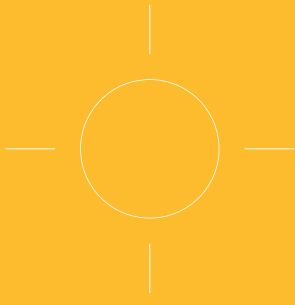


FOCUS ON



**EPIDEMIOLOGY**

Public health issues have now emerged as social issues in their own right. Hardly a day goes by without the results of another epidemiological study coming to the public's attention, giving rise to comments in the media and the blogosphere or motivating political decisions. Some studies reveal that exposure to certain environmental factors increases the risk of disease. There is debate as to the possible, presumed or proven impact that diet, pesticides, atmospheric pollution or electromagnetic waves may have on health.

With such an abundance of information on topics concerning the interaction between health and environment, there is a growing need for standard “tools” for interpreting available information correctly. It was with this in mind that Inserm decided to produce this document, which sets out to clarify a number of common and sometimes complex notions encountered in epidemiology, drawing on concrete examples to define and illustrate them. This publication, though not exhaustive, is intended to provide interested readers – who are not necessarily specialists – with a few keys for a clearer understanding of the results of epidemiological studies.



Jim Borgman, published in the *Cincinnati Enquirer*

### What is epidemiology exactly?

Epidemiology seeks to **measure the frequency of a health event** in a given population and, at the same time, determine its biological, medical, environmental and socio-economic **causes**.

Ultimately, its aim is to identify factors (atmospheric pollutants, diet, etc.) considered to influence the onset of the health event so that they can be limited or eliminated.

Epidemiologists achieve this by gathering data based on the observation of populations of individuals in good or poor health in order to estimate the levels of exposure to influencing factors.

First, however, they make a hypothesis as to the possible causes of the pathology (“factor X increases the risk of onset of Y”), which they then seek to confirm. The hypothesis (from the Greek hypo meaning “under” and thesis meaning “proposition”) is a vital part of the scientific approach adopted for generating new knowledge. It is built on a very wide range of existing data, such as the increased incidence of a pathology, human toxicology or animal experimentation.

**Health event:** disease or health disorder.  
**The frequency of a health event** can be characterised in two ways:

- **Prevalence:** proportion of a given population affected by the health event at a given time. It is expressed as the number of cases in a total population.

> *In 2008, some 800,000 people were affected by Alzheimer's disease in France (out of an estimated total population of 60 million). The prevalence is therefore more than 13 cases per 1,000 (or 1.3%).*

**A disease is considered rare when its prevalence is less than 1 case per 2,000.**

- **Incidence:** percentage of new cases of the health event studied in a given population over a given period of time. Incidence is expressed as a number of cases, a rate or percentage of the total population per unit time.

> *During the week from 6 April to 12 April 2009, 33 new cases of influenza per 100,000 inhabitants were counted in France (source: Inserm Sentinelles network).*

Different types of epidemiological study are used to determine trends in health event frequency and identify the **related risk factors**.

**Risk factors:** the characteristics of an individual (e.g. age, gender, existing illnesses, genetic characteristics, lifestyle or dietary habits) that may be associated with the onset of a particular disease.

> *A high cholesterol level is a risk factor for cardiovascular diseases in humans.*

## Different types of epidemiological study

Epidemiologists use three main types of study: those which gather information on health events affecting individuals, those which gather information concerning population sub-groups and, lastly, those which provide an overview of a series of studies already published (literature reviews, meta-analyses, technical reports). The randomised trial is another type of epidemiological study, though it is used more rarely. The first two types of study and the randomised trial are described in detail below.

## Studies that gather information from individual volunteers

**Cohort studies** consist in monitoring a group of individuals recruited at a time when they were unaffected by the health event in question. The goal is to **measure the onset of new cases of the health event** among this group of persons, while recording individual risk factors, and to compare onset trends of new cases among exposed and non-exposed individuals. Many cohort studies are prospective studies.

