Chapter IV

TRANSVERSAL STUDIES

Pierre Duhaut, Jean Schmidt

In clinical practice, most issues are addressed with reference to the frequency of the event under consideration. These are fractions, or proportions, which give an idea of the frequency of these clinical events, with the number of cases in the numerator, and the population from which these cases come in the denominator.

The first of the measures of frequency, discussed in this chapter, is prevalence.

In medicine, counting events, whether beneficial or adverse (death, illness, disability, discomfort, dissatisfaction and their opposite), is the essential prelude to any subsequent analysis or interpretation. This is the fundamental objective of the prevalence study, or cross-sectional study, the advantages and limits of which will appear at the end of the chapter. It is indeed difficult, beyond observation and counting, to establish the temporal sequence of the events considered. Moreover, estimating causality is always a hazardous process in this context... but it is a good starting point.

Cross-sectional studies, also called prevalence studies, are so named because they analyze the presence of a given factor or of a particular disease in a population P at a specific time t, without reference to the past and without follow-up in the future. They represent the equivalent of a rigorously and scientifically constructed survey, or a photographic snapshot of a precise situation in the population studied.

Cross-sectional studies are primarily descriptive, not analytical like case-control studies, cohort studies or randomized trials. They are particularly useful for providing precise quantitative knowledge on the distribution of a disease or a risk factor in a population, its frequency, and the subgroups of the population that are more particularly affected.

The results of cross-sectional studies are therefore important in two main areas of application:

- The implementation of public health programs, preventive or curative, by making it possible to define the groups of the population in which the program must be applied as a priority (age groups, urban population or rural population, men or women, geographic, etc.). The optimal definition of the scope of the program allows an optimal allocation of the human and material resources devoted to it and represents one of the essential conditions for its effectiveness. For example, the study of the prevalence of resistant and non-resistant forms of malaria in the various regions of the world allows the adequate implementation of WHO malaria control programs, and is at the origin of the type of prophylactic advice given to travellers.
- The observation in a cross-sectional study of associations between a pathological state and one or more conditions, that can be assumed to be causal leads to the formulation of etiological hypotheses to be tested in other studies of a different nature (biological or epidemiological). For example, the association between seropositivity for hepatitis B and

hepatocarcinoma in South-East Asia led to the carrying out of case-control studies, then cohort studies which proved the causal relationship, in conjunction with biological studies showing the integration of the virus genome into the DNA of neoplastic cells.

Descriptive studies include:
- on the one hand, case reports, case series and ecological studies;
- on the other hand, prevalence studies, which represent a particular type of descriptive study, on the borderline of analytical studies (which are case-control studies, cohort studies and, on the experimental side, randomized trials) .

I - CASES-REPORTS, CASE SERIES AND ECOLOGICAL STUDIES

A - CASES REPORTS

Case reports, describing an unusual observation, are often the first step in recognizing a new disease or risk factor. For example, the association of thromboembolism and estrogen-progestins was reported for the first time in 1961 in a patient, widely discussed in larger series before being the subject of multiple studies, including case-controls studies, proving its reality .

B - THE CASE SERIES

They represent the next step by grouping together different similar observations and thus establishing the probable existence of a pathological entity. In some cases, they can very strongly suggest an etiological factor.

For example, Thomas Hodgkin in 1832 had identified 7 patients with similar tumor abnormalities of the spleen and lymph nodes, 70 years before the Sternberg cell was described as pathognomonic of the disease and the nosological entity could thus be formed. Closer to home, the diagnosis of Pneumocystis carinii pneumonia with oral candidiasis in 4 young subjects with no particular history, male homosexuals, led to the discovery of AIDS and already carried the germ of recognition of one of the risk factors for disease .

Case series and reported cases, however, most often reflect the experience and observation of an author and do not allow conclusions to be drawn that can be generalized to other cases. Moreover, and despite their indisputable usefulness, case series and a fortiori reported cases do not make it possible to establish the frequency of a disease: an incidence or prevalence study would be necessary for this. Nor do case series make it possible to statistically assess the importance of a risk factor that they may possibly suggest. A comparison group would be needed here.

