characteristics using ratios, i.e., attack rate ratio, hence risk ratio (RR) or differences in attack rates, hence risk differences (RD).

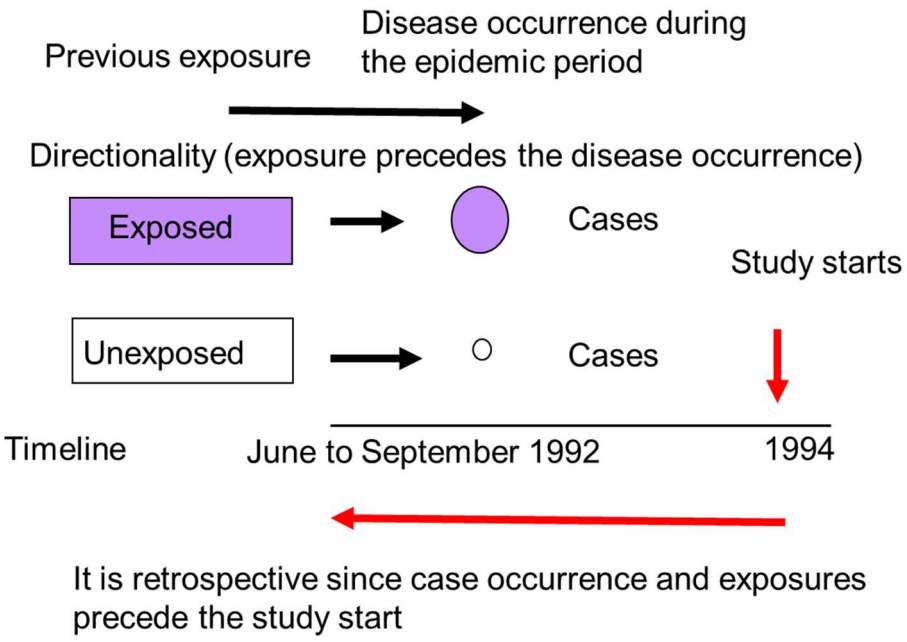
Often the descriptive phase of an outbreak investigation and even the group-based, i.e., ecological studies, do not answer the questions one has about the preventable risk factors of the agent, host, environment including those related to social determinants and those related to health services, which influence the occurrence of the disease or condition of interest. For this reason, individual-based studies such as cohort or case-control studies are necessary. Keep in mind that some determinants or

determinants of the health-disease process cannot be fully evaluated with data at the individual level, but at the ecological level [5].

Overall design

Retrospective cohort studies collect information on exposures once the exposure has occurred, and also once all or most of the cases have already occurred (Figure 1). That is, what makes these cohort studies "retrospective" is that both the exposure and the cases have already occurred at the time the study starts.

Figure 1. Features of the retrospective cohort study



Outbreak of human immunodeficiency virus infection in a hemodialysis unit

For general design illustration purposes and of other topics to be reviewed later, we now return to the example visited in this series of the HIV epidemic in a dialysis unit that occurred in a university hospital in Colombia in 1992 [6]. Field evaluation of the data available in the records of the dialysis unit led to identify that the risk was concentrated among those who received repeated dialysis (Table 1).

Table 1. Dialysis characteristics, availability of serum for antibody testing to HIV, results of testing, January 1992 to December 1993 (epidemic period), dialysis unit in Colombia

Dialysis	Number of patients			
	No. (serum available)	Serologic status		
		Negatives	Converters	Seropositives Probable/definitive
CAPD*	4 (4)	4	0	0
Acute dialysis	9 (8)	8	0	0
Chronic dialysis	29 (23)	10	9	1/3
Total	42 (35)	22	9	1/3

*Chronic ambulatory peritoneal dialysis
Source: [Reference 6]

The study was concentrated among 19 of these 23 subjects. For reasons unrelated to the study, the head of the dialysis unit was himself the dean of the medical school and was conducting a research protocol whereby he obtained sera each time patients attended their hemodialysis sessions. This first led to the discovery of the outbreak because a microbiologist serendipitously used a serum from these patients as a control for an HIV test, and when it was positive, she ended up not only notifying the outbreak to the National Institute of Health, but also sending all the samples from the serum bank.

Patients on dialysis with or without HIV seroconversion during the epidemic period were similar in number of blood transfusions received, exposed to surgical procedures, renal transplantation, dental procedures, or endoscopy (Table 2). In contrast, the rate of HIV seroconversion was higher in patients who received dialysis during the same 4 months in which the patient 22 from the dialysis unit, a patient who was positive on the first serum sample taken the first time he attended the unit (May to October 1992) (90% vs. 0%; P<0-0001) or the months in which the dialysis unit reprocessed dialysis filters (January 1992 to August 1993) (60% vs. 0%, P=0.05).

Table 2. Comparison of the risk of HIV seroconversion among patients of dialysis unit with and without exposure to various risk factors

Exposure	Yes/No	Seroconversión	Risk Ratio (95% CI)	Risk Difference (95% CI)	*P*- value
Received hemodialysis					
With patient #22 (May - October 1992)	Yes	9/10	∞ (NC*, ∞)	90% (60.0, 100.0)	0.0001
	No	0/9			
Reprocessed filters (Jan 1992 - August 1993)	Yes	9/15	∞ (NC, ∞)	60% (35.2%, 84.8%)	0.05
	No	0/4			
>2 blood transfusions	Yes	4/6	1.7 (0.7, 4.2)	28.2% (-17.9%, 74.3%)	0.3
	No	5/13			
Procedures					
Renal transplant	Yes	2/3	1.5 (0.6, 4.0)	22.9% (-35.7%, 81.5%)	0.5
	No	7/16			
Dental	Yes	4/7	1.4 (0.5, 3.5)	15.5% (-30.6%, 61.5%)	0.6
	No	5/12			
Endoscopy	Yes	3/6	1.4 (0.4, 2.9)	12.0% (-44.5%, 34.8%)	0.9
	No	6/13			

*NC: not calculable
Source: [Reference 6]

If we say that an exposure (variable X) occurs before, i.e., that it precedes a health effect, such as the disease or condition under study (variable Y), we say that causal directionality is well established. When we do a retrospective cohort study, the time at which both exposure (X) and disease (Y) occur is investigated. Like all cohort studies, the retrospective cohort study maintains a clear causal directionality, that is, the timing of exposures preceding the occurrence of the disease is clearly established and measures of occurrence such as risk or attack rate can be measured directly. This is the case of the study in the example of HIV in the dialysis unit.