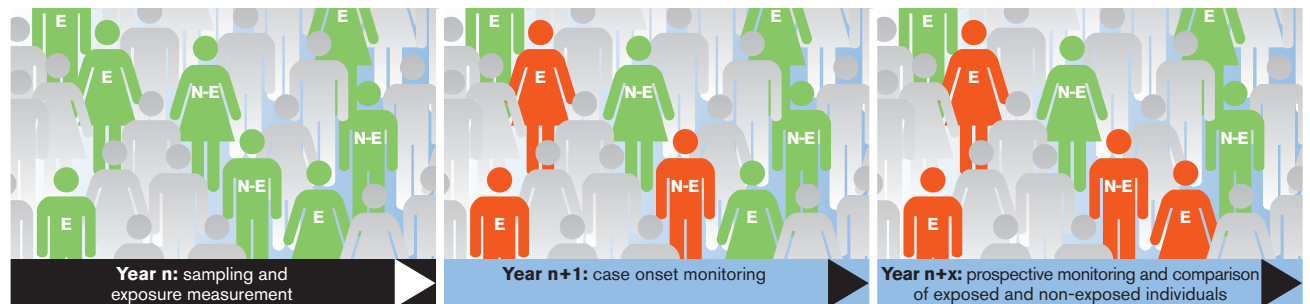
### Examples of cohorts

- > **The E3N cohort in France:** a group of 100,000 women insured with the MGEN (*Mutuelle générale de l'éducation nationale*) who have been monitored for the onset of colorectal and breast cancer since 1990.
- > **The Framingham cohort** is a group of about 5,300 men and women recruited in the town of Framingham in the United States between 1948 and 1952. The group was set up to monitor the onset of cardiovascular diseases. Monitoring of this cohort is still in progress.
- > **The Gazel cohort** was set up in France in 1989. It is made up of about 20,000 employees of EDF-GDF, the national electric and gas utilities, and is used to monitor a range of health issues.

**Case-control studies** consist in recruiting a group of individuals with the condition (health event) being monitored and another group (called the controls) who are unaffected by the condition. **The level of exposure to one or more risk factors** among the two groups is assessed through questionnaires or assays. Case-control studies are retrospective by definition as participants are recruited after the onset of the disease.

> *The Interphone study is a case-control study coordinated by IARC (the International Agency for Research on Cancer). Its purpose is to determine whether exposure to radiation from mobile telephones is associated with the risk of cancer. Persons who had developed certain types of tumour were recruited for the study (thus after the problem had been diagnosed) and questioned about their past use of the mobile telephone. Another group of individuals who had not developed similar tumours were also recruited and asked the same questions about their past use of the mobile telephone.*

**Cohort study (usually prospective)**



**N-E:** Non-exposed individuals **E:** Exposed individuals

**Red square:** Individuals affected by the health event **Green square:** Individuals unaffected by the health event

**Case-control study (always retrospective)**



**Red square:** Cases **Green square:** Controls

**Studies that gather information concerning population sub-groups**

**Time-series studies** consist in monitoring changes in the exposure level to a given factor (atmospheric pollutants) in a single location (a city for example) and comparing the results with trends in new cases of a health event. It only concerns the short-term effects of exposure (within a time frame of a few days or weeks) and can only be used when daily records of “cases” (such as hospital admissions) are available.


**Ecological studies** compare the onset frequency of a pathology at the same time among several populations presenting different risk factors.

> *Studies comparing average alcohol consumption in a number of geographical regions with the incidence of liver cancer in the same regions show that the disease has a higher incidence in regions with high alcohol consumption.*

**Randomised trials, a type of intervention study, are rarely used in epidemiology**


Unlike observation studies, which are used to gather information under the normal living conditions of the individuals concerned, “**intervention**” studies compare the impact of interventions that are specially planned for the study of individuals or groups at different times.

**Randomised trials** are an example of this type of study. Before the study begins, two groups of individuals, identical in terms of age, gender and other risk factors regarding the disease, are formed using a random draw method (which is why the trial is called “randomised”). One group will be “exposed” (to a drug for example), the other will be a “non-exposed” or placebo group. Both groups are monitored over time so that the onset of the health event in question (e.g. increased blood pressure or the onset of a particular pathology) can be compared among the two groups. It is the predominant type of study in clinical epidemiology and ensures that the exposed and non-exposed groups are comparable right from the start. For obvious ethical reasons, however, randomised trials are not used to characterise the long-term effect of any exposure thought to have a detrimental effect on health. Observational studies are more appropriate in this case.

 **CAUTION! A RISK FACTOR IS NOT NECESSARILY RESPONSIBLE FOR THE ONSET OF THE ILLNESS.**

A case-control study or cohort study alone cannot demonstrate whether a relationship of cause and effect exists between a risk factor and the onset of a pathology.

Proving that a risk factor is not only associated with the onset of the disease but is also responsible for it is quite another story!