C - CORRELATION STUDIES

Correlation studies, or ecological studies, allow analysis on a larger scale. They establish the comparison between the importance (or the frequency) of a supposed risk factor within a population and the prevalence or the incidence of the supposed secondary disease, from data already available at the level of this population (Is the incidence of lung cancer in various regions of the globe proportional to the amount of cigarettes smoked in these regions? Is the incidence of cardiovascular disease proportional to the amount of animal fat ingested?).

Ecological studies therefore do not use data collected at the individual level, but averages calculated at the level of a population.

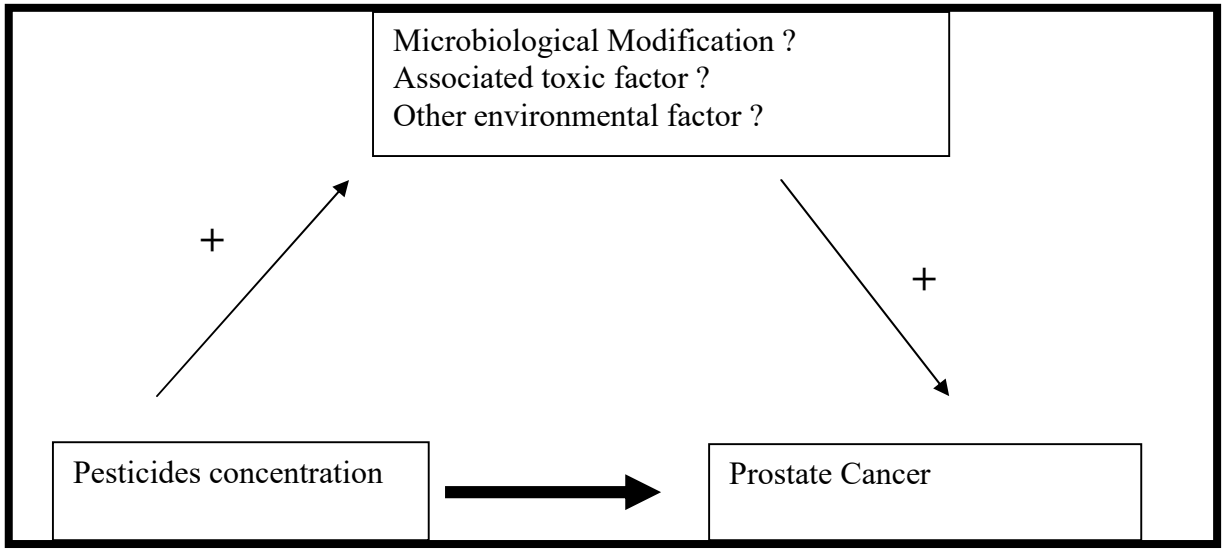
On the one hand, ecological studies introduce the notion of comparison: it is necessary to have data from different populations in order to be able to establish a correlation between the importance of a risk factor and the importance of the disease studied in each population. They also use the notion of frequency of the risk factor and of the disease. They are cross-sectional insofar as they superimpose two types of data (frequency of the risk factor and frequency of the disease) collected over the same period of time. Finally, they are easy to carry out in a limited time, because they use descriptive statistical data that has already been collected and published. They make it possible to put forward interesting hypotheses, such as the possible role of pesticides in the pathogenesis of prostate cancers, the incidence of which differs according to the degree of exposure in Martinique.

However, they have significant flaws that make their interpretation hazardous:

Data are averages describing the characteristics of a population. They do not make it possible to know whether the person exposed is actually the one who developed the disease, and therefore whether the risk factor should be considered as such. If there is a correlation between the concentration of herbicides and prostate cancer, is it really the people – at the individual level – most subjected to pesticides who develop the most prostate cancer? Nor do ecological studies allow control of confounding factors, even if a multivariate analysis can be performed.

For example, one can imagine that the quantity of pesticides present in the soil and water can modify the microbiological environment, and promote the emergence of a viral or other carcinogen. Where will the carcinogen implicated in the pathogenesis of prostate cancer be? In herbicides, the modified factors of the microbiological environment, or even any other environmental factor in the broad sense of the term, including dietary factors, toxic factors, indicators of standard of living - including a whole set of social determinants at a large scale, - some of which of individual, familial or societal consumption-, all correlated to the previous ones? What are the interactions between these different factors (fig. 1)?