The Dutch Famine Cohort Study

The characteristics of a planned retrospective cohort study are briefly illustrated by the famous 1944 Dutch famine study at the end of World War II. Epidemiologists Mervin Susser and Zena Stein led this famous study [7], which they envisioned as "analogous to animal experiments in which nutritional deprivation lowers fetal growth, including the head, and perhaps causes brain cell depletion." Dr. Susser had served as a pilot with the Allied forces and both Dr. Susser and Stein were persecuted in their native South Africa for their interest in the health of the black majority and their political support of the struggle against Apartheid. Both emigrated in the 1960s to the USA. Knowing that there were records of births in hospitals in both rationing (of approximately 1,500 calories per day) affected and unaffected areas of Holland during the famine (October 1944 to May 1945), they set out to examine what effects such an experience had on anthropometric indicators at birth. They then set out and succeeded in matching the birth records with the military service records of 400,000 Dutchmen, and thus established the effects of in utero starvation on neurodevelopment. What makes this cohort study retrospective, or historical, is that both the famine exposure and the effects occurred twenty years before the study began. However, there is no doubt that the famine temporally preceded the occurrence of child developmental deficits. The researchers identified that birth weight loss was most significant when experiencing famine in the third trimester of pregnancy. Surprisingly, their studies established that there was no significant effect of in utero exposure to starvation on cognitive development at age 18. Other studies nested within those of the authors have identified an increased risk of chronic disease among those who experienced famine in the fetal stage of life.

This retrospective or historical cohort study design in a shortened form is applied in the study of outbreaks of food poisoning occur, say, among attendees of a social gathering, such as a banquet, a party, or among customers at a cafeteria or restaurant. Likewise, when an outbreak occurs among patients in a hospital ward as we saw in the example of the outbreak of HIV in the dialysis unit. We will

later review other scenarios in which the retrospective cohort study is used in field epidemiology. The very unplanned nature of outbreak studies means that almost all outbreak studies are retrospective: they depend on a health jurisdiction having reported the unusual occurrence of cases. Most times those reporting to immediately also identify the common experience or the group that experienced an elevated risk.

Let us advance that during high-frequency epidemics affecting the general population, the study of residents of a neighborhood or a school or military institution may serve as study populations in which the retrospective cohort study is developed.

It will be of interest to readers of this section to note that often studies of primarily clinical interest of patients with a disease are published and referred to as cohort studies. During the COVID-19 pandemic many reports of such cohort studies were published. It is perhaps appropriate to call such investigations case series studies since they lack data from the population at risk.

Requirements

Again, the requirements are that the attack rate of Y is relatively high (≥10%), all or most of the cases had already occurred, and that cohort members are easy to enumerate and collect data on exposure to X prior to the occurrence of the outbreak or epidemic, so that there is no doubt that the exposure temporally precedes the occurrence of the disease or condition of interest.

It is necessary to have a way to measure exposure either directly by structured interview using a questionnaire, or a self-administered questionnaire, by consulting records such as menu orders if it is a banquet, product orders, or medical records or charts if it is an outbreak in a hospital, measurement of vaccine-derived antibodies or vaccination records. It is also necessary to be able to measure the occurrence of the disease or condition under study using an operational case definition and to do so in

a way that is complete both retrospectively and during and at least some time after the conclusion of the investigation.

Assessment of association

The measures of association used in retrospective cohort studies, as mentioned above, are the attack rate ratio, or risk ratio (RR), and the attack rate difference, or risk difference (RD). If there is information on follow-up time and date/time of occurrence of the disease or condition of interest (Y), one could estimate incidence rates, and the measures of association one can use are both the incidence rate ratio and the incidence rate difference. This type of detailed (person-time) information is either not available or the epidemic period is relatively short so in most cases there is no justification for using incidence rates over risk or attack rates.

The attack rate ratio or risk ratio (RR) among the exposed over the unexposed, i.e., $\frac{Risk\ in\ the\ exposed}{Risk\ in\ the\ unexposed}$, is the measure of association most often employed in retrospective cohort studies.

One of the advantages of cohort studies over other epidemiological study designs is that the attack rates in the study population can be obtained directly from the observations at hand. The attack rates or risk in the unexposed, subtracted from the attack rates or risk in the exposed, that is, the risk difference, (RD), is $Risk\ in\ the\ exposed - Risk\ in\ the\ unexposed \times k$, where k is a constant usually 100, is the other measure of association or effect employed in cohort studies. Note that in contrast to the risk ratio, the risk difference is a direct measure of disease occurrence, in this case risk, whereas the risk ratio is a relative measure. Although there are two advantages of using the risk difference over the risk ratio, one that it provides an absolute measure of risk and the second that it allows estimates to be obtained even when there are cells with zero as illustrated by the example in Table 2. The difference in attack rate or risks, divided by the attack rate or risk in the exposed,

$\frac{Risk\ in\ the\ exposed - Risk\ in\ the\ unexposed}{Risk\ in\ the\ exposed} \times 100$, is also called the attributable risk percent or simply attributable risk (AR), which is a measure of potential impact.