SEE THE STUDY  
TABLE ON P.12 

## How can we prove that a factor causes the onset of a disease?

If a study can demonstrate that the risk of disease increases with exposure to the factor in question (strong association) and if that study does not include any significant biases (see page 9), then the factor in question is considered a risk factor. It does not, however, provide irrefutable proof that a causal relationship exists between the risk factor and the onset of the disease.

Epidemiological studies identify associations between exposure factors and the risk of onset of a disease. The strength of these associations is quantified through various statistical measurements, including **relative risk**.

**Exposure:** refers to a person's contact with a pollutant found in the environment. It generally depends on the level of the pollutant in each microenvironment (or environmental compartment: indoor air, outdoor air, food, drinking water) where the persons monitored live and their habits (such as diet).

**Dose:** quantity of pollutant absorbed by the body (it can be expressed in g, mg, mg by unit weight (of the person in question) or, in the case of radiation, by unit of energy per kg of tissue, or Gray). It depends on exposure and certain physiological (e.g. respiratory rate) or behavioural characteristics of the individual.

**Dose-effect relationship:** relationship between exposure to a substance (dose) and changes observed in physiological function or health (effect).

Starting with known exposure levels and the dose-effect relationship, the impact of the factor can be determined for a given population, for example, in terms of the number of new cases that can be attributed to this factor in France each year.

**Relative risk:** ratio of the incidence in an exposed group versus a non-exposed group.

➤ *If the risk of death is 2 cases per 100,000 non-exposed persons monitored for one year, and 3 cases per 100,000 exposed persons monitored for one year, then the relative risk is  $3/2 = 1.5$ . The annual risk of death is therefore 50% higher in the exposed population than in the non-exposed group.*

Various arguments involving several disciplines are taken into consideration to prove that an associated risk factor is, in fact, a causal factor:

**the toxicological effects** of exposure to the risk factor, as observed in animals,

**a dose-effect relationship** between exposure to this factor and the onset of the health event,

**the revelation of the biological and physiological action mechanisms** of the associated factor.

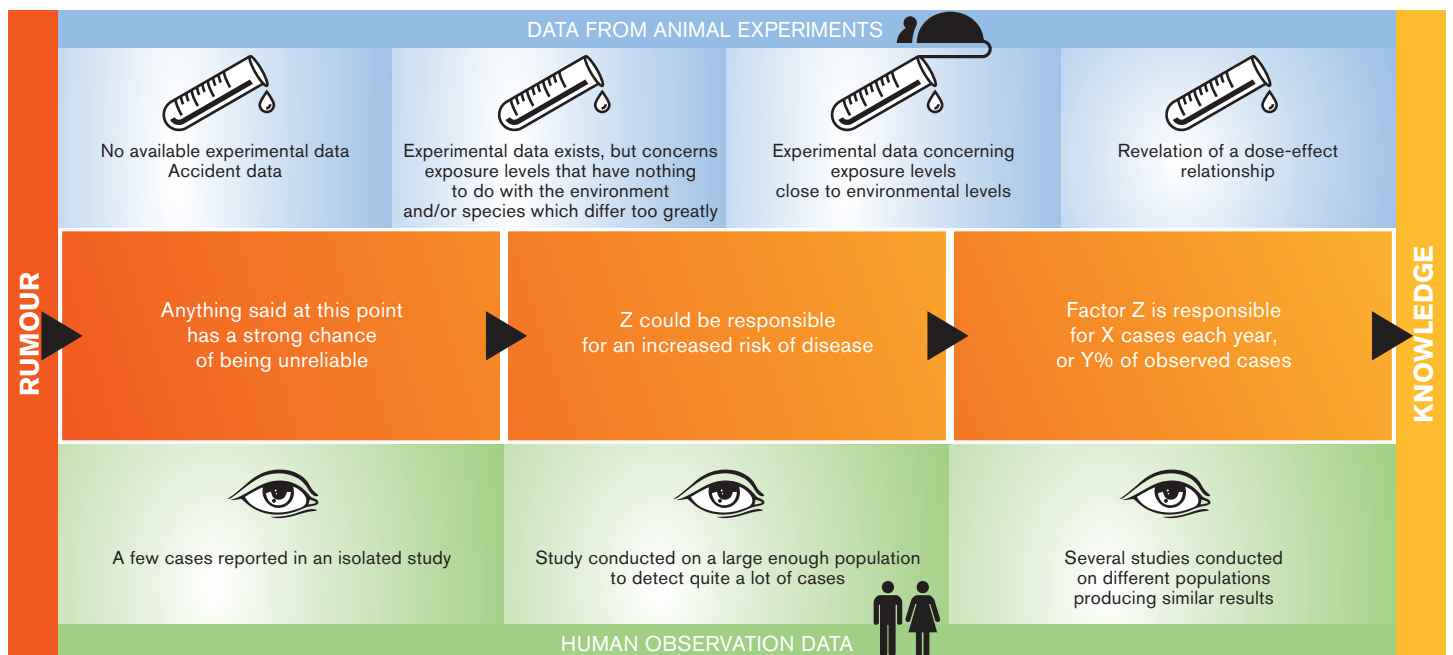
Characterising the dose-effect relationship is key to the search for a causal relationship.