Fig. 1 - Theoretical example of a confounding factor in the association of pesticides and prostate cancer.



Finally, a real risk factor may not be identified by an ecological study if it is "diluted" by the specific characteristics of the population. For example, the relationship between saturated fat intake and coronary disease is easy to demonstrate in relatively old populations, where the incidence of coronary disease is high. The same relationship in populations with a younger average age, consuming large amounts of saturated fat, could go unnoticed by the "dilution" effect of the age group at risk in other segments of the population.

Ecological studies, allowing large-scale comparisons, however suffer from the lack of individual information: what and who are we studying precisely? What are the confounding factors that can explain the observed association by their relationship with each of the associated variables? To what extent does the association of averages calculated in a population describe the association of risk factor - disease actually present at the level of the affected individual?

II - PREVALENCE STUDIES

A - INCIDENCE VERSUS PREVALENCE

Prevalence indicates the percentage of people with the disease at a given time in a given population:

$$\text{Prevalence} = \frac{\text{Number of sick people at time t}}{\text{Total population considered}}$$

Incidence indicates the percentage of new cases diagnosed over a period of time in a given population:

$$\text{Incidence} = \frac{\text{New cases diagnosed during a period p}}{\text{Total population considered}}$$

Prevalence is most often expressed as the number of cases/100,000 inhabitants, and incidence as the number of cases/100,000 inhabitants/year.

Cross-sectional studies do not take into account the time variable. Studies determining the incidence of a disease are not, by definition, cross-sectional studies. The incidence is determined by cohort studies.

Prevalence is higher than incidence in case of chronic disease, and lower in case of acute disease (curable or not).

Examples:

- The prevalence of rheumatoid arthritis is far greater than the number of new cases diagnosed each year (incidence). It therefore gives a better idea of the burden of the disease and its social and individual consequences than the incidence.
- The number of people poisoned by amanita phalloides on April 30 (prevalence) is on the other hand lower than the annual incidence of poisoning, and only gives a very incomplete idea, even without value, of the importance of the problem. However, the prevalence in a given region makes it possible to estimate the number of resuscitation beds needed to accommodate affected subjects.

- The number of smokers in France (prevalence) makes it possible to fully measure the social importance of the phenomenon, a major decision-making element in the discussion of the advisability of an anti-smoking campaign. The number of people starting to smoke during the previous 6 months and following a warning campaign against the harmful effects of tobacco (incidence) makes it possible to measure a trend and assess the effectiveness of the campaign.

Prevalence and incidence therefore each have their own usefulness, and provide different and complementary information. Prevalence studies combine the respective qualities of case series and ecological studies, while eliminating some of their shortcomings:

- the collection of data is done at the level of individuals, and makes it possible to identify possible confounding factors and to control them;
- the conditions of selection of the group studied allow calculations of descriptive statistics;
- there is a comparison group. However, we will see that we must be wary of.

B - CONSTITUTION OF A PREVALENCE STUDY

It proceeds in successive stages, similar to those of any epidemiological study:

- what is the question ? How to ask the question?
- what population is it aimed at?
- how to select a representative sample of this population?
- how to quantify and analyze the data?
- how to interpret the results?
- in view of the answers given to these questions or the problems they raise: ultimately, is the structure of a prevalence study adapted to the problem to be solved? If not, what other type of epidemiological study would it be better to choose?

1 - Question and population, sampling, bias

The question can be simple (what is the prevalence of disease X in population Y) or double (what is the prevalence of disease X in population Y, and is there an association with the factor Z?).

The population must be defined very precisely both geographically and in terms of individual characteristics (age, sex, etc.).

Examples:

- Study of the prevalence of hemochromatosis in the Picardy region. Any modification of the geographical area can be at the origin of different results: the distribution of the genes of the HFE system, more important in the populations of Nordic origin, close to that observed in Brittany, can no longer be the same if the study is extended to Champagne or the neighboring Ile-de-France.
- Study of the prevalence of allergic asthma in children in the Paris region. The inclusion or exclusion of neighboring rural departments is likely to profoundly modify the figures, because the distribution of allergens is not ubiquitous: the vegetation differs considerably in departments with a strong agricultural dominance (Picardy), livestock (Normandy) or mainly urban (Ile-de-France).