Since one of the possible sources of error is chance, i.e., that the observed differences are due to chance alone, by using statistical methods one wishes to set limits on this uncertainty. We do this in two ways. The first is by doing a statistical test which for categorical data is usually the chi-square test for the case of sufficient numbers, or a Fisher test for a situation with limited observations. Such tests compare the observed numbers with those expected under the so-called null hypothesis (H_0), produce a value for each test statistic which in turn has an associated *P*-value. The *P*-value measures the probability of the data given that the null hypothesis was true. *P*-values range from 0 to 1, and a cutoff point of less than 0.05 is used by convention to reject the H_0, i.e., that such a hypothesis is implausible given the observations at hand. In the case of RR, the null hypothesis, H_0, is RR=1; whereas, in the case of RD, the H_0 is RD=0. The second approach, preferred by epidemiologists and in this Journal, is to obtain a confidence interval around RR or RD. If the 95% confidence interval (CI) does not include the null value of RR or RD depending on whether one measure or the other is chosen, it provides the same answer as the *P* value. Furthermore, the CI gives us the point value of RR or RD that maximizes the probability of the data and the range most consistent with them, i.e., the CI is much more informative than the *P*-value.

Returning to the example in Table 2, for the risk of HIV infection for patients on chronic dialysis at the university hospital while patient 22 was on dialysis, the hazard ratio is undefined since the value of the denominator is zero, i.e., the rate for those not exposed was zero. The risk difference does not have this problem and is calculated as 90%. The 90% risk difference divided by the attack rate in the exposed (90%) and expressed as a percentage is 100%. This is the AR.

Interpretation of measures of association

Exposures that are deleterious to health increase the risk in the exposed and as the RR and DR values of the exposures are harmful, they are greater than 1 and 0, respectively. Exposures may have a protective effect on the risk of developing a disease or health-related condition. For example, vaccines intended as exposures to which people are exposed not only voluntarily but on the recommendation of health professionals, notably epidemiologists and other professionals, when and if effective decrease the risk of infection and disease or its serious outcome after immunization with them. The values of RR and RD if protective would be less than 1 and less than 0, respectively.

Since these risk measures (RR and RD) are different, i.e., the former is a measure of relative association and the latter is absolute, their interpretation is also different. One expresses RR as the number of times the risk of the exposed folds over the risk of the unexposed. In Table 2, exposure to being dialyzed simultaneously with patient 22, the risk of those patients increased infinitely or indefinitely compared to the risk of those not exposed, and we could also say that it was statistically significant with a *P*-value of 0.0001. If we were to refer to the RR of having been infected with HIV by having received more than two blood transfusions, we would say that the risk of those who received more than two blood transfusions was apparently 1.7 times the risk of those who did not, but which confidence interval did not allow us to establish that this difference could have been due to chance alone. The latter we assert because the 95% CI, (0.7, 4.2), included the null value and consistently the *P*-value was 0.3. One could also say that apparently the risk of HIV infection among those who received two or more transfusions had an excess risk of 70% (RR-1) relative to the risk of those who did not receive two or more transfusions, and again we can say that such an apparent excess may have been due to chance.

A common error in the interpretation of RRs is the incorrect use of the comparative “more” instead of the multiplicative doubled, tripled, quadrupled, or and so or nearly doubled, tripled, quadrupled and so on. It is also possible to say X-fold increase or increase in so many times the risk compared to that of the unexposed. We should explain that a RR of 2 does not mean that there is twice "more" risk of disease or condition “X", but that the risk of “X” was doubled among those exposed compared to the risk of those not exposed. If one insists on using the comparative more .. than, one should subtract 1 from the RR, i.e., a RR of 2, is correctly interpreted to mean that the risk of the exposed increased by 100% relative to the risk of the unexposed. Similarly, a RR of 5 means that the risk of the exposed is 400% more than the level of the unexposed. The same applies to the comparative greater ... than or less ... than, since it only implies that there is a difference, unless it is specified that it is X number of times greater or less. This is because RR is a measure on the multiplicative scale and not the additive scale.

The RD values are measured on the additive scale, i.e., they are not relative but absolute since the difference results in a risk value on the original scale. So, one of the simplest ways to interpret it is as excess risk or cases by the constant *k*, used to express risk, in the case of attack rates, these are generally per hundred people, in this case exposed. For example, returning to Table 2, exposure to being dialyzed simultaneously with patient 22 conferred an excess risk for 9 out of 10 patients so exposed to develop HIV infection and we could also say that this excess was statistically significant with a *P* value of 0.0001.

As we had written earlier, the ratio of RD over the rate of the unexposed, AR, is a measure not of association, but of potential impact, which is interpreted as the fraction of the disease or condition among the exposed that is due to their exposure. In the example in Table 2, the AR associated with receiving chronic dialysis at the same time as patient 22 was 100% (i.e., $\frac{90\%-0\%}{90\%} \times 100$) and can be

interpreted to mean that none of the HIV infections would have occurred in the dialysis unit had they not been transfused simultaneously with patient 22. This measure is very useful for evaluating the efficacy or effectiveness of an intervention such as vaccines (vaccine efficacy or effectiveness), but also of drugs and other procedures, technologies and methods of prevention, control, and health promotion.

Types of study settings

Expanding the scenarios in which the retrospective cohort study is employed during the course of an epidemic that we mentioned earlier, we can list the following settings:

Attendees at a discrete, one-time event (banquet [9], family/social gathering [10], religious services [11], concerts [12], transportation [13, 14], or other similar).

Common continuous exposure in institutions (schools [15], hospitals [6], other health facilities such as nursing homes [16], childcare centers [17], barracks [18], prisons [19]).

Workplaces (factories [20], offices [21]).

Homes when attack rates are high in the general population as in the COVID-19 pandemic [22] or vector-borne disease outbreaks as occurred in a Venezuelan equine encephalitis epidemic [23].