Gathering all this data is, in fact, part of a long, drawn-out process and five to ten years can elapse between the time the research hypothesis is formulated and the acquisition of definitive proof that a causal relationship exists between the risk factor and the onset of a disease.

 **TO USE A BOXING METAPHOR, PROOF IS OBTAINED “ON POINTS” IN EPIDEMIOLOGY. NEVER BY KNOCK-OUT!**

Like boxers, epidemiologists have to go more than one round to reach their goal. Strong evidence is built on a body of results acquired from a number of epidemiological, toxicological and clinical studies, not from a single study. Why? Because in practice, observational studies come up against methodological problems all along the way, making it impossible for them to produce irrefutable data on their own. Consequently, all studies need to be compared with others to lessen the impact of any errors and to test other, competing hypotheses.

Both in and out of the laboratory, the road from rumour to knowledge is long and there are several stops along the way





## Uncertainty is at the heart of any epidemiological study

Since epidemiology concerns individuals observed under “normal” living conditions and not under laboratory conditions, any conclusions reached must be qualified. First, there is rarely just one **cause of a disease**. Second, **low exposure doses** are hard to detect and represent another obstacle to the conclusive detection of risk factors. Last, the process of gathering and analysing observation data can lead to **errors** that epidemiologists must strive to identify, then take into account.

Apart from some special cases (fulminant viruses, exposure to very high doses of pollutants), the vast majority of diseases have multiple causes (multifactorial pathologies).

The cause of a health event can be seen as a constellation of components acting together and/or instead of one another. This rules out a “one cause-one effect” approach.

The different types of possible nested causes are not mutually exclusive but are explained at different levels. The cause of the disease can thus be cellular, genetic, behavioural, environmental and societal at one and the same time (see box).

### Genetics and environment

A molecular biologist will consider that lung cancer can be caused by the mutation of the p53 gene (genetic cause), a cell biologist will describe cancer as a deregulated cell cycle (cellular cause due to genetic mutation), while an epidemiologist will see the disease as being caused by an environmental factor (passive exposure to tobacco smoke possibly causing a genetic mutation). The existence of a strong genetic element in a disease does not mean that the cause cannot be environmentally related.

In this respect, the epidemiologist's point of view can be compared with that of the geneticist, psychologist, biologist or sociologist.



### It can be very hard to demonstrate that a factor has absolutely no effect on the risk of disease.

A study that fails to show any significant increase in the risk of a given disease in an exposed population does not prove that the factor in question is absolutely harmless. Absence of proof of a risk is not proof of its absence.

This difficulty is inherent in statistical tools. When a risk is low, it may be impossible for a study – no matter how well designed and extensive – to demonstrate the responsibility of that risk. It may, at best, conclude that, where it exists, the excess risk associated with a factor is below a certain value.

> **Smoking and lung cancer:** *Some people may smoke all their lives and die without cancer (so the cause is not sufficient). Conversely, some people suffer from lung cancer even though they have never smoked or been significantly exposed to tobacco smoke in a passive manner (so cause is not necessary). And yet smoking is the prime cause of lung cancer. Cigarette manufacturers have long used these apparent paradoxes in an attempt to refute the established fact that smoking is responsible for the onset of cancer.*

> **Multifactorial pathology:** *The onset of asthma is promoted by exposure to atmospheric pollutants caused by road traffic. But it can also be accentuated by exposure to pollutants inside the home or to natural allergens like pollen.*

**The risk of error is an inherent part of any epidemiological study.** The important thing is to identify and make allowance for such risks. They are related to the effects of variability between each individual. These effects are known as biases.