If the population considered is large, sampling becomes necessary (fig. 2). It must be representative of the initial population and of sufficient size to allow valid conclusions to be drawn. The size estimate depends on the question being asked and the assumed prevalence of the factors being measured.

Reference population

Not exposed Not sick	Not exposed Sick
Exposed Not sick	Exposed Sick

Fig. 2 - Reference population and its subgroups

The representativeness of the sample can only be ensured by drawing lots, provided that any person in the initial population has a probability equal to that of any other person of being drawn at random. This supposes having a complete and current list of the population, where each individual appears only once, under a single number or identification code. These "perfect" lists are rare in practice when looking at the general population. The usual databases (telephone directory, census data, list of insured persons, etc.) only come close. They are even more difficult to obtain if one wishes to study a particular sociological sub-group. The list of employees of a company represents a privileged example of a perfect list when studying an occupational disease.

The next step is to define the cases, and the difficulties are not unique to cross-sectional studies. The problem of defining exposure arises if the study is not limited to defining a prevalence, but seeks to measure the association of the disease with a presumed risk factor. Here again, the difficulties are not specific to cross-sectional studies: knowing when a subject is subject to a risk factor proceeds from the same questions as in case-control studies and cohort studies.

The type of the "exposure" variable, on the other hand, makes the interpretation of the results more or less risky, and three scenarios may arise:

- Exposure is fixed over time and does not a priori influence the age of onset and the length of progression of the disease.

These risk factors must be characteristics present at birth and not undergoing modification during life.

In the study of the relationship between ankylosing spondylitis and HLA B27 antigen, the question of temporality does not arise, the HLA B27 antigen being present before the onset of the disease and its nature not varying during life. A cross-sectional study can easily show the higher prevalence of the antigen in the sick population than in the healthy population.

- Exposure is fixed over time, but can influence the age of onset or the duration of disease progression.

As before, the risk factor must be present at birth and not undergo any modification during life. The question of temporality does not arise, since the factor considered necessarily precedes the disease (if it is acquired during life). But the influence of the risk factor on the

age of onset or the duration of disease progression may be the source of a selective survival bias.

A cross-sectional study aimed at testing the association between trisomy 21 and acute leukemia in *the adult population* could show a lower frequency of trisomy in leukemia patients and wrongly conclude that trisomy protects against leukemia, because children with trisomy, who are more at risk than children without Down syndrome from developing leukemia and dying from it young, would not be included in the cross-sectional study because of their early death (fig. 3).