Generally, mostly outside the practice of field epidemiology, there are other settings in which planned cohort studies of longer follow-up periods are used, either retrospective or prospective, including pregnancy cohort studies [24], occupational cohort studies [25], studies of populations suffering from unusual exposures such as famine [e.g., 7,8] or atomic bombing [26], among others [1, pp. 44-46].

Case ascertainment and follow-up

During disease outbreaks it is common that almost all cases have already been reported. Sometimes a lack of completeness of case-reporting requires a more thorough search, involving contacting all or a

random sample of the at-risk population. The decision to survey all members of the universe (e.g., banquet attendees or factory employees or inhabitants of all households in an affected community) should be based on what is feasible. More important is the validity of the study (i.e., the absence of bias) in a sample that is feasible to enumerate, and care should be taken not to compromise such validity by an ambitious plan that produces estimates of association that could be biased. A major source of potential bias derives from conducting the investigation too late when the exposure can no longer be documented. For example, in a food poisoning outbreak where more than 2-3 days have passed and people cannot remember what they ate and implicated foods have been ruled out. For example, in a study conducted in Guizhou, China, after receiving a report of 200 cases of diarrhea on May 12, 2012, among university students [27]. Epidemiologists initiated the retrospective cohort study immediately. Most of the cases occurred between May 8 and 11. The epidemic was caused by *Aeromona hydrophyla* which has an incubation period of 1-2 days. The investigation identified a dose-response relationship with the amount of cucumber salad. Although they studied 902 students (AR=14%), many could not recall what they had eaten, and the report does not clarify whether the students were personally interviewed or whether the questionnaire was self-administered. Only 10% of the cases had microbiological tests. This bacterium is waterborne, and although cucumbers may have been the vehicle one would think that water may have been involved, given that there were histories of similar events associated with cafeteria water tank failure. One can reason that in these circumstances a study of 150 to 200 randomly selected students interviewed more carefully might have provided better information.

A surveillance system, perhaps active if warranted and feasible, should be in place to identify the occurrence of new cases in addition to cases already identified. These short-term, active surveillance systems, in general, are important to achieve a more accurate identification of the occurrence of the disease or health-related condition under investigation, as well as to monitor the occurrence into the

near future since many epidemic diseases, especially those spread from person-to-person, but also in those that are vehicle-borne, may have secondary or subsequent waves.

When collecting information on cases, the date of onset should be investigated as precisely as possible, and with as fine a precision as required according to the apparent incubation period of the disease. If it is food poisoning, it is likely that we will need the time and date of onset. There are diseases such as influenza or dengue fever where affected persons remember very accurately the time it started. Of course, this is of value if the inquiry is timely. There will be prevalent cases of certain conditions with chronic duration, such as HIV, where distinguishing them from new cases is important because prevalent cases are not part of the "at risk population", i.e., they are not susceptible.

The cohort study, whether prospective or retrospective, requires what is called a complete follow-up of the cohort, that is, the identification of the most comprehensive disease occurrence equally among both exposed and unexposed. Obviously, differentially identifying the occurrence of cases among the different exposure categories will introduce systematic errors or so-called biases. For example, if there were a greater effort to search for cases among the exposed than the unexposed, this would obviously artificially inflate the RR estimate by error. Alternatively, if the imbalance leads to a more thorough investigation of the unexposed, this could underestimate the association. The fieldwork of a retrospective cohort study should avoid the application of different case-finding methods according to exposures.

Exposure assessment

Often the exposures to be evaluated seem deceptively numerous. One should be guided by clues on the plausible biological or sociological causes of the disease or health-related event, the recall of patients during the exploratory phase of the investigation and resist the temptation to measure everything and all possible causes. As a rule, such approach would result in a dilution of the quality of

the information. Include possible factors that will be associated with disease occurrence in the absence of exposure and that may be associated with exposure, as well as those that one would expect to modify the effect of exposure on the risk. In the example of the dialysis unit HIV study, we noted that there was one patient who was retrospectively known to be already positive for the virus from the first visit to the dialysis unit (patient 22) and the task was to identify the visit dates of each patient to establish a timeline, which we have shown before how to do in a previous installment [28]. But HIV is transmitted by sexual intercourse, intravenous drug use, receipt of blood and organs, and among other risk factors these were studied in detail. Age, sex, marital status, occupation, and schooling were also abstracted by interview and from medical records.

In practice, in many outbreaks, exposure and health status are measured at the same time through the application of a questionnaire that has a section on health status and another on exposures. It is important to use questionnaires that have already been tested. They usually require calibration or adjustments to the situation under study and preferably being field tested before their application. As mentioned above, questionnaires are generally administered during an interview in person or by telephone. If there is reason to believe that the information collected by self-administration of the questionnaire is reliable, easy to answer, and minimize the possibility that answers are left blank, then this method or others such as the internet can be used, although it requires a minimum of literacy from the participant and consider the difficulties if a question may require the ability to do mental calculations. All these issues should be carefully evaluated. Field work should be multidisciplinary, recognizing the complementarity of the different professions.

Most of the rigor of our science lies to a large extent on the quality of the instrument we use. The importance of the questionnaire, as well as observation, cannot be overemphasized. The questionnaire should ask questions regarding exposures during the relevant period (i.e., recall or reference period),

use visual aids or references to holidays to enhance recall, while avoiding suggesting responses and introducing bias.

Unfortunately, there are many outbreak investigations that lack an epidemiological study component. It is common that instead of evaluating exposures using a questionnaire, other professionals, mainly microbiologists, prefer to rely on the swabbing of environmental samples, or the genotypic relationship between strains obtained from patients, neglecting the study of the population at risk and its measurement. Field work should be multidisciplinary, recognizing the complementarity of the different professions.