### Variability

Identifying every individual suffering from a disease in order to determine the frequency of that disease in a population is often hard to do for reasons of logistics and cost. Statistical theories show that the frequency of the disease can simply be studied in a sub-group (or sample) of the population, provided that this sample is obtained by a random-draw method from the entire population (this is known as random sampling).

The result of random sampling is only valid as an average. Sampling must be carried out several times and the average value of the different samples calculated to obtain the true frequency of the disease. The difference between the value obtained from a sample and the “true” value, which can only be obtained by studying the entire population, is due to **random fluctuations**. The smaller the sub-group and/or the rarer the disease studied, the greater these fluctuations will be.

Statistics methods can define a **confidence level** according to the size of the sample.

**Random fluctuations:** If, for example, there is a 5% frequency of asthma in a population at a given time and random sampling is carried out on 100 individuals in this population, the number of asthma sufferers drawn with the random sample will not always be 5. There may be 7 in one sample, then 4, 3 or 6 in other samples. None of these values is correct, but they are all randomly distributed about the “true” value of 5.

### Biases

Other errors apart from those induced by sampling fluctuations need to be taken into account. Biases are among them. There are three types of bias: selection bias, measurement bias and confounding bias.

**Selection bias:** this type of bias covers a number of situations where the effect of exposure to the risk of disease has been incorrectly estimated. This can be for a number of reasons: inappropriate selection of populations for comparison, no data available for one part of the population or loss of contact with a large number of individuals during the study.

#### > An example of selection bias

*A case-control study is conducted to assess the impact of smoking on lung cancer. All the **cases** of lung cancer diagnosed at a hospital chest and lung unit are recruited and compared with **controls** considered as representative of the general population, and recruited from the ENT department of the same hospital where they are undergoing treatment for health disorders unrelated to cancer. As smoking can be a risk factor in ENT disorders, the control population is probably not representative of the general population.*

**The confidence interval (CI)** defines a minimum and a maximum value between which the true value for an entire population is to be found for a given risk of error. The larger the sample, the smaller the interval and thus the more accurate the estimation will be.

> *If 4 persons from a sample population of 100 are affected by a disease, it can be said that the true frequency value of this disease in the population is between 0 and 8 (with a 5% risk of error). In this case, the 95% confidence interval is between 0 and 8.*

> *If 52 persons from a sample population of 1,000 suffer from the disease, then the CI is between 3.8 and 6.4, which gives a more accurate estimation.*

It is likely that the “control” group here includes a higher proportion of smokers than the general population, thus reducing the differences between cases and controls with regard to smoking. This is a **selection bias**. There is a risk therefore that the study will underestimate the impact of smoking on the risk of lung cancer if this bias is not identified and taken into consideration.

**Measurement bias:** This type of bias occurs in the acquisition (questionnaire, varying degrees of accuracy of the system used, observer, etc.), recording or transmission of data on the disease, exposure or any other characteristic of the persons studied.

➤ **An example of measurement bias**

If a questionnaire is used in a study to estimate alcohol consumption in a context where heavy drinking is viewed negatively by society, the participants in the study will tend to underdeclare their alcohol consumption. This leads to a measurement error on alcohol consumption. If undetected, this **measurement error** leads to a bias in the estimated impact of alcohol on health.



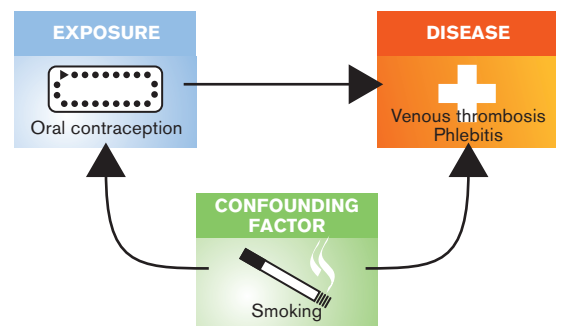
**A measurement of exposure or a biological parameter can be imperfect without necessarily distorting the estimation of the association between the factor in question and the health event.**

Different approaches can be adopted to determine the impact of these measurement errors on study results, or even correct them.

**Confounding bias:** This refers to cases where exposure and the health event are simultaneously and independently influenced by an external factor that has not been taken into consideration.

**The existence of potential biases does not invalidate a study.** Confounding biases, in particular, can generally be corrected through a statistical approach that makes it possible to compare individuals who are exposed to the factor of interest and those who are not, as from the time when the information on these confounding factors was acquired.