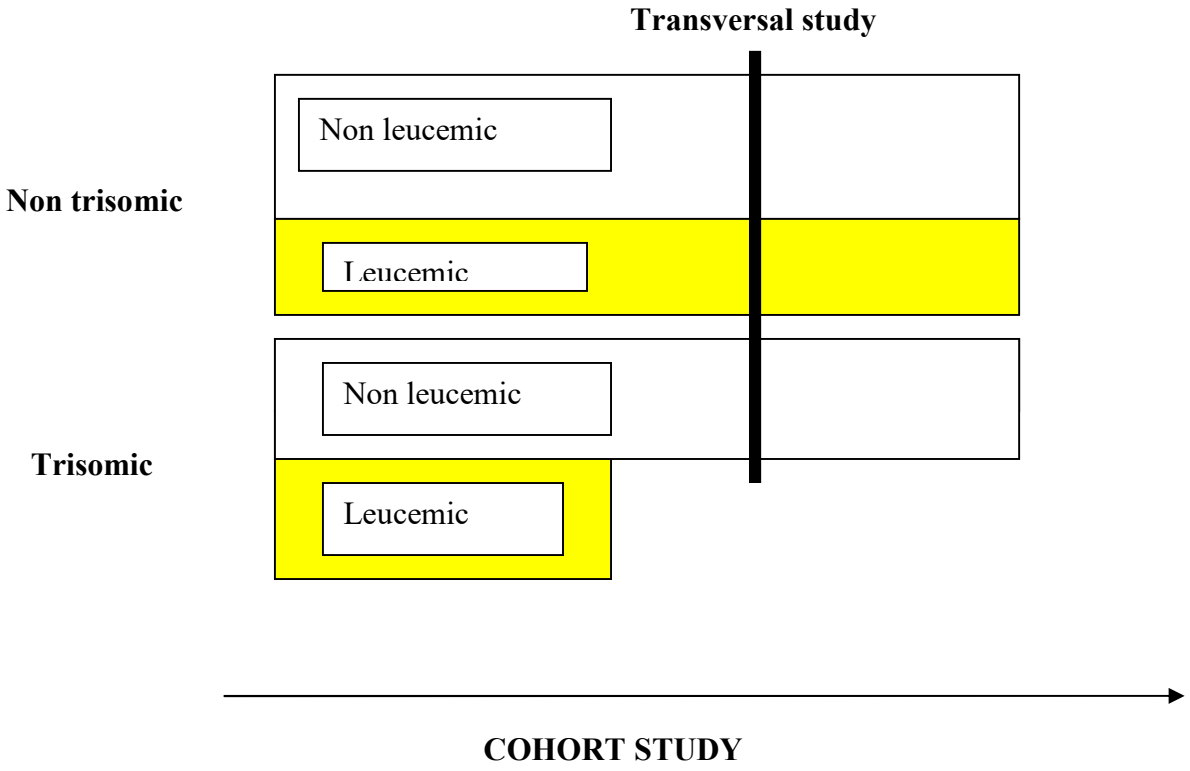


Fig. 3 - Association between trisomy 21 and acute leukaemia: representation of a cross-sectional study in an adult population (vertical line) and of a cohort study (horizontal arrow).

The cohort study would have shown the increased risk of acute leukemia in subjects with Down syndrome. The cross-sectional study, which cannot include Down's syndrome leukemia, wrongly concludes that Down's syndrome has a protective effect and therefore eliminates the causal relationship.

In practice, it is often easy to know whether the assumed risk factor is fixed over time. It is much more difficult to know precisely whether it influences the age of onset and the duration of disease progression, and differentiating between cases 1 and 2 is not always easy. The suspicion of selective survival bias therefore also exists when one thinks, without being able to be absolutely sure, that one is in situation 1.

- *The exposure is not fixed in time.*

It is a risk factor acquired at some point in life, and the intensity of which may vary over time. This is the case for the majority of risk factors studied in pathology (nutritional factors, smoking, alcoholism, viral, bacterial, parasitic contamination, occupational exposure, accidental poisoning, etc.).

The cause and effect relationship is very difficult to establish here, because there are two major problems:

- *The temporal sequence:* disease and supposed risk factor are determined at the same time. Which preceded the other? The answer may be easy if one can reliably recognize in the past an exposure that occurred on a specific date (nuclear accident and prevalence of congenital malformations in the affected population). It can be much more difficult, even impossible to obtain by a cross-sectional study alone, in other situations and especially when the pathophysiology of a disease remains mysterious. In the association between the presence of anti-nuclear antibodies and the manifestations of lupus disease, are anti-nuclear antibodies the cause of the lesions observed, or do they only appear as a consequence of cell destruction caused by factor X, bringing intracellular antigens into contact with the immune system, which can then produce antibodies against the antigens thus exposed?

The question has not yet been definitively resolved, even if it seems that the lifting of antibodies is a harbinger of the progressive resumption of the disease, which is perhaps only the final, apparent stage of the initial destructive process revealed more early by the re-ascent of the level of antibodies.

- *The accuracy of the exposure measurement:* when exposure varies over time, what is better to measure? Exposure at the time of the cross-sectional study, which can be defined with maximum accuracy, but which, concomitant with the pathological state, is not necessarily that which induced the disease? Or exposure in the past, more likely to have induced the disease, especially when there is a long latency period, but whose determination is based on the memories of the subjects and is often imprecise?

Example: prevalence of cardiovascular diseases and content of short-chain fatty acids in the diet. Quantifying the lipid content of the subjects' diet at the time of the cross-sectional study is possible. But does the current diet reflect the diet of past years, a real risk factor? On the other hand, how to measure the lipid content of the subjects' diet 5, 10, or 20 years before the study was carried out?

Recall bias is one of the very important limiting factors of cross-sectional studies.