Sample size and statistical power

If the number of exposed subjects is limited, an attempt should be made to study all members of the cohort and it should be considered that there will be limitations inherent to the situation in which the outbreak occurred in order to adequately estimate the associations. Since outbreak field epidemiological investigations are not planned in advance, more often than not, rather than determining the number of subjects needed to test a hypothesis, it is a matter of knowing the probability that the hypothesis test would detect an association if it existed. In other words, the evaluation of the power of the study should be considered especially if an association that one would expect to have found was not found.

If a sample is to be drawn and a decision made as to how many persons to include, as a rule of thumb, 140 persons with half exposed and half unexposed in the cohort, it produces adequate estimates, with at least an 80% probability (i.e., 0. 8 as a fraction of 1, on the Y axis in the figures) of detecting the association, that is statistical power, if the RR to be estimated is 2.0 and above, provided the AR in the unexposed is at least 30%, but if the AR in this group is 20% we need at least 80 in each group, and if the AR in the unexposed is around 10% one has to study more than 200 in each group (Figure 2). Of

course for RR greater than 3 and hence for stronger associations one needs at least 63 subjects per group to reach a statistical power of 80% (Figure 3). Note that the chosen value of 80% power does not guarantee that associations will be detected if they exist.

When sampling it is not always obvious how to select the sample so that we are left with a 1:1 ratio of exposed to unexposed. Sampling should aim to have as many exposed as unexposed, in a balance close to 1:1, to reduce the sample size to the minimum necessary [29].

Figure 2. Statistical Power of Studies of 25 to 200 exposed and unexposed persons comparing attack rates from 10% to 40% in the unexposed and a RR of size 2

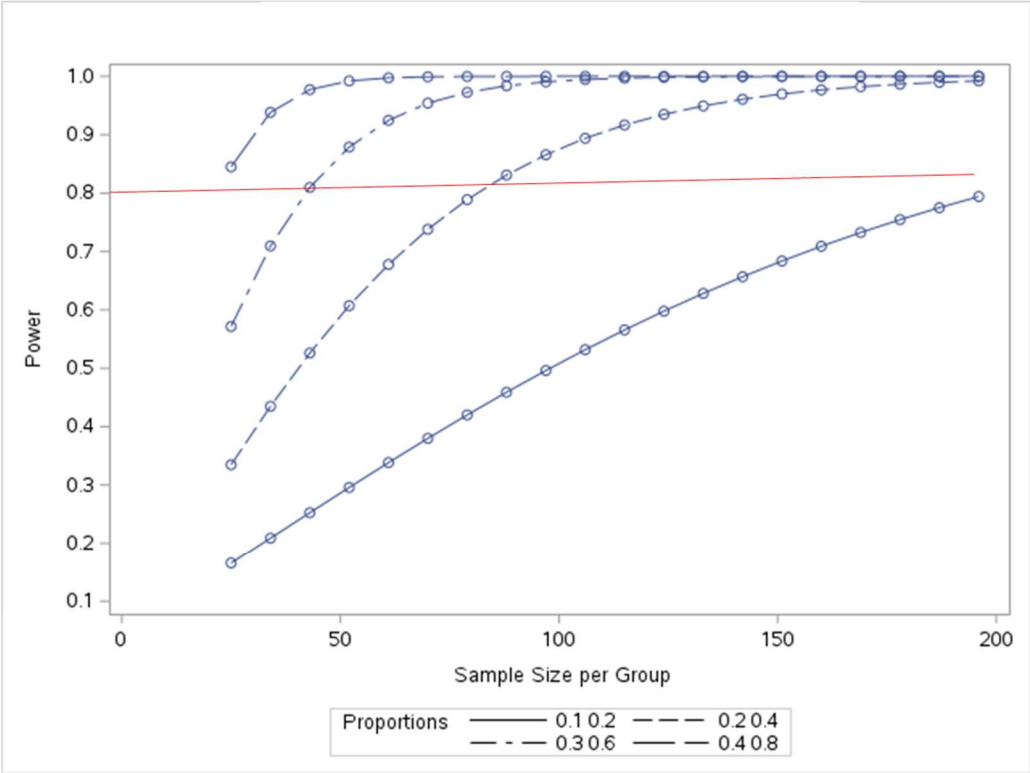
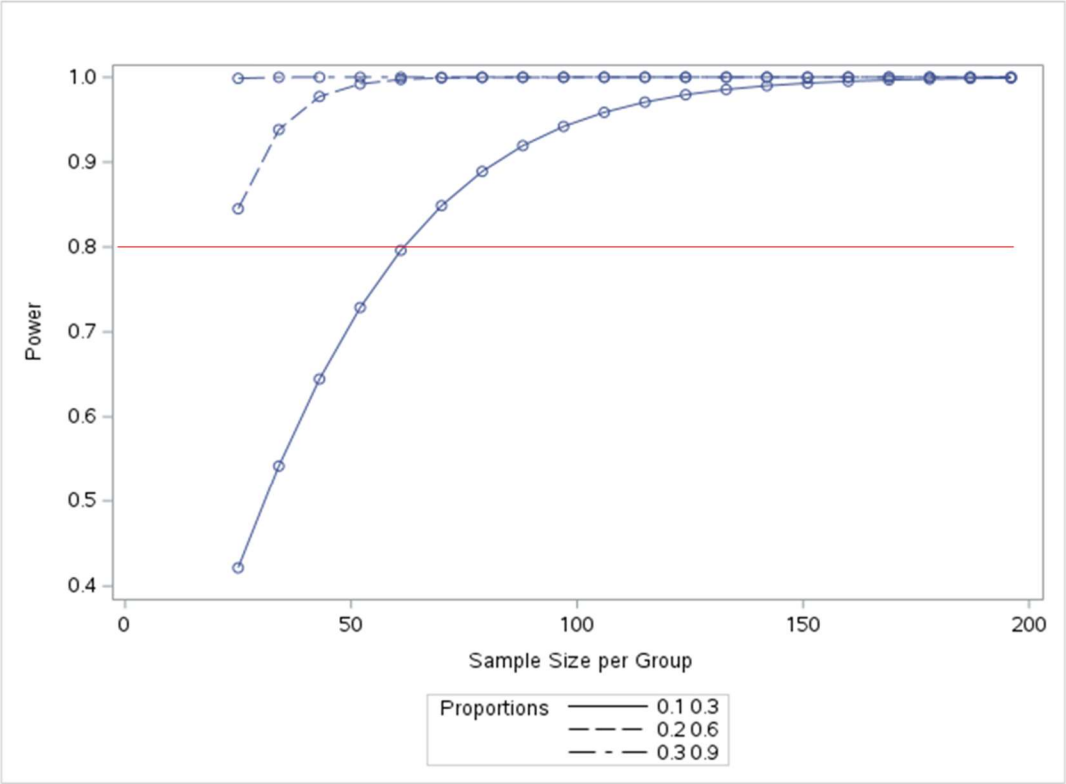