**Example of a confounding factor in the case of oral contraception**

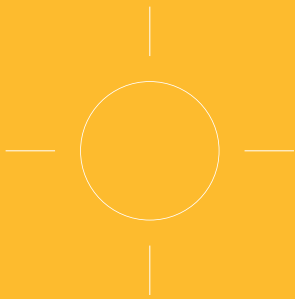


**Oral contraception is used more frequently by smokers. Smoking increases the risk of venous thrombosis. This makes it hard to determine whether the increased risk of thrombosis in oral contraceptive users is not wholly or partly related to smoking. For this reason, smoking is said to be the confounding factor in this relationship.**

We have already seen that uncertainty is key to conducting and understanding epidemiological studies.

Technically, some aspects of uncertainty, namely those due to sampling fluctuations, are quantified by the confidence interval. Other sources of uncertainty are related to study biases that cannot be quantified or corrected and which, in some cases, are simply not known.

More generally, we must remember that uncertainty is an inherent part of any scientific approach, where progress is achieved by being able to challenge facts that were previously considered as certainties.



# EPIDEMIOLOGY

This document is a collective work with contributions from the DELICE group, which was set up at the initiative of Séverine Ciancia (head of Inserm's press department, DISC) to help explain epidemiology. The group is made up of Sandrine Blanchard, Assistant Managing Editor of the weekly *Le Monde Magazine*; Gérard Bréart, Director of the *Institut thématique multi-organismes Santé publique*, Dominique Donnet-Kamel, in charge of the *Associations de malades* mission, Inserm-DISC; François Faurisson, Scientific and Medical Adviser at Eurordis, the European Organisation for Rare Diseases; Rémy Slama, epidemiologist, Inserm unit 823, Avenir team on "Environmental epidemiology applied to the study of human fecundity and reproduction"; Alfred Spira, Director of *GIS-IReSP-Institut de recherche en santé publique*.

**December 2009**



Département  
Information scientifique  
et communication

[presse@inserm.fr](mailto:presse@inserm.fr)

**Comparative table of studies**

	Individuals		Populations	
	<b>COHORT STUDY</b>	<b>CASE-CONTROL STUDY</b>	<b>TIME-SERIES STUDY</b>	<b>ECOLOGICAL STUDY</b>
<b>PARTICIPANTS</b>	Individuals unaffected by the health event when recruited for the study	Persons with the disease (cases) and without the disease (controls) when recruited for the study	Population in which risk factors are measured at different periods	Two or more populations in which risk factors are measured at the same time
<b>INFORMATION GATHERED</b>	<ul style="list-style-type: none"> <li>■ Risk factors: initial and/or during the study</li> <li>■ Onset (incidence) of event</li> </ul>	<ul style="list-style-type: none"> <li>■ Level of exposure to the risk factor</li> <li>■ Presence of health event</li> </ul>	<ul style="list-style-type: none"> <li>■ Population exposure indicator at different periods</li> <li>■ Onset of event</li> </ul>	<ul style="list-style-type: none"> <li>■ Risk factors</li> <li>■ Onset of event</li> </ul>
<b>ADVANTAGES</b>	<ul style="list-style-type: none"> <li>■ Precise, unbiased measurement</li> <li>■ Possibility of studying the involvement of the risk factor in diseases other than those initially studied</li> </ul>	<ul style="list-style-type: none"> <li>■ Fewer participants required than for a cohort study</li> <li>■ Several risk factors can be studied at the same time</li> <li>■ Lower cost</li> </ul>	<ul style="list-style-type: none"> <li>■ Exhaustive study of the population concerned</li> <li>■ Easier to set up than an individual study</li> </ul>	<ul style="list-style-type: none"> <li>■ Exhaustive study of the population concerned</li> <li>■ Easier to set up than an individual study</li> </ul>
<b>DRAWBACKS</b>	<ul style="list-style-type: none"> <li>■ Large number of participants required</li> <li>■ Higher cost</li> <li>■ Long wait before obtaining results</li> </ul>	<ul style="list-style-type: none"> <li>■ Frequent exposure measurement biases (retrospective)</li> </ul>	<ul style="list-style-type: none"> <li>■ Only the short-term effects of exposure can be studied</li> </ul>	<ul style="list-style-type: none"> <li>■ Hard to make effective allowance for confounding factors</li> </ul>

**CONTINUED  
FROM P.5**