Finally, in most cases, the acquired risk factor, which varies over time, influences the age of onset and the duration of disease progression. To the bias of memory and the problem of temporality, there is therefore added the bias of selective survival, and the interpretation of the cross-sectional study is all the more uncertain.

2- Measurements made, mathematical expression

The results can be expressed in table form (Table 1).

	ILL	NOT ILL	
EXPOSED	a	b	a + b
NO EXPOSED	c	d	c + d
	a + c	b + d	a + b + c + d

Table 1: Expression of the results for a prevalence study

In this presentation, the prevalence (the only really rigorous measurement authorized by this type of study), is written:

$$\text{Prevalence} = \frac{a + c}{a + b + c + d}.$$

We can also define a prevalence rate, which answers the following question: how often is the disease more frequent in exposed subjects than in non-exposed subjects, in the population examined in the cross-sectional study?

It is necessary to calculate the prevalence of the disease in the exposed subjects (Prevalence 1), and in the unexposed subjects (Prevalence 2):

$$\text{Prevalence 1} = \frac{a}{a + b} \qquad \text{Prevalence 2} = \frac{c}{c + d}.$$

$$\frac{\text{Prevalence 1}}{\text{Prevalence 2}} = \frac{a}{a + b} \times \frac{c + d}{c}$$

It should be noted that the prevalence rate is not the equivalent of a relative risk, which would answer the following question: how many times the exposed subjects have a greater risk of being affected by the disease than unexposed subjects? The relative risk calculated in a cohort study or approximated in a case-control study measures the "pathogenic power" of the exposure factor.

In order for the prevalence rate to approach the relative risk, it would be necessary to:

- there is no selective survival bias;
- there is no recall bias;
- that there is a real causal relationship between the supposed risk factor and the disease, which is impossible to prove by an isolated cross-sectional study, due to its very structure; in the cross-sectional study, there is a juxtaposition of a supposed risk factor and the disease, but we do not know which preceded the other.
- that the duration of the disease in the exposed subjects is the same as the duration of the disease in the unexposed subjects, so that exposed and unexposed sick subjects have the same chance of being included in the cross-sectional study in as patients. If this were not the case, we would risk finding ourselves in the following situation (fig. 4): none of the unexposed

subjects who contracted the disease were included as patients in the cross-sectional study, and yet the risk of contracting the disease for the unexposed (3/3) is the same as for the exposed subjects (3/3). The comparison of prevalences would lead to the false conclusion that the disease does not exist in unexposed subjects and therefore that they are not at risk for the disease considered.