Figure 3. Statistical Power of Studies of 25 to 200 exposed and unexposed persons comparing attack rates from 10% to 30% an a RR of size 3



Data analysis

As a general rule the analysis should use simple methods that highlight the state of the relationship between exposures (X's), using a reference group (the unexposed) (X=0). The most commonly used, simple and easy to communicate is the 2 x 2 table, in which the attack rates between exposed and unexposed are compared. In food-borne outbreak studies it is called the consumption and non-consumption table. They also serve to present the desired measure of association, either the RR or the RD; then, we estimate the confidence interval with the level of statistical significance, usually 95%, and the *P*-value associated with the test statistic.

For exposure variables measured on a continuous scale, for example, the amount of water available in dwellings is a protective factor for the risk of diarrheal disease. One can examine the distribution in the entire cohort and examine whether it can be grouped into quintiles, quartiles, or tertiles and then

examine or some other cutoff points and contrast the ARs according to the categories created from the continuous variables. The same approach can be used for ordinal variables if such analysis. One can examine whether risk increases or decreases concomitantly with *j* levels of exposure (i.e., X=0, X=1, X=2...X*j*). The analysis can be as simple as describing whether there is variation in RR or RD using a level of this categorical variable as a referent. It may or may not be justified by performing a Mantel's chi-square test for trend [30] which assumes a linear trend, and by applying a polynomial regression with inflection points [31].

Pertussis outbreak in San Bartolomé Quialana, Mexico

To illustrate the 2x2 analysis we will use data from a field study of a pertussis outbreak in San Bartolomé Quialana, municipality of Tlacolula, Oaxaca, in 1988 [32]. In this outbreak, 22 sick children had been reported at the beginning and were treated at the health unit of the Mexican Institute of Health in that locality, so, together with a trainee from the Applied Epidemiology Residency program, advised by the author, recently graduated from that program, and a state epidemiologist, we conducted a census of 280 of the 364 houses in the village, identifying a total of 125 probable cases, confirmed by compatible symptomatology and the epidemiologic link. The overall AR was 7.4%, but among the 269 children under five years of age, there were 68 cases for an AR of 25.3%. All 17 pertussis deaths were in this age group. Among the under-fives, only 15.0% were found to have received the complete vaccination schedule with DPT, the biologic then used for prophylaxis. The data on vaccination history according to the national vaccination booklet are presented in Table 3. The AR in children with a complete vaccination schedule was about half the risk of those without vaccination. If we consider using the AR, the RD should be reversed by considering the unvaccinated as exposed and subtracting from the AT among them (31.9%) the AR of the vaccinated (17.9%), an AR of 14%, which is divided by the AR of

the vaccinated, allows us to estimate the AR which is an estimate of the vaccine effectiveness AR or VE=$\frac{14.0\%}{31.95}\sim 44.0\%$, since it is the proportion of the disease prevented among the vaccinated.

Table 3. Pertussis risk according to DPT vaccination among 7 to 59 month-old children in San Bartolomé Quialana, Tlacolula, Oaxaca, México, March to July 1988.

DPT vaccination	Pertussis cases (Attack Rate %)	Non-Cases	Total	Risk Ratio (95% CI)	*P*-value
Complete	5 (17.9)	23	28	0.56 (0.24, 1.29)	0.14
Incomplete	44 (31.9)	94	138	1 (Referente)	
Total	49 (29.6)	117	166	-	

Vaccine effectiveness = 44.0% (-29.0%, 76.0%)

Source: Reference [32]

Outbreak of Diarrheal Disease after the 1985 Earthquake in Mexico City

To illustrate the use of categorical or ordinal variables in the analysis of data from a retrospective cohort, we will use data from an outbreak of diarrheal disease after the September 19, 1985, earthquake in Mexico City [33]. A good part of the city had been without water service immediately after the earthquake and a government agency, CONASUPO, distributed undisinfected water in bags used to distribute milk. The outbreak was identified by the author in a survey of some dwellings in a low-income area without piped water supply including the Workers Neighborhood (Colonia Obrera), which led the Applied Epidemiology Residency program to conduct surveys in that and other areas of the city in October and December 1985. The distribution of attack rates according to water availability in low-

income area dwellings without water in October 1985 is presented in Table 4, which shows that diarrhea ARs did not increase gradually as water availability decreased, but rather as they reached a critical level of less than 10 liters of water per day.

Table 4. Risk of diarrheal disease in households of low economic income without piped water after the September 19, 1985, earthquake in Mexico City.