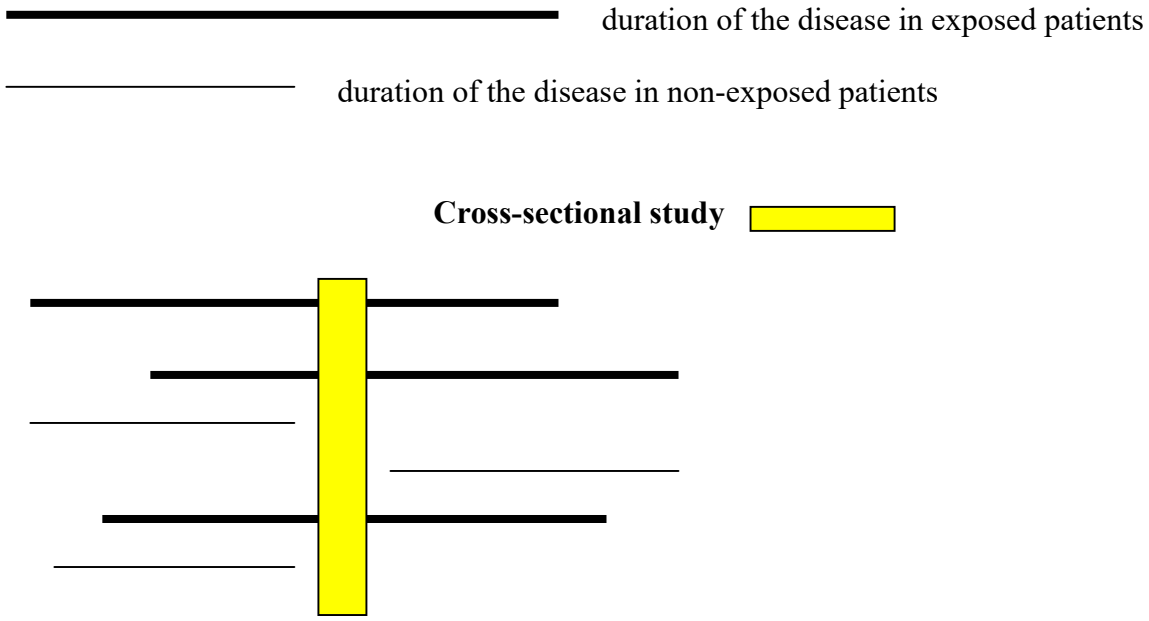


Fig. 4 - Influence of disease duration in a cross-sectional study

On the other hand, the interesting information, namely the increase in the duration of the disease in the exposed subjects, is not obtained by the cross-sectional study.

Abusively interpreting a prevalence rate therefore leads to erroneous conclusions. Since the four conditions detailed above are rarely met, one should not infer from the prevalence rate any causal relationship or the importance of a relative risk. The prevalence rate therefore only answers the question, quickly posed, quickly resolved, of the relative frequency, at a time t, of the disease among subjects exposed and not exposed to a supposed risk factor. It leaves many other questions unresolved. The observation of a difference in frequency can however serve as a hypothesis for a case-control study, a cohort study or biological experiments aimed at confirming or invalidating the pathogenic role of the risk factor.

III - ADVANTAGES AND WEAKNESSES OF CROSS-SECTIONAL STUDIES

A- Advantages

- Cross-sectional studies are the only ones that can establish prevalence. This measurement is particularly useful for assessing the extent of a phenomenon, the social repercussions of a disease, its geographical distribution. It is necessary in order to be able to adjust the number and quality of healthcare structures to the needs encountered in the population.
- They have a comparison group and thus make it possible to study the association between a pathological state and a supposed risk factor.
- They make it possible to study simultaneously the association between several pathological states and several supposed risk factors. Thus, they serve as generators of hypotheses for more elaborate studies such as case-control studies or cohort studies.
- They can represent a first step in a cohort study (subject inclusion phase).
- They can be carried out in a relatively short period of time, and are therefore inexpensive.
- Possible confounding factors can be controlled by stratifying sick and healthy subjects according to the element of confounding.

B - Weaknesses

- They do not make it possible to establish the temporal sequence of events. Finding an association between a pathological condition and a supposed risk factor therefore does not allow us to deduce a cause and effect relationship.
- They do not make it possible to estimate an association when the disease is rare in the population, because they would require too large a sample size to be able to include a sufficient number of sick subjects.
- They are subject to the possibility of selective survival bias.
- They are subject to memory bias.
- They are subject to the always possible existence of unforeseen confounding factors.
- The prevalence does not make it possible to estimate the incidence, and the prevalence ratio does not make it possible to estimate the relative risk.
- Finally and above all, we must beware of any abusive, often tempting, interpretation.

References

1. Ganem D, Prince AM. Hepatitis B infection- Natural history and clinical consequences. NEJM 2004;350:1118-1129.

2. Tyler ET. Oral contraception and venous thrombosis. JAMA 1963 ;185 :131-132.

3. WHO Collaborative Study of Cardiovascular Disease and Steroid Hormone Contraception. Effect of different progestagens in low oestrogen oral contraceptives on venous thromoembolic disease. Lancet 1995;346:1582-1588.

4. Gottlieb MS, Schroff R, Schanker HM, Weisman JD, Fan PT, Wolf RA, Saxon A. Pneumocystis carinii pneumonia and mucosal candidiasis in previously healthy homosexual men: evidence of a new acquired cellular immunodeficiency.NEJM 1981;305:1425-31.

5. Belpomme D, Irigaray P, Ossondo M, Vacque D, Martin M. Prostate cancer as an environmental disease: an ecological study in the French Caribbean islands, Martinique and Guadeloupe. Int J Oncol 2009;34:1037-1044.

6. Merryweather-Clarke AT, Pointon JJ, Jouanolle AM, Rochette J, Robson KJ. Geography of HFE C282Y and H63D mutations. Genet Test 2000;4:183-98.

.