Availability of water (20 liter buckets/day)	Diarrhea cases (Attack Rate %)	Non-Cases	Total	Risk Ratip (IC 95%)	*P*-value*
<10	20 (7.7)	239	259	8.4 (3.7, 18.8)	
10 – 20	4 (0.9)	428	433	1.0 (0.3, 3.3)	<0.001
>20	8 (0.9)	859	867	1 (Referente)	
Total	32 (2.1)	1,527	1,559	-	

*Chi-square 2 df

Source: Reference [33]

Table 4 shows that there is no concomitant trend or variation per se between the decrease in water supply, i.e., the RR did not increase to say intermediate values, say between 2-4 for those who had between 10 and 20, but remained at the null value of 1, and only rose in 17% of the cohort with less than 10 buckets per day. Indeed, collapsing the 10+ buckets per day categories would yield the same result (RR=8.4). This example demonstrates a situation where it would be inappropriate to use a Mantel trend test using the assumption that the relation is linear, because there is no appreciable difference between the referent and the first level of exposure.

The analysis of data from a retrospective cohort would not be complete without considering the relation of the exposure variable of interest and the risk of the disease or condition of interest to variables that may confound or modify the association and which we will call Z. There are two stages of this analysis, which should preferably be performed sequentially: first the stratified analysis, followed by a multivariate analysis if it is considered that there is a basis for it. Leaving the details of such analyses for another occasion, in brief, stratified analysis requires disaggregating the association between X and Y by the levels of the Z's variables. Typically, age is a potential confounder because exposure is generally associated with age and in turn is usually an independent risk factor of the X variable of interest on the risk of the Y variable. Other potentially confounding variables are factors associated with the severity of the disease or condition of interest Y. Once the risk of Y is stratified by X by Z levels, and what is the weighted average across Z levels. The most popular of the methods for obtaining this summary measure of association measures according to the levels of the variable being stratified is the Mantel-Haenszel method [34]. If the value of the Mantel-Haenszel RR or RD, which we often call "adjusted" by Z, is similar (approximately within 10%) [35]. to the value of the unadjusted RR or RD, which we call "raw," there is no evidence of confounding, and one could ignore the stratified estimates. One has to observe whether the RR or RD changes according to Z levels and if it changes significantly there may be either the presence of effect modification, or as is also known, interaction.

It is important to note that confounding should be avoided because its presence distorts or biases the association estimates, i.e., produces invalid RR or RD values. One should try to avoid disregarding it if confounding is present in the field study data. In contrast, interaction or effect modification is a state of nature that one would want to describe if it has not been adequately described before.

In future installments of this section, we will return to the topics of epidemiological analysis, as well as validity, bias, confounding, and interaction.

References

1. Lash TL, VanderWeele TJ, Haneuse S, Rothman KJ. *Modern Epidemiology*.4th ed.Philadelphia PA: Wolters Kluwer, 2021.
2. Dwyer DM, Strickler H, Goodman RA, Armenian HK. Use of case-control studies in outbreak investigations. *Epidemiol Rev.* 1994;16(1):109-123. doi:10.1093/oxfordjournals.epirev.a036137.
3. Cardenas VM, Ramírez DR, Suárez Rangel GI. Analytic epidemiological studies as part of an epidemic investigation. *Am J of Field Epidemiol*. 2023; 1(2): 34–44. doi.org/10.59273/ajfe.v1i2.7849
4. Koopman JS, Longini IM Jr. The ecological effects of individual exposures and nonlinear disease dynamics in populations. *Am J Public Health.* 1994;84(5):836-842. doi:10.2105/ajph.84.5.836
5. Schaffner W, LaForce FM. Training field epidemiologists: Alexander D. Langmuir and the epidemic intelligence service. *Am J Epidemiol*. 1996;144(8 Suppl):S16-S22. doi:10.1093/aje/144.supplement_8.s16
6. Velandia M, Fridkin SK, Cárdenas V, et al. Transmission of HIV in dialysis centre*. Lancet.* 1995;345(8962):1417-1422. doi:10.1016/s0140-6736(95)92603-8
7. Stein Z, Susser M. The Dutch Famine, 1944–1945, and the Reproductive Process. I. Effects on Six Indices at Birth. Pediatr Res 1975; 9: 70–76. doi.org/10.1203/00006450-197502000-00003
8. Stein Z, Susser M, Saenger G, Marolla F. *Famine and human development: The Dutch hunger winter of 1944-1945.* 1975. New York: Oxford University Press.

9. Kemble SK, Westbrook A, Lynfield R, et al. Foodborne outbreak of group a streptococcus pharyngitis associated with a high school dance team banquet--Minnesota, 2012. *Clin Infect Dis*. 2013;57(5):648-654. doi:10.1093/cid/cit359
10. Haselow DT, Safi H, Holcomb D, et al. Histoplasmosis associated with a bamboo bonfire--Arkansas, October 2011. *MMWR Morb Mortal Wkly Rep*. 2014;63(8):165-168.
11. Nordt SP, Minns A, Carstairs S, et al. Mass sociogenic illness initially reported as carbon monoxide poisoning. *J Emerg Med.* 2012;42(2):159-161. doi:10.1016/j.jemermed.2011.01.028
12. Botelho-Nevers E, Gautret P. Outbreaks associated to large open air festivals, including music festivals, 1980 to 2012. *Euro Surveill*. 2013;18(11):20426. Published 2013 Mar 14. doi:10.2807/ese.18.11.20426-en
13. de Laval F, Chaudet H, Gorgé O, et al. Investigation of a COVID-19 outbreak on the Charles de Gaulle aircraft carrier, March to April 2020: a retrospective cohort study. *Euro Surveill.* 2022;27(21):2100612. doi:10.2807/1560-7917.ES.2022.27.21.2100612
14. Walker LJ, Codreanu TA, Armstrong PK, et al. SARS-CoV-2 infections among Australian passengers on the Diamond Princess cruise ship: A retrospective cohort study. *PLoS One*. 2021;16(9):e0255401. Published 2021 Sep 7. doi:10.1371/journal.pone.0255401
15. Simone B, Atchison C, Ruiz B, et al. Investigating an outbreak of *Clostridium perfringens* gastroenteritis in a school using smartphone technology, London, March 2013. *Euro Surveill*. 2014;19(19):20799. Published 2014 May 15. doi:10.2807/1560-7917.es2014.19.19.20799
16. Bouza E, Pérez-Granda MJ, Escribano P, et al. Outbreak of COVID-19 in a nursing home in Madrid. *J Infect.* 2020;81(4):647-679. doi:10.1016/j.jinf.2020.06.055
17. Gingrich GA, Hadler SC, Elder HA, Ash KO. Serologic investigation of an outbreak of hepatitis A in a rural day-care center. *Am J Public Health*. 1983;73(10):1190-1193. doi:10.2105/ajph.73.10.1190

18. de Laval F, Chaudet H, Gorgé O, et al. Investigation of a COVID-19 outbreak on the Charles de Gaulle aircraft carrier, March to April 2020: a retrospective cohort study. *Euro Surveill*. 2022;27(21):2100612. doi:10.2807/1560-7917.ES.2022.27.21.2100612
19. Hutchinson JA, Wheeler C, Mohle-Boetani JC. Outbreak epidemiologically linked with a composite product of beef, mechanically separated chicken and textured vegetable protein contaminated with multiple serotypes of *Salmonella enterica* including multidrug-resistant *infantis*, California 2016. *Epidemiol Infect.* 2018;146(4):430-436. doi:10.1017/S0950268817002941
20. Finci I, Siebenbaum R, Richtzenhain J, et al. Risk factors associated with an outbreak of COVID-19 in a meat processing plant in southern Germany, April to June 2020. *Euro Surveill*. 2022;27(13):2100354. doi:10.2807/1560-7917.ES.2022.27.13.2100354
21. Nicolay N, Boland M, Ward M, et al. Investigation of Pontiac-like illness in office workers during an outbreak of Legionnaires' disease, 2008. *Epidemiol Infect.* 2010;138(11):1667-1673. doi:10.1017/S0950268810000403
22. McCarthy KL, James DP, Kumar N, et al. Infection control behaviours, intra-household transmission and quarantine duration: a retrospective cohort analysis of COVID-19 cases. Aust *N Z J Public Health*. 2022;46(6):730-734. doi:10.1111/1753-6405.13282
23. Rivas F, Diaz LA, Cardenas VM, et al. Epidemic Venezuelan equine encephalitis in La Guajira, Colombia, 1995. *J Infect Dis.* 1997;175(4):828-832. doi:10.1086/513978
24. Cardenas VM, Cen R, Clemens MM, et al. Use of Electronic Nicotine Delivery Systems (ENDS) by pregnant women I: Risk of small-for-gestational-age birth. *Tob Induc Dis*. 2019;17:44. Published 2019 May 21. doi:10.18332/tid/106089

25. Selikoff IJ. Lung cancer and mesothelioma during prospective surveillance of 1249 asbestos insulation workers, 1963-1974. *Ann N Y Acad Sci.* 1976;271:448-456. doi:10.1111/j.1749-6632.1976.tb23146.x
26. Ozasa K, Shimizu Y, Suyama A, et al. Studies of the mortality of atomic bomb survivors, Report 14, 1950-2003: an overview of cancer and noncancer diseases [published correction appears in Radiat Res. 2013 Apr;179(4):e40-1]. Radiat Res. 2012;177(3):229-243. doi:10.1667/rr2629.1
27. Zhang Q, Shi GQ, Tang GP, Zou ZT, Yao GH, Zeng G. A foodborne outbreak of Aeromonas hydrophila in a college, Xingyi City, Guizhou, China, 2012. *Western Pac Surveill Response J.* 2012;3(4):39-43. Published 2012 Dec 19. doi:10.5365/WPSAR.2012.3.4.018
28. Cardenas VM, Suárez-Rangel GI, Ramírez DR. Timelines in epidemiological outbreak investigations. *Am J Field Epidemiol.* 2023; 1(1): 71–76. https://doi.org/10.59273/ajfe.v1i1.7521
29. Walter SD. Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. *Am J Epidemiol*. 1977;105(4):387-397. doi:10.1093/oxfordjournals.aje.a112395
30. Mantel N. Chi-Square Tests with One Degree of Freedom; Extensions of the Mantel-Haenszel Procedure. *JASA* 1963; 58:303, 690-700, DOI: 10.1080/01621459.1963.10500879
31. Greenland S. Dose-response and trend analysis in epidemiology: alternatives to categorical analysis. *Epidemiology.* 1995;6(4):356-365. doi:10.1097/00001648-199507000-00005
32. Peña MP, Cárdenas VM, Ruiz C, Stetler HC, López O, Ibarra J, Sapiaín LA, Villafán F. Estudio de un brote de tosferina en San Bartolome Quialana, Oaxaca, 1988. *Boletín Mensual de Epidemiología del Sector Salud de México* 1989 4; (8): 112-118.
33. Ruiz Matus C, Cárdenas Ayala V, Koopman J, Herrera Bastos E, Montesano Castellanos R, Hinojosa M. Enfermedad diarreica después de los sismos de 1985 en México [Diarrheal disease after the 1985 earthquakes in Mexico]. *Salud Publica Mex.* 1987;29(5):399-405.

34. Mantel N, Haenszel W. Statistical aspects of the analysis of data from retrospective studies of disease. *J Natl Cancer Inst.* 1959;22(4):719-748.
35. Greenland S. Modeling and variable selection in epidemiologic analysis. *Am J Public Health*. 1989;79(3):340-349. doi:10.2105/ajph.79.3.